Mémoire présenté
pour l'obtention du diplôme de Statisticien mention Actuariat
et l'admission à l'Institut des Actuaires
le 26/09/2023

Par : **Antoine Heranval**

Titre : **Introduction d'une méthode de tarification pour des risques extrêmes**

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : Mission Risques Naturels

Signature :

Directeur du mémoire en entreprise :

Nom :
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Résumé

Le marché de l'assurance est étroitement lié aux risques extrêmes, caractérisés par leur rareté mais aussi par leur potentiel catastrophique. Ces risques peuvent avoir des conséquences économiques et humaines désastreuses. L'assurance joue un rôle central dans l'atténuation et la gestion de ces risques, en offrant la possibilité d'avoir une protection financière face à des événements tels que les catastrophes naturelles majeures, les pandémies ou plus récemment les attaques cyber. Pour faire face à ces risques, les assureurs peuvent proposer des polices d'assurance spécifiques qui aident à couvrir les pertes financières liées à ces événements exceptionnels. Cependant, la mise en place et la gestion de ces polices posent des défis particuliers. La nature imprévisible de ces risques et leur impact potentiellement considérable rendent la tarification difficile. Les assureurs doivent également faire face à un manque de données historiques fiables sur les événements extrêmes. Étant donné leur caractère rare, il y a souvent peu de données disponibles pour les évaluer. Cela rend la modélisation statistique traditionnelle et la projection des pertes potentielles complexes.

Dans ce mémoire, nous présentons une méthode visant à tarifer ou à prédire le coût d'un risque extrême. Pour cela, nous combinons des informations individuelles, peu nombreuses, avec des données collectives. Par "extrême", nous entendons que les distributions des pertes aléatoires sont caractérisées par une queue lourde. Notre méthode combine la théorie des valeurs extrêmes et la théorie de la crédibilité bayésienne.

Nous proposons deux applications de notre méthode. La première application porte sur le risque cyber, tandis que la seconde application concerne l'estimation rapide du coût des inondations peu après leur survenue. Bien que la partie consacrée aux catastrophes naturelles soit plus développée, en raison de notre collaboration avec la Mission Risques Naturels (MRN), un groupement technique de France Assureurs, l'application au risque cyber démontre la capacité de généralisation de notre méthode. Nous illustrons comment, même avec des données limitées, il est possible de proposer un tarif cohérent pour évaluer un risque.

Mots-clés : *risques extrêmes*, *tarification*, *théorie des valeurs extrêmes*, *crédibilité bayesienne*,*arbres de régression*,*catastrophes naturelles*, *risque Cyber*.

Abstract

The insurance market is closely linked to extreme risks, characterized by their rarity but also their catastrophic potential. These risks can have disastrous economic and human consequences. Insurance plays a central role in mitigating and managing these risks by offering the possibility of financial protection against events such as major natural disasters, pandemics, or more recently, cyber attacks. To address these risks, insurers can propose specific insurance policies that help cover financial losses related to these exceptional events. However, the establishment and management of these policies pose particular challenges. The unpredictable nature of these risks and their potentially considerable impact make pricing difficult. Insurers also face a lack of reliable historical data on extreme events. Given their rare nature, there is often little data available to evaluate them. This makes traditional statistical modeling and projecting potential losses complex.

In this thesis, we present a method aimed at pricing or predicting the cost of an extreme risk. To achieve this, we combine limited individual information with collective data. By "extreme," we mean that the distributions of random losses are characterized by heavy tails. Our method combines the theory of extreme values and Bayesian credibility theory.

We provide two applications of our method. The first application focuses on cyber risk, while the second application concerns the rapid estimation of flood costs shortly after their occurrence. Although the section dedicated to natural disasters is more developed due to our collaboration with the Mission Risques Naturels (MRN), a technical group within France Assureurs, the application to cyber risk demonstrates the generalizability of our method. We illustrate how, even with limited data, it is possible to propose a consistent pricing to assess a risk.

Keywords : *pricing*, *extreme value theory*, *Regression trees*, *Bayesian credibility theory* ,*natural disaster*, *cyber risk*.

Note de Synthèse

Le marché de l'assurance est étroitement lié aux risques extrêmes, caractérisés par leur rareté mais aussi par leur potentiel catastrophique. Ces risques peuvent avoir des conséquences économiques et humaines désastreuses. L'assurance joue un rôle central dans l'atténuation et la gestion de ces risques, en offrant la possibilité d'avoir une protection financière face à des événements tels que les catastrophes naturelles majeures, les pandémies ou plus récemment les attaques cyber. Pour faire face à ces risques, les assureurs peuvent proposer des polices d'assurance spécifiques qui aident à couvrir les pertes financières liées à ces événements exceptionnels. Cependant, la mise en place et la gestion de ces polices posent des défis particuliers. La nature imprévisible de ces risques et leur impact potentiellement considérable rendent la tarification difficile. Les assureurs doivent également faire face à un manque de données historiques fiables sur les événements extrêmes. Étant donné leur caractère rare, il y a souvent peu de données disponibles pour les évaluer. Cela rend la modélisation statistique traditionnelle et la projection des pertes potentielles complexes. Dans ce mémoire, nous présentons une méthode visant à tarifer ou à prédire le coût d'un risque extrême. Pour cela, nous combinons des informations individuelles, peu nombreuses, avec des données collectives. Par "extrême", nous entendons que les distributions des pertes aléatoires sont caractérisées par une queue lourde. De plus, les événements que nous cherchons à estimer ne se sont jamais produits pour un assuré donné, ou se sont produits seulement quelques fois. Dans ce deuxième cas, nous pouvons affiner l'évaluation préalable avec cet historique.

Nous proposons deux applications de notre méthode. La première application porte sur le risque cyber, tandis que la seconde application concerne l'estimation rapide du coût des inondations peu après leur survenue. Bien que la partie consacrée aux catastrophes naturelles soit plus développée, en raison de notre collaboration avec la Mission Risques Naturels (MRN), un groupement technique de France Assureurs, l'application au risque cyber démontre la capacité de généralisation de notre méthode. Nous illustrons comment, même avec des données limitées, il est possible de proposer un tarif cohérent pour évaluer un risque.

Pour cela, nous utilisons la théorie des valeurs extrêmes qui cherche à analyser la queue de distribution de variables aléatoires. L'objectif est de quantifier les scénarios extrêmes, pour lesquels la valeur de ces variables aléatoires est élevée par rapport aux valeurs typiques. D'un point de vue statistique, les événements climatiques, notamment les catastrophes naturelles, sont souvent associés à de tels événements. Lorsqu'ils surviennent, ces événements peuvent prendre des valeurs très faibles ou très élevées et entraîner des conséquences considérables. De même, cela s'applique à certains événements liés au cyber. La théorie des valeurs extrêmes s'avère donc particulièrement appropriée pour la tarification des risques extrêmes.

Dans le cadre de cette théorie, nous adoptons l'approche des dépassements de seuil, également

connue sous le nom de Peaks-Over-Threshold (PoT), dont le résultat fondamental a été établi par Balkema & De Haan (1974). Cette approche repose sur l'utilisation des observations qui ont dépassé un seuil spécifique.

Considérons des variables aléatoires Y_1, Y_2, ..., Y_n, indépendantes et identiquement distribuées (i.i.d), avec une fonction de répartition inconnue F. Dans la suite, nous noterons $\overline{F}$ la fonction de survie associée, définie comme $\overline{F}(y) = \mathbb{P}(Y_i > y)$ pour tout y.

Dans l'approche PoT (Peaks-Over-Threshold), une observation est considérée comme extrême si elle dépasse un seuil préalablement choisi, noté u. Lorsqu'une observation est jugée extrême, on définit l'excès correspondant comme la différence entre cette observation et le seuil u. En 1975, Pickands III démontre que si $\overline{F}$ satisfait la propriété suivante :

$$\lim_{t\to+\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0}, \forall y > 0,$$

avec $\gamma_0 > 0$, alors

$$\lim_{u\to+\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}_{\sigma_{0u},\gamma_0}(z)| = 0,$$

où $\sigma_{0u} > 0$ et $\overline{H}_{\sigma_{0u},\gamma_0}$ est la fonction de survie d'une loi non dégénérée qui appartient nécessairement à la famille des lois de Pareto généralisée (GP) avec

$$\overline{H}_{\sigma_{0u},\gamma_0}(z) = \left(1 + \gamma_0 \frac{z}{\sigma_{0u}}\right)^{-1/\gamma_0}, z > 0,$$

σ_{0u} est un paramètre d'échelle et $\gamma_0 > 0$ un paramètre de forme, appelé indice de queue, reflétant l'épaisseur de la queue de F. Plus γ_0 est grand, plus la queue de distribution est lourde. Cette théorie nous permet donc d'obtenir la distribution des événements extrêmes. Cependant, pour pouvoir fixer un coût, nous la relions à la théorie de la crédibilité. Dans la théorie de la crédibilité, un assuré est associé à un facteur de risque θ qui suit une distribution a priori que nous appellerons p. Dans le cadre le plus simple, les pertes individuelles subies par cet assuré sont supposées être indépendantes et identiquement distribuées, notées $(Y_1, ..., Y_n)$ conditionnellement à $\theta = t$, avec g_t représentant leur densité.

Une approche de crédibilité adaptée au contexte des risques extrêmes doit satisfaire la condition selon laquelle $\int g_t(y)p_i(t)dt$ correspond à la densité d'une distribution de Pareto généralisée. Cette condition garantit que la prime de crédibilité reflète adéquatement les caractéristiques des événements extrêmes et leur distribution. Ainsi, en choisissant correctement la distribution a priori, il est possible de construire une approche de crédibilité qui est cohérente avec la théorie des valeurs extrêmes. On peut montrer que dans ce cas, la prime de crédibilité est donnée par :

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

avec

$$r = \frac{1}{\gamma},$$

$$\lambda = \frac{\sigma}{\gamma}.$$

$\sum_{i=1}^{n} Y_i$ et n correspondent quand à eux à l'historique des données. L'un des défis consiste alors à estimer au mieux les paramètres γ et σ qui décrivent l'expérience collective. Pour cela, nous proposons d'utiliser une méthode basée sur des arbres de régression. Cette méthode nous permet d'intégrer des covariables tout en préservant une excellente interprétabilité des résultats. Grâce à l'utilisation des arbres de régression, nous sommes en mesure d'estimer les valeurs optimales de γ et σ qui correspondent au mieux aux données observées, et qui permettent d'obtenir des primes de crédibilité précises et fiables.

Pour tenir compte des caractéristiques de l'assuré $\mathbf{X} \in \mathbb{R}^d$, qui peuvent influencer son groupe de risque spécifique, nous souhaitons que $\mathbf{X}$ ait un impact sur la distribution a priori utilisée pour déterminer la prime de crédibilité. Nous supposons l'existence de fonctions $\mathbf{x} \to r(\mathbf{x})$ et $\mathbf{x} \to \lambda(\mathbf{x})$ (et par conséquent, de fonctions $\mathbf{x} \to \gamma(\mathbf{x})$ et $\mathbf{x} \to \sigma(\mathbf{x})$) qui décrivent l'hétérogénéité entre les classes d'assurés. Ces fonctions permettent de modéliser la relation entre les caractéristiques de l'assuré $\mathbf{X}$ et les paramètres γ et σ. En utilisant ces fonctions, nous pouvons estimer les paramètres γ et σ en fonction des caractéristiques de chaque assuré. Cela nous permet d'ajuster la prime de crédibilité en tenant compte de l'hétérogénéité entre les classes d'assurés et de mieux refléter les risques individuels.

Pour calibrer l'a priori, nous considérons alors que nous disposons de $(Z_1, \mathbf{X}_1, \cdots, Z_N, \mathbf{X}_N)$, i.i.d. réplications de $(Z_1, \mathbf{X}_1)$. La calibration de l'a priori consiste ainsi à estimer les fonctions de régression $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$, en supposant que $Z_1|\mathbf{X}_1 = \mathbf{x}_1$ est distribué selon une distribution de Pareto généralisée avec des paramètres $(\gamma(\mathbf{x}_1), \sigma(\mathbf{x}_1))$.

Dans ce mémoire, nous présentons une méthode introduite par Farkas, Lopez & Thomas (2021) qui utilise des arbres de régression pour créer des classes homogènes en termes de distributions extrêmes. Une caractéristique intéressante de cette méthodologie est la création d'un nombre fini de classes de risque, pour lesquelles les valeurs $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$ sont constantes. Cette approche permet de regrouper les assurés en fonction de leurs caractéristiques communes, facilitant ainsi la tarification et la gestion des risques. De plus des résultats théoriques validant l'utilisation de cette méthode sont prouvés dans Farkas, Heranval, Lopez & Thomas (2021).

Nous présentons ensuite la première application de notre méthode pour estimer le coût d'un événement Cyber. Nous nous appuyons sur la base de données de la Privacy Rights Clearinghouse (PRC), largement utilisée pour l'étude du risque cyber. Dans un premier temps, nous allons appliquer la méthode CART GPD pour regrouper les événements en classes homogènes en termes de distribution extrême, ce qui nous fournira la base pour la théorie de la crédibilité. Pour illustrer la tarification, nous utiliserons un exemple fictif, car malheureusement, nous n'avons pas accès à des données de sinistralité permettant de confronter cette méthode dans le cadre du risque Cyber.

Nous commençons donc par appliquer la méthode CART GPD pour créer des classes qui sont homogènes dans leurs comportements extrêmes. Dans la base de données PRC, nous utilisons le nombre d'enregistrements, qui est la variable que nous cherchons à prédire. Pour chaque événement, nous utilisons les informations suivantes :

— les types de violation,
— les types d'organisation,
— les sources.

L'arbre obtenu avec cette méthode est présenté dans la figure 5.1. Notre arbre comporte 8 feuilles, avec des séparations basées sur toutes les variables fournies. Cet arbre nous permet de déduire la distribution des événements en fonction de certaines covariables. On peut observer que tous les paramètres de forme γ sont supérieurs à 1, ce qui illustre les valeurs élevées prises par les nombres

d'enregistrements. Toutes les classes ne sont pas semblables et la répartition est hétérogène comme on peut s'y attendre. On a deux classes "peu extrêmes" (même si le paramètre de forme reste supérieur à 1) qui regroupent 65% des événements et une classe au contraire très extrême avec un γ de 3.35 qui regroupe 2% des événements. On a une répartition qui permet de bien discriminer les événements. Cependant, il est important de noter que cette répartition est très dépendante des données et en particulier de la qualité des données. Elle présente donc forcément des incertitudes.

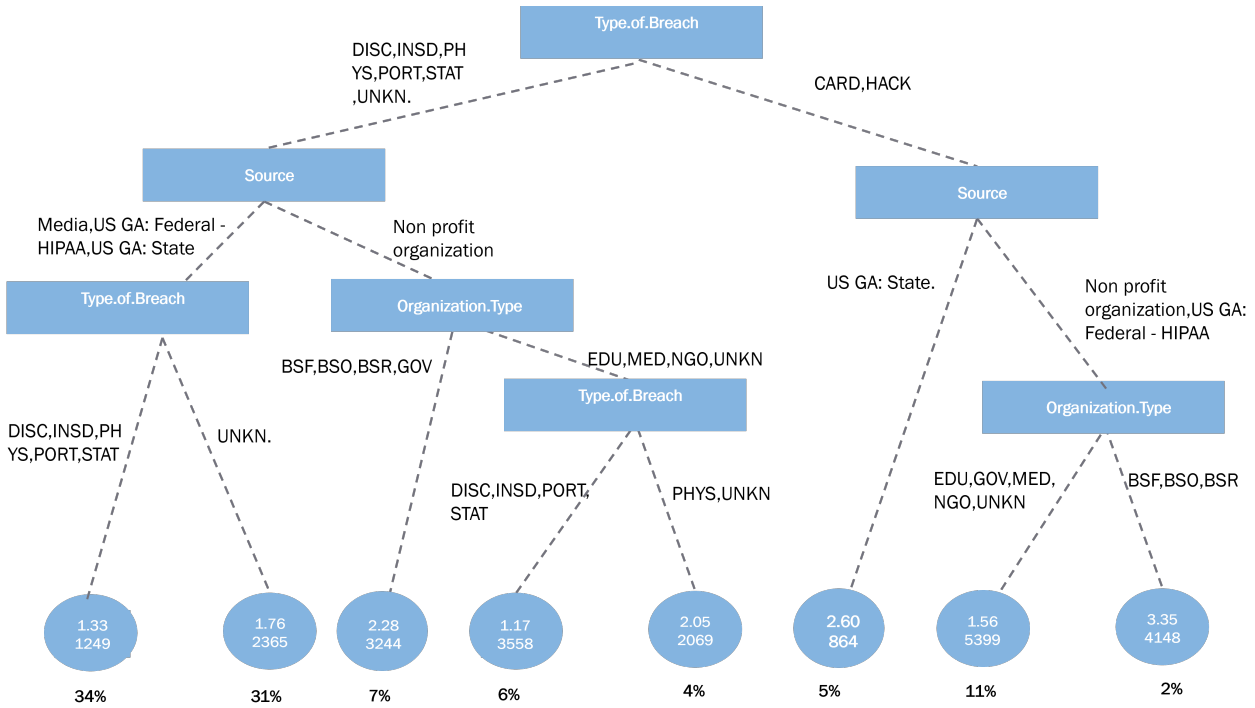


FIGURE 1 – Arbre de régression GP obtenu pour les événements cyber. Pour chaque feuille, la valeur du paramètre de forme γ (première ligne) et du paramètre d'échelle σ (deuxième ligne) sont indiquées. Le pourcentage d'observations attribué à chaque feuille est aussi indiqué.

Pour la deuxième illustration, nous appliquons notre méthode pour estimer le coût des événements d'inondation. Afin d'avoir un point de comparaison, nous introduisons également une méthode basée sur la fréquence et la gravité. Notre objectif se concentre sur l'estimation des conséquences des inondations, sans prendre en compte la modélisation du phénomène lui-même. Notre démarche s'inscrit dans le cadre d'une mission d'appui à France Assureurs, visant à dimensionner les réponses en cas de gestion de crise liée aux événements naturels. Ainsi, nous cherchons à estimer le coût d'un événement d'inondation rapidement après sa survenue.

La principale contrainte de notre approche est de pouvoir fournir une estimation rapide du coût des événements d'inondation. Cependant, nous avons également accordé une attention particulière au développement d'une méthode compréhensible, facile à utiliser et dotée de paramètres contrôlables. Conscients de la difficulté de cet exercice et de l'existence d'incertitudes inhérentes à toute estimation, nous avons cherché à laisser une part de contrôle aux gestionnaires de risques. Nos méthodes reposent avant tout sur l'expertise métier et la comparaison avec les données historiques recueillies par la MRN. Ainsi, nous avons privilégié une approche qui intègre ces éléments pour fournir une estimation aussi précise que possible.

L'arbre obtenu avec cette méthode est présenté dans la figure 6.2. Notre arbre comporte 6 feuilles,

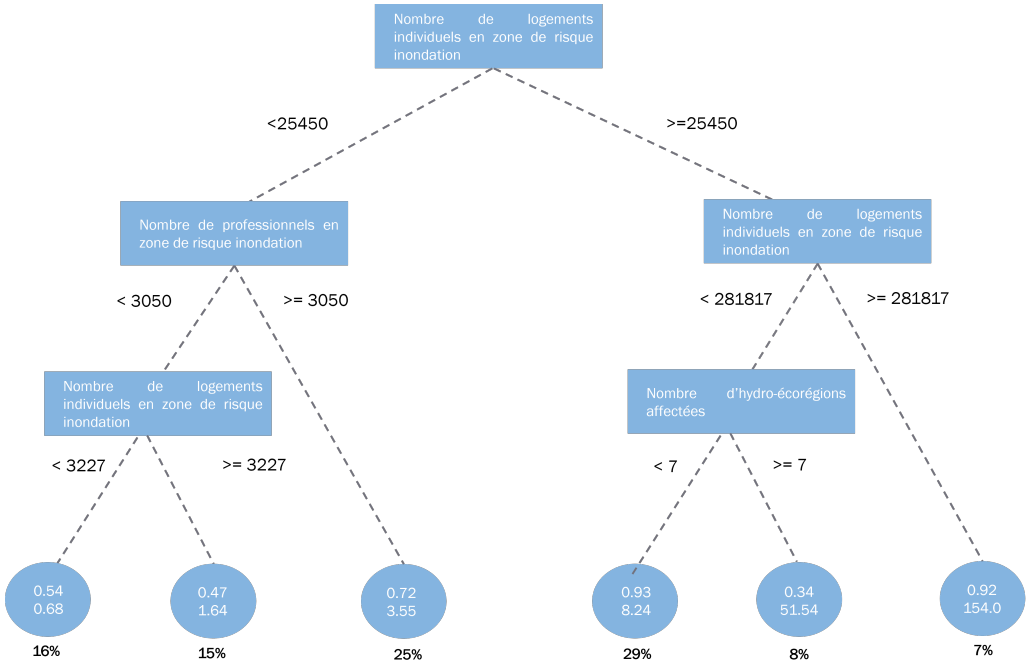


FIGURE 2 – Arbre de régression GPD obtenue pour les événements inondations. Pour chaque feuille on indique le paramètre de forme γ (première ligne), le paramètre d'échelle σ à 10^{-5} (deuxième ligne). Les pourcentages d'observations dans chaque feuille sont aussi présentés.

avec des séparations basées sur 3 critères :

— le nombre de logements individuels en zone de risque inondation,
— le nombre de professionnels en zone de risque inondation,
— le nombre d'hydro-écorégions affectées.

Cette répartition semble cohérente. Les deux premières covariables reflètent l'exposition aux inondations ainsi que la densité de population de la zone touchée, tandis que la troisième covariable représente l'étendue de l'événement. Le cas le plus extrême correspond à la feuille la plus à droite, qui a un paramètre de forme de 0.92 et contient 7% des événements. Cette feuille correspond à une proportion importante de logements individuels touchés et à une zone étendue. Nous pouvons ensuite appliquer la théorie de la crédibilité avec la formule suivante :

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \frac{\sum_{k=1}^{n} y_{i,k} + (\frac{pi\sigma_j}{\gamma_j})}{n + \frac{1}{\gamma_j} - 1}.$$

Sur la base de test, nous constatons que près de la moitié des communes que nous tentons d'estimer n'ont pas d'historique de sinistres ($n = 0$), et seulement 20% des communes ont connu plus d'un événement passé ($n > 1$). Cette méthode est donc particulièrement adaptée, car l'expérience des événements passés n'est pas suffisante pour fournir une estimation fiable. L'information apportée par les classes du CART GPD permet d'enrichir l'estimation dans ces cas.

En résumé, au jour J, nous avons un événement qui correspond à une liste de communes impactées. Nous calculons les variables d'entrée pour cet événement, qui seront utilisées pour la classification CART GPD. Ensuite, pour chaque commune, nous pouvons estimer un coût total en fonction de la

classe obtenue. Le coût total correspond à la somme des coûts des communes. Nous obtenons des résultats encourageants qui seront utiles à la fédération dans son processus de gestion. En effet, ces méthodes ont été créées dans une logique d'aide à la décision. Nous obtenons des résultats encourageants qui seront utiles à la fédération dans son processus de gestion. En effet, ces méthodes ont été créées dans une logique d'aide à la décision.

Dans ce mémoire, nous avons introduit une méthode permettant d'estimer le coût des événements extrêmes en tenant compte de l'historique des données, même incomplet, ainsi que des covariables. Cela est particulièrement précieux dans une perspective de gestion des risques, car le coût est construit étape par étape, et des informations importantes sont accessibles à chaque étape. En effet, en plus de fournir une estimation du coût, nous estimons la distribution des événements extrêmes. Nous avons présenté deux applications avec des résultats encourageants. Cependant, cette méthode pourrait encore être améliorée en incorporant des informations sur l'intensité de l'aléa, notamment pour les inondations, en utilisant des variables météorologiques. Actuellement, notre étude se base uniquement sur des variables indiquant l'exposition et l'ampleur des événements, sans prendre en compte leur intensité. Bien que l'utilisation de telles informations à l'échelle de l'événement puisse être complexe, elles représentent l'une des principales perspectives pour améliorer nos estimations. Une application intéressante consisterait à étudier l'influence des variables météorologiques sur le coût en utilisant cette méthode. Nous pourrions ainsi estimer l'évolution potentielle du coût selon différents scénarios du GIEC. Notre méthode de tarification offre une approche exhaustive pour la tarification des risques extrêmes, en tenant compte des problèmes liés à la disponibilité des données, qui sont partiellement pris en compte dans cette approche bayésienne, ainsi que de la nature extrême des événements. Elle pourrait s'avérer utile pour la tarification des risques émergents, et pourrait ainsi être testée pour le risque cyber et les évènements climatiques au sein des compagnies d'assurance.

Synthesis note

The insurance market is closely linked to extreme risks, characterized by their rarity as well as their catastrophic potential. These risks can have devastating economic and human consequences. Insurance plays a central role in mitigating and managing these risks, offering the possibility of financial protection against events such as major natural disasters, pandemics, or more recently, cyber attacks. To address these risks, insurers can offer specific insurance policies that help cover the financial losses associated with these exceptional events. However, the establishment and management of these policies pose particular challenges. The unpredictable nature of these risks and their potentially considerable impact make pricing difficult. Insurers also have to contend with a lack of reliable historical data on extreme events. Given their rarity, there is often limited data available to evaluate them. This makes traditional statistical modeling and projecting potential losses complex.

In this thesis, we present a method aimed at pricing or predicting the cost of an extreme risk. To achieve this, we combine limited individual information with collective data. By "extreme", we mean that the distributions of random losses are characterized by heavy tails. Furthermore, the events we aim to estimate have either never occurred for a specific policyholder or have only happened a few times. In the latter case, we can refine the prior assessment with this historical data. We offer two applications of our method. The first application focuses on cyber risk, while the second application pertains to the rapid estimation of the cost of floods shortly after their occurrence. Although the section devoted to natural disasters is more elaborate due to our collaboration with the Mission Risques Naturels (MRN), a technical group of France Assureurs, the application to cyber risk demonstrates the generalizability of our method. We illustrate how, even with limited data, it is possible to provide a consistent pricing for risk assessment.

For this purpose, we use the extreme value theory, which seeks to analyze the tail distribution of random variables. The objective is to quantify extreme scenarios, where the value of these random variables is high compared to typical values. From a statistical standpoint, climatic events, especially natural disasters, are often associated with such occurrences. When they happen, these events can take on very low or very high values and lead to considerable consequences. Similarly, this applies to certain events related to cyber risks. The theory of extreme values is thus particularly suited for the pricing of extreme risks.

Within the framework of this theory, we adopt the threshold exceedance approach, also known as Peaks-Over-Threshold (PoT), whose fundamental result was established by Balkema & De Haan (1974). This approach relies on the use of observations that have exceeded a specific threshold. Consider random variables Y_1, Y_2, ..., Y_n, independent and identically distributed (i.i.d), with an unknown cumulative distribution function F. Throughout, we will denote $\overline{F}$ as the associated survival function, defined as $\overline{F}(y) = \mathbb{P}(Y_i > y)$ for any y.

In the Peaks-Over-Threshold (PoT) approach, an observation is considered extreme if it surpasses a pre-chosen threshold, denoted as u. When an observation is deemed extreme, the corresponding excess is defined as the difference between that observation and the threshold u. In 1975, Pickands III demonstrated that if $\overline{F}$ satisfies the following property:

$$\lim_{t\to+\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0}, \forall y > 0,$$

with $\gamma_0 > 0$, then

$$\lim_{u\to+\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}_{\sigma_{0u},\gamma_0}(z)| = 0,$$

where $\sigma_{0u} > 0$ et $\overline{H}_{\sigma_{0u},\gamma_0}$ is the survival function of a non-degenerate distribution that necessarily belongs to the family of generalized Pareto (GP) distributions with

$$\overline{H}_{\sigma_{0u},\gamma_0}(z) = \left(1 + \gamma_0 \frac{z}{\sigma_{0u}}\right)^{-1/\gamma_0}, z > 0,$$

σ_{0u} is a scale parameter, and $\gamma_0 > 0$ is a shape parameter, referred to as the tail index, reflecting the thickness of the tail of F. The larger γ_0 is, the heavier the tail of the distribution.

This theory thus allows us to obtain the distribution of extreme events. However, in order to determine a cost, we connect it to the theory of credibility. In credibility theory, an insured individual is associated with a risk factor θ following a prior distribution which we will call p. In the simplest framework, individual losses suffered by this insured are assumed to be independent and identically distributed, denoted as $(Y_1, ..., Y_n)$ conditionally on $\theta = t$, where g_t represents their density. A credibility approach suited for the context of extreme risks must satisfy the condition that $\int g_t(y)p_i(t)dt$ corresponds to the density of a generalized Pareto distribution. This condition ensures that the credibility premium adequately reflects the characteristics of extreme events and their distribution. Thus, by appropriately choosing the prior distribution, it is possible to construct a credibility approach that is consistent with extreme value theory. In this case we can find that the credibility premium is given by:

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

with

$$r = \frac{1}{\gamma},$$

$$\lambda = \frac{\sigma}{\gamma}.$$

$\sum_{i=1}^n Y_i$ and n corresponding to the data history. One of the challenges then is to accurately estimate the parameters γ and σ that describe the collective experience. To achieve this, we propose using a method based on regression trees. This approach allows us to use covariates while maintaining excellent interpretability of the results. Through the use of regression trees, we can estimate the optimal values of γ and σ that best fit the observed data, thus enabling the generation of accurate and reliable credibility premiums.

To take into account the characteristics of the insured $\mathbf{X} \in \mathbb{R}^d$, which can influence their specific risk group, we aim for $\mathbf{X}$ to impact the a priori distribution used to determine the credibility premium. We assume the existence of functions $\mathbf{x} \to r(\mathbf{x})$ and $\mathbf{x} \to \lambda(\mathbf{x})$ (and consequently, functions $\mathbf{x} \to \gamma(\mathbf{x})$ and $\mathbf{x} \to \sigma(\mathbf{x})$) that describe the heterogeneity among insured classes. These functions allow modeling the relationship between the insured's characteristics $\mathbf{X}$ and the parameters γ and σ. By using these functions, we can estimate the parameters γ and σ based on each insured's characteristics. This enables us to adjust the credibility premium while considering the heterogeneity among insured classes and better reflecting individual risks.

To calibrate the prior, we consider that we have $(Z_1, \mathbf{X}_1, \cdots, Z_N, \mathbf{X}_N)$, i.i.d. replications of $(Z_1, \mathbf{X}_1)$. Calibrating the prior involves estimating the regression functions $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$, assuming that $Z_1|\mathbf{X}_1 = \mathbf{x}_1$ is distributed according to a generalized Pareto distribution with parameters $(\gamma(\mathbf{x}_1), \sigma(\mathbf{x}_1))$. In this thesis, we present a method introduced by Farkas, Lopez & Thomas (2021) that uses regression trees to create homogeneous classes in terms of extreme distributions. An interesting feature of this methodology is the formation of a finite number of risk classes, for which the values $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$ are constant. This approach allows grouping policyholders based on their common characteristics, thus facilitating risk pricing and management. Additionally, theoretical results validating the use of this method are demonstrated in Farkas, Heranval, Lopez & Thomas (2021).

Next, we present the first application of our method to estimate the cost of a Cyber event. We rely on the Privacy Rights Clearinghouse (PRC) database, widely used for studying cyber risk. Initially, we will apply the CART GPD method to group events into homogeneous classes in terms of extreme distribution, which will provide the basis for credibility theory. To illustrate the pricing, we will use a fictional example since, unfortunately, we lack loss data access to validate this method in the context of Cyber risk.

Thus, we begin by applying the CART GPD method to create classes that are homogeneous in their extreme behaviors. In the PRC database, we employ the number of records, which is the variable we aim to predict. For each event, we utilize the following information:

— types of breaches,
— types of organizations,
— sources.

The tree obtained using this method is presented in Figure 5.1. Our tree has 8 leaves, with splits based on all the provided variables. This tree allows us to deduce the distribution of events based on certain covariables. It can be observed that all the shape parameters γ are greater than 1, which illustrates the high values taken by the numbers of records. Not all classes are similar, and the distribution is heterogeneous as expected. There are two "less extreme" classes (even though the shape parameter remains above 1) that group 65% of the events, and on the contrary, a very extreme class with a γ of 3.35 that encompasses 2% of the events. The distribution allows for a good discrimination of events. However, it's important to note that this distribution is highly dependent on the data, particularly the data quality. As a result, it inherently carries uncertainties.

For the second illustration, we apply our method to estimate the cost of flood events. To provide a point of comparison, we also introduce a method based on frequency and severity. Our focus is on estimating the consequences of floods, without considering the modeling of the phenomenon itself. This approach aligns with our mission to support France Assureurs in sizing responses to natural events-related crisis management. Thus, our aim is to estimate the cost of a flood event after its occurrence.

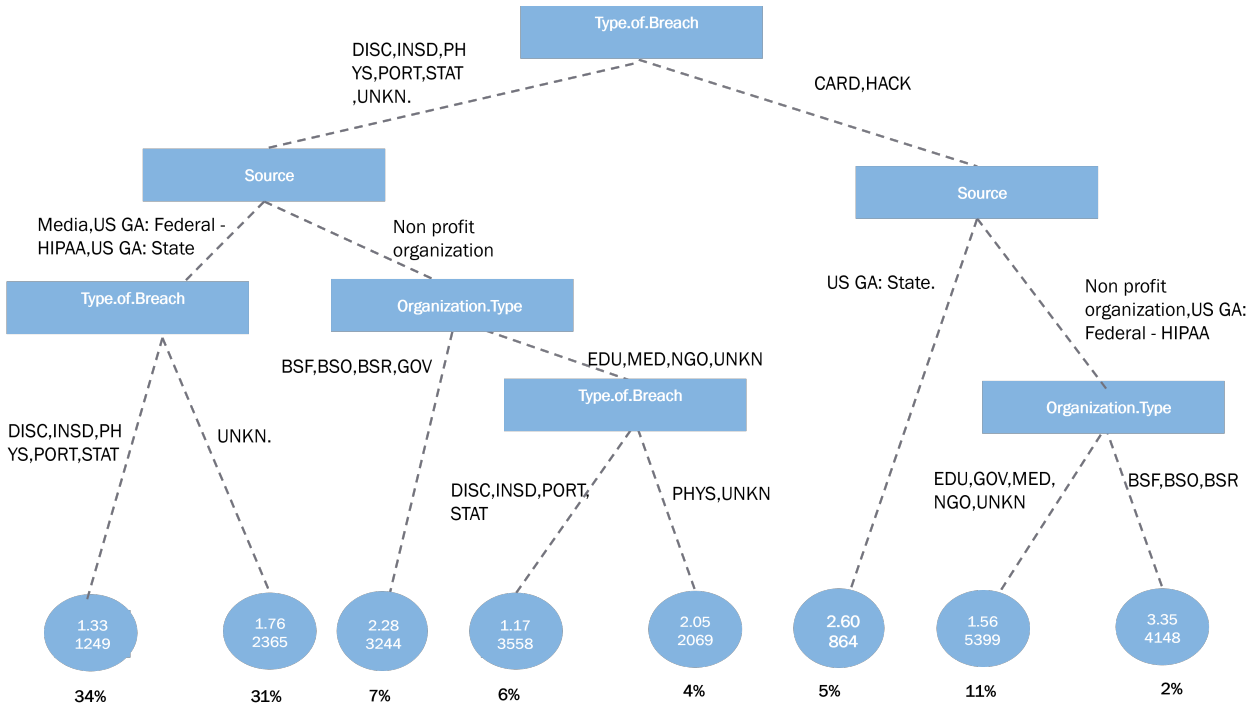


Figure 3 – Regression GP tree obtained for cyber events. For each leaf, the value of the shape parameter γ (first row) and the scale parameter σ (second row) is provided. The percentage of observations assigned to each leaf is also indicated.

The primary constraint of our approach is to be able to provide a quick estimation of the cost of flood events. However, we have also given special attention to developing a method that is understandable, user-friendly, and equipped with controllable parameters. Recognizing the challenge of this task and the inherent uncertainties in any estimation, we have aimed to leave room for risk managers' control. Our methods primarily rely on domain expertise and comparison with historically robust data collected by the MRN. Thus, we have favored an approach that integrates these elements to provide the most accurate estimation possible. The tree obtained using this method is presented in Figure 6.2. Our tree has 6 leaves, with separations based on 3 criteria:

— the number of individual households in flood risk zones,
— the number of professionals in flood risk zones,
— the number of hydro-ecoregions affected.

This repartition appears consistent. The first two covariates reflect exposure to floods as well as population density in the affected area, while the third covariate represents the extent of the event. The most extreme case corresponds to the rightmost leaf, which has a shape parameter of 0.92 and contains 7% of the events. This leaf corresponds to a significant proportion of individual homes affected and a widespread area. We can then apply the credibility theory using the following formula:

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \frac{\sum_{k=1}^{n} y_{i,k} + (\frac{pi\sigma_j}{\gamma_j})}{n + \frac{1}{\gamma_j} - 1}.$$

Based on the test dataset, we observe that nearly half of the cities we are attempting to estimate have no historical claims records ($n = 0$), and only 20% of cities have experienced more than one past event ($n > 1$). This method is thus particularly suitable since the experience from past events is

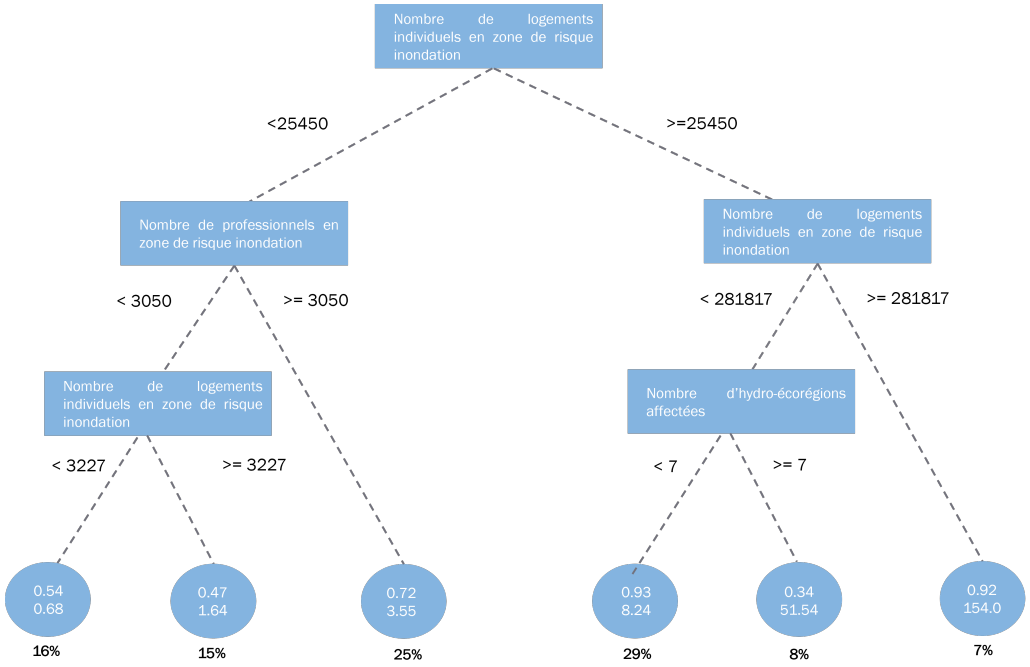


Figure 4 – Regression tree obtained for flood events using the GPD method. For each leaf, the shape parameter γ (first line) and the scale parameter σ at 10^{-5} (second line) are indicated. The percentages of observations in each leaf are also presented.

insufficient to provide a reliable estimate. The information provided by the CART GPD classes helps enhance the estimation in these cases. In summary, we have an event that corresponds to a list of impacted cities. We calculate input variables for this event, which will be used for the CART GPD classification. Then, for each city, we can estimate a total cost based on the obtained class. The total cost corresponds to the sum of city costs. We obtain promising results that will be useful for the federation in its management process. Indeed, these methods were developed with a decision support logic.

In this dissertation, we have introduced a method for estimating the cost of extreme events, taking into account historical data, even when incomplete, and covariates. This is particularly valuable from a risk management perspective, as the cost is constructed step by step, and important information is accessible at each stage. Indeed, beyond providing a cost estimation, we estimate the distribution of extreme events. We have presented two applications with promising results. However, this method could be further enhanced by incorporating information about the intensity of the hazard, especially for floods, using meteorological variables. Currently, our study relies solely on variables indicating exposure and event magnitude, without considering their intensity. While utilizing such event-level information may be complex, it represents a significant prospect for improving our estimations. An interesting application would involve studying the influence of meteorological variables on costs using this method. This way, we could estimate the potential cost evolution under different scenarios from the IPCC.

Our pricing method offers a comprehensive approach for pricing extreme risks, accounting for challenges related to data availability, partially addressed in this Bayesian approach, as well as the extreme nature of events. It could prove useful for pricing emerging risks and could be tested for cyber risks and climate-related events within insurance companies.

Remerciements

Je tiens à remercier mes directeurs de thèse, Olivier Lopez et Maud Thomas, qui m'ont permis d'accomplir ces travaux, en particulier à Olivier pour son encadrement et la relecture de ce mémoire.

Je tiens également à exprimer ma sincère gratitude envers Sarah Gerin, Eric Petitpas et Lilian Pugnet pour leur encadrement au sein de la Mission Risques Naturels. Ce fut un réel plaisir de travailler avec vous. Un grand merci à Lilian d'avoir continué à assurer le suivi.

Je souhaite également exprimer ma reconnaissance envers toutes les personnes qui ont contribué à la réalisation et au bon déroulement de mes travaux. En particulier, je tiens à remercier toutes les sociétés d'assurance qui ont partagé leurs données avec la MRN et France Assureurs. Votre soutien opérationnel a été déterminant pour la réussite de ces travaux. Je tiens également à remercier l'équipe de la Privacy Rights Clearinghouse qui nous a mis à disposition des données pour le domaine de la Cyber.

Un grand merci à l'ENSAE qui m'a permis de mener à bien ce mémoire tout en occupant le poste de coordinateur. Merci en particulier à Caroline Hillairet pour son accueil chaleureux. Enfin, je tiens à remercier l'ISUP et les équipes administratives pour la gestion de mon dossier.

Contents

Introduction

Le marché de l'assurance est étroitement lié aux risques extrêmes, caractérisés par leur rareté mais aussi par leur potentiel de catastrophe. Ces risques peuvent avoir des conséquences économiques et humaines désastreuses. L'assurance joue un rôle central dans l'atténuation et la gestion de ces risques, en offrant la possibilité d'avoir une protection financière face à des événements tels que les catastrophes naturelles majeures, les pandémies ou plus récemment les attaques cyber.

Pour faire face à ces risques, les assureurs peuvent proposer des polices d'assurance spécifiques qui aident à couvrir les pertes financières liées à ces événements exceptionnels. Cependant, la mise en place et la gestion de ces polices posent des défis particuliers. La nature imprévisible de ces risques et leur impact potentiellement considérable rendent la tarification difficile. Les assureurs doivent également faire face à un manque de données historiques fiables sur les événements extrêmes. Étant donné leur caractère rare, il y a souvent peu de données disponibles pour les évaluer. Cela rend la modélisation statistique traditionnelle et la projection des pertes potentielles complexe.

Dans ce mémoire, nous présentons une méthode visant à tarifer ou à prédire le coût d'un risque extrême. Pour cela, nous combinons des informations individuelles, peu nombreuses, avec des données collectives. Par "extrême", nous entendons que les distributions des pertes aléatoires sont caractérisées par une queue lourde. De plus, les événements que nous cherchons à estimer ne se sont jamais produits pour un assuré donné, ou se sont produits seulement quelques fois. Dans ce deuxième cas on peut affiner l'évaluation préalable avec cet historique.

Ce mémoire propose deux applications de notre méthode. La première application concerne l'estimation rapide du coût des inondations rapidement après leur occurrence, tandis que la seconde application porte sur le risque cyber. Bien que la partie consacrée aux catastrophes naturelles soit plus développée, étant donné notre collaboration avec la Mission Risques Naturels, un groupement technique de France Assureurs, l'application au risque cyber démontre la capacité de généralisation de notre méthode. Nous illustrons comment, même avec des données limitées, il est possible de proposer un tarif cohérent pour évaluer un risque.

Dans une première partie nous allons présenter ces risques extrêmes, les mécanismes d'assurance possibles et la tarification associée. Nous présenterons ensuite les deux théories mathématiques au coeur de notre méthode, à savoir, la théorie des valeurs extrêmes et la théorie de la crédibilité bayésienne et nous expliquerons comment elles peuvent être combinées pour la tarification des risques. La partie suivante se concentrera sur l'estimation du "prior", c'est-à-dire la connaissance préalable des risques. Nous présenterons une méthode qui repose sur l'utilisation de covariables et qui offre des résultats théoriques nouveaux. Enfin, nous présenterons les applications de notre méthode, en commençant par une introduction des données nécessaires à nos études. Nous mettrons en évidence comment notre approche peut être appliquée, notamment aux catastrophes naturelles et au risque cyber. Nous présenterons dans ce mémoire les différents aspects de notre méthode, depuis les bases théoriques jusqu'à ses applications pratiques.

Chapter 1

Assurance et risques extrêmes

Les risques extrêmes sont un enjeu majeur pour le marché de l'assurance en raison de leur nature imprévisible et de leurs conséquences potentiellement dévastatrices. Les événements tels que les catastrophes naturelles majeures, les pandémies, les attaques cyber, pour ne citer qu'elles, peuvent avoir un impact considérable sur les sociétés, tant sur le plan économique que sur le plan humain. Ce mémoire se concentre exclusivement sur les catastrophes naturelles et le risque cyber, qui font l'objet de nos applications.

L'augmentation de la fréquence et de l'intensité des événements climatiques, tels que les sécheresses et les inondations, et les conséquences associés font partie des préoccupation principales aujourd'hui. Ces phénomènes peuvent causer des pertes humaines désastreuses, financières massives et provoquer des perturbations durables dans de nombreux secteurs.

La menace des attaques cyber s'est aussi intensifiée, avec des incidents de grande ampleur touchant des organisations publiques et privées. Ces attaques peuvent paralyser des infrastructures critiques, voler des données sensibles et perturber les opérations normales des entreprises, avec des conséquences économiques et sociales importantes.

Face à ces risques extrêmes grandissants, il est capital de renforcer les stratégies d'atténuation et de gestion en place. Cela implique notamment une évaluation précise des risques, le renforcement des mécanismes d'assurance adaptés, ainsi que la mise en œuvre de mesures de prévention et de résilience visant à réduire les pertes potentielles.

Dans ce chapitre, nous commençons par présenter ces deux risques et leur évolution potentielle. Ensuite, nous abordons les mécanismes d'assurance possibles pour ces risques. Enfin, nous examinons les difficultés rencontrées lors de la tarification.

1.1 Catastrophes Naturelles

1.1.1 Impact assurantiel

La France n'est pas épargnée par les événements climatiques extrêmes et elle a toujours connu des catastrophes naturelles d'ampleur, des exemples marquants incluent les tempêtes extrêmes de Lothar et Martin en 1999, ou les inondations historiques de la Seine en 1910. L'histoire récente le confirme, elle a été, dans les 15 dernières années, soumise à des événements significatifs de tempêtes, de grêle,

d'inondations, de submersions marines, mais aussi aux effets dévastateurs sur le bâti, d'épisodes répétés de sécheresse géotechnique, et même dans ces territoires d'Outre-Mer aux cyclones. Le graphique, 1.1 illustre clairement la situation. Nous pouvons observer des événements de grande ampleur, engendrant des coûts très élevés, tels que plusieurs tempêtes coûtant plusieurs milliards d'euros, le cyclone Irma ou les inondations de la Seine et la Loire en 2016. De plus, on constate également une fréquence élevée de sinistres, avec une sinistralité attritionnel qui en cumul pèsent lourd dans la charge. La sinistralité global est en augmentation, passant d'une moyenne annuelle d'environ 2 milliards d'euros sur la période 1989-2009 à 2,8 milliards sur la période 2010-2019, et même à 3,8 milliards pour la période 2016-2019. Selon les projections et études réalisées par la profession, cette évolution de la sinistralité risque de s'intensifier à l'avenir. En effet, l'augmentation des richesses et les effets du changement climatique vont aggraver les conséquences des catastrophes naturelles et alourdir le bilan économique.

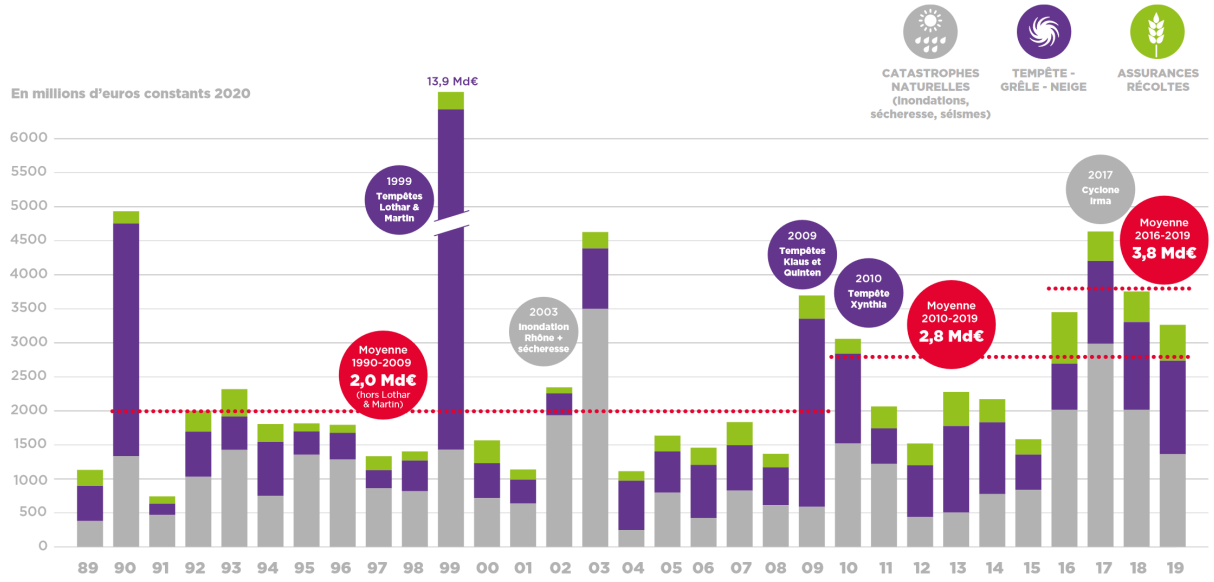


Figure 1.1 – Répartition annuelle de la charge des sinistres et principaux événements (Source : *Etude : Changement climatique et assurance à l'horizon 2040* (2021))

1.1.2 Les effets du changement climatique

Selon une récente étude publiée par France Assureurs (FA), *Etude : Changement climatique et assurance à l'horizon 2040* (2021), le montant des sinistres liés aux événements naturels pourrait atteindre 143 milliards d'euros cumulés entre 2020 et 2050, ce qui représente une augmentation de 93% par rapport à la période 1989-2019. Cela équivaut à une augmentation de 69 milliards d'euros. Il convient de rappeler que cette augmentation s'ajoute à un montant déjà élevé, avec une charge moyenne de sinistres liés aux événements naturels de 2,4 milliards d'euros par an pour 416 000 sinistres sur la période de 1989 à 2019. Au cours des dernières années, on observe une augmentation significative avec une moyenne annuelle de 3,8 milliards d'euros pour la période 2016-2019, marquée par des sécheresses importantes et des événements extrêmes tels que le cyclone Irma ou les inondations de la Seine en 2016. Cette étude prend en compte l'augmentation des richesses qui est le premier facteur d'augmentation, suivi de près par le changement climatique. Si l'on se concentre uniquement sur les effets du changement climatique, les travaux menés par la Caisse Centrale de Réassurance (CCR) et Météo-France montrent que dans le scénario le plus pessimiste du GIEC (SSP5-8.5), la sinistralité

climatique due aux catastrophes naturelles augmenterait de 50% d'ici 2050, toutes choses égales par ailleurs (richesse et population). Selon cette étude, le dérèglement climatique entraînerait donc une augmentation moyenne annuelle des sinistres de 650 millions d'euros d'ici 2050. Cette évaluation ne tient pas compte de la réduction de la sinistralité due aux mesures actuelles de prévention des risques naturels.

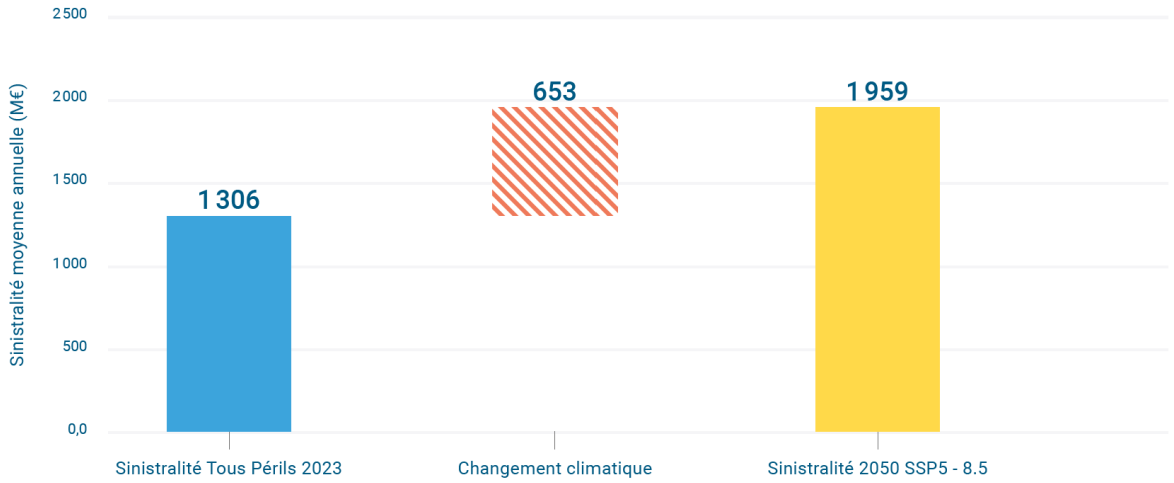


Figure 1.2 – Projection de la sinistralité annuelle moyenne des risques climatiques à l'horizon 2050, en tenant compte de la réforme Cat Nat et des effets du changement climatique selon le scénario RCP 8.5, (Source : *Rapport au Ministre de l'économie, des finances et de la souveraineté industrielle et numérique sur le régime d'indemnisation des catastrophes naturelles* (2022))

1.1.3 Les inondation

Dans ce mémoire, nous nous intéressons principalement aux inondations, qui représentent un enjeu majeur en France. En effet, une vaste partie du territoire français est exposée aux risques d'inondation, ce qui en fait l'un des aléas naturels les plus fréquemment mentionnés. Selon les données de la Mission Risques Naturels, environ 21 000 communes se trouvent en zone d'exposition aux inondations, parmi lesquelles 34 700 ont rapporté au moins un sinistre d'inondation depuis 1990. Les données concernant les zones potentiellement inondables indiquent que près d'une personne sur quatre habite dans une zone à risque d'inondation.

Le risque d'inondation touche une vaste partie du territoire français, ce qui se reflète dans les montants d'indemnisation. Entre 2016 et 2020, on observe en moyenne 700 millions d'euros d'indemnisations par an. Parmi ces montants, une part importante est attribuable à des événements extrêmes, tels que les inondations des bassins de la Seine moyenne et de la Loire en mai-juin 2016, qui ont causé à elles seules 1 500 millions d'euros de dommages, selon *L'assurance des événements naturels en 2020* (2022). Ces inondations ont touché une vaste zone où des enjeux majeurs étaient concentrés. Cependant, il y a également des inondations plus localisées et intenses qui restent gravées dans les esprits. On peut mentionner celles survenues en octobre 2020 dans le sud de la France à la suite de la tempête Alex, avec des conséquences désastreuses pour les vallées de la Roya et de la Vésubie. Ces inondations ont entraîné un lourd bilan humain et ont également causé près de 200 millions d'euros de dommages selon la CCR. La France est exposée à différents types d'inondations, et il peut parfois être difficile de les séparer ou

de déterminer l'unique phénomène générateur, car certains événements peuvent impliquer plusieurs types d'inondations simultanément. Cependant, on peut distinguer plusieurs catégories d'inondations, comme présenté dans, *Inondations, s'informer pour mieux se protéger* (2019) :

- Les inondations par débordement de cours d'eau, de type "crues lentes de plaine", se produisent lorsque les fleuves ou les rivières sortent lentement de leur lit mineur et inondent leur lit moyen, voire leur lit majeur. Ces inondations sont relativement fréquentes. Elles se caractérisent par leur lenteur et sont assez prévisible, elles surviennent après de longues périodes de pluie dans des cours d'eau déjà hauts. Elles causent généralement peu de pertes humaines directes.
- Les inondations par débordement de cours d'eau, de type "crues rapides et torrentielles", se produisent principalement dans les zones montagneuses à la suite de fortes précipitations. Elles se caractérisent par des inondations soudaines et dévastatrices, qui peuvent entraîner des pertes humaines. La montée des eaux peut être très rapide et moins prévisible.
- Les inondations par ruissellement se produisent lorsque les précipitations ne peuvent pas s'infiltrer dans le sol ou lorsque le sol est déjà saturé en eau. Suite à des fortes précipitations, concentrées sur plusieurs jours ou rapides sur plusieurs heures, les eaux ruissellent dans des zones habituellement sèches. Ce phénomène est aggravé par l'urbanisation et l'imperméabilisation des sols, qui empêchent l'absorption du surplus d'eau. En milieu urbain, cela peut provoquer des écoulements rapides avec des vitesses élevées, entraînant des dommages humains et matériels significatifs. En milieu rural, cela peut se transformer en coulées de boue, causant aussi des dégâts importants. Les inondations par ruissellement touchent l'ensemble du territoire et sont difficiles à prévoir en raison de leur nature complexe. De nombreux paramètres entrent en jeu. Il est cependant, intéressant de noter que les modèles ont récemment bénéficié d'une amélioration grâce à l'utilisation de données satellitaire.
- Les submersions marines sont des inondations qui se produisent dans les zones côtières, suite à des conditions météorologiques et océaniques défavorables. Elles peuvent être causées par des débordements du niveau de la mer, des vagues puissantes ou la rupture de systèmes de protection. En raison de sa façade maritime et de ses côtes basses, la France est particulièrement exposée à ce risque. Les submersions marines sont des inondations rapides et de courte durée qui surviennent généralement lors de tempêtes et de marées spécifiques.
- Les inondations par remontée de nappe se produisent lorsque le niveau de la nappe phréatique atteint la surface du sol. Elles surviennent généralement lors d'événements pluvieux exceptionnels qui entraînent une surcharge anormale des nappes phréatiques dans des conditions particulières. Ces inondations peuvent causer des dommages dans les sous-sols, les garages semi-enterrés ou les caves, ainsi que des fissurations dans les immeubles et des dommages au réseau routier et ferroviaire.

Les dommages causés par les inondations sont nombreux et diversifiés, et ils varient en fonction du type d'inondation. Au niveau des bâtiments individuels, les dommages les plus courants concernent les embellissements, tels que les revêtements muraux et les réseaux. Cependant, dans le cas d'inondations torrentielles ou intenses, la structure du bâtiment peut également être affectée.

Selon l'étude réalisée par la France Assureurs, *Etude : Changement climatique et assurance à l'horizon 2040* (2021), le coût des dommages liés aux inondations devrait augmenter pour atteindre 50 milliards d'euros entre 2020 et 2050. Cette augmentation est principalement due à la croissance de la richesse, qui entraîne une concentration plus importante d'entreprises, de logements et d'infrastructures. Les mêmes conclusions sont tirées pour les submersions marines, avec une augmentation estimée à 87% par rapport à la période précédente. Dans ce cas, l'impact du changement climatique est plus prononcé et contribue à hauteur de $6,5$ milliards d'euros sur les 54 milliards prévus.

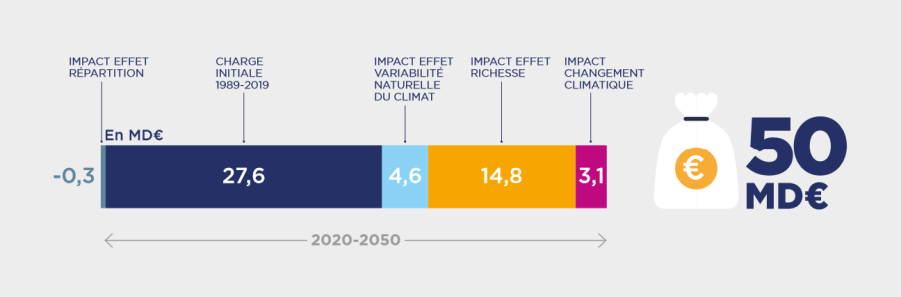


Figure 1.3 – Évolution supposée du coût lié aux inondations pour la période 2020-2050 (Source : *Etude : Changement climatique et assurance à l'horizon 2040* (2021))

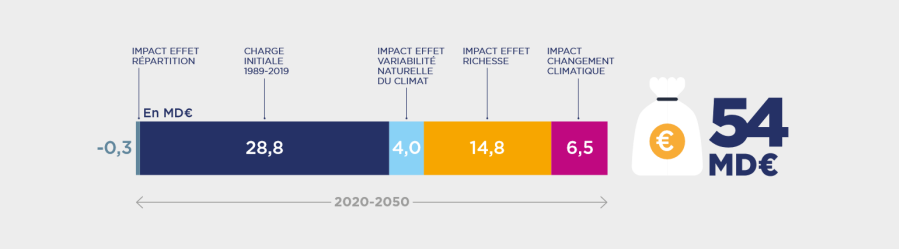


Figure 1.4 – Évolution supposée du coût lié aux submersions marines pour la période 2020-2050 (Source : *Etude : Changement climatique et assurance à l'horizon 2040* (2021))

1.2 Risque Cyber

1.2.1 Introduction

La cybercriminalité représente un défi majeur à l'heure, où les industries et les services publics s'appuient fortement sur les outils numériques. Le nombre d'incidents et d'attaques cyber a connu une forte augmentation ces dernières années, *Etat de la menace rançongiciel* (2021)), notamment en raison de la pandémie de Covid-19 et de l'essor du télétravail Lallie et al. (2021). Les avancées récentes dans la génération automatique de texte, notamment avec Chat GPT peuvent aussi renforcer les vulnérabilités. Ce phénomène croissant de cybermalveillance, qui englobe des actes tels que le piratage, le vol de données, l'usurpation d'identité et le détournement d'objets connectés, suscite de profondes inquiétudes du fait de ses conséquences dévastatrices, en particulier pour les entreprises. Le risque cyber peut émaner de divers acteurs, tels que les États, les groupes soutenus par des États, les groupes criminels, voire des individus isolés agissant en tant que hackers. L'accessibilité croissante d'outils et de logiciels malveillants sur Internet a alimenté l'émergence d'un marché du "crimeware as a service", permettant aux criminels d'acquérir des outils sophistiqués, *Climat, cyber, pandémie : le modèle assurantiel mis au défi des risques systémiques* (2022). Parmi les formes de cybermalveillance, les rançongiciel connaissent une expansion rapide. Ces programmes malveillants chiffrent les données des systèmes d'information, les rendant soit inaccessibles à leurs propriétaires, soit inutilisables. De plus, certaines attaques peuvent passer inaperçues pendant plusieurs mois, n'étant détectées que lorsque les dommages sont devenus irréversibles. Deux attaques notoires, à savoir Wannacry et NotPetya, sont des exemples frappants d'attaques par rançongiciel massives. Elles ont touché environ 300 000 machines dans 150 pays pour Wannacry et des millions de machines dans 65 pays pour NotPetya, en l'espace de quelques jours. Les pertes causées par ces attaques sont considérables, bien qu'il soit difficile de

les quantifier avec précision. Les estimations varient de plusieurs centaines de millions à 4 milliards de dollars pour Wannacry, selon *Internet Organized Crime Threat Assessment* (2018), et dépassent les 10 milliards de dollars pour NotPetya, selon *Cost of Cyber Incidents Study* (2020). Les dommages résultant de ces attaques ne sont généralement pas uniquement liés aux demandes de rançon. Dans le cas de Wannacry, le montant des rançons est connu, car les transactions effectuées sont publiques, bien qu'il soit difficile de retracer leur destination. Toutefois, le coût des rançons est négligeable par rapport aux pertes totales. Ces pertes comprennent notamment la perte d'activité ainsi que les risques juridiques et de réputation. Cela est particulièrement vrai lorsqu'une attaque de "double extorsion" se produit : le blocage des données est précédé de leur exfiltration, avec la menace de les divulguer. Si des données personnelles sont touchées, cela peut entrer dans le champ d'application du Règlement Général sur la Protection des Données (RGPD). Par ailleurs, des informations confidentielles ou des secrets industriels peuvent être révélés. À ce sujet, il est intéressant de souligner un aspect essentiel : l'évaluation précise de l'ampleur de ce phénomène est complexe en raison de la sous-déclaration des incidents qui se produisent. Cette sous-déclaration peut être attribuée à une méconnaissance des enjeux ou à une volonté délibérée de préserver les intérêts des victimes. De plus, les enjeux géopolitiques, stratégiques et sécuritaires jouent également un rôle majeur. Dans le cadre de notre étude, cet aspect revêt une importance particulière, car il influence considérablement la disponibilité des données nécessaires à la tarification. Il est important de noter que les données pertinentes sont rares, et ces facteurs ne font qu'accentuer davantage cette problématique.

1.2.2 Les développements de l'assurance

Au cours des années 1990, les premières offres d'assurance cyber ont été introduites. Ces offres étaient des formules combinant des logiciels de sécurité et une couverture d'assurance. Des collaborations entre les éditeurs de logiciels de sécurité et les compagnies d'assurance ont permis le développement de ces offres Lelarge & Bolot (2009). Au fil du temps, le marché de l'assurance cyber s'est étendu, en particulier aux États-Unis, avec l'émergence de polices spécifiquement conçues pour couvrir les risques cyber. Ces polices comprennent des clauses adaptées à l'évolution constante des cyberattaques et à la complexité croissante des systèmes d'information, Hillairet & Lopez (2022). C'est aujourd'hui un marché en forte croissance, *Insurance 2020 beyond: Reaping the dividends of cyber resilience* (2020). Selon une étude de PwC, les primes devraient passer d'environ 2,5 milliards de dollars aujourd'hui à 7,5 milliards de dollars d'ici la fin de la décennie *Insurance 2020 beyond: Reaping the dividends of cyber resilience* (2020). Les polices d'assurance cyber couvrent généralement différents aspects du risque cyber, tels que les frais de notification des violations de données, les coûts de restauration des systèmes informatiques, les pertes financières directes et indirectes, les dépenses liées à la gestion de crise et à la communication, ainsi que les frais juridiques et de responsabilité. Les polices d'assurance cyber sont régulièrement mises à jour pour prendre en compte les nouvelles formes de cyberattaques et les changements dans le paysage de la cybersécurité.

En France, l'étude "Lumière sur la Cyber-Assurance" (LUCY) réalisée par l'AMRAE (Association pour le Management des Risques et des Assurances de l'Entreprise) offre une vision précise des primes collectées ainsi que des pertes subies. Cette étude, *"LUCY : LUmière sur la CYberassuranc* (2022), s'appuie sur des données provenant de courtiers, et constitue l'une des rares sources fiables dans le domaine. Selon le dernier rapport, les assureurs ont imposé des conditions de renouvellement strictes aux entreprises après une année 2020 difficile. Cela s'est traduit par des hausses importantes des taux de primes, l'introduction de franchises élevées, atteignant en moyenne près de 4 millions d'euros pour les grandes entreprises et 228 000 euros pour les ETI, ainsi qu'une réduction notable des capacités de couverture. En conséquence, la protection offerte aux entreprises a considérablement diminué. Toutefois, ces mesures ont été fructueuses pour l'industrie de l'assurance, car le marché

de l'assurance cyber est redevenu rentable, avec des résultats techniques se rapprochant de ceux de 2019. Le ratio sinistres sur primes (S/P) s'établit à 88% pour l'ensemble du marché en 2021. Dans un système équilibré, les primes sont conçues pour absorber les coûts des sinistres, et un ratio S/P inférieur à 100% témoigne d'une anticipation adéquate de l'évolution du risque. En 2020, ce ratio était de 167%. Les grandes entreprises constituent le principal moteur du marché de l'assurance cyber, représentant 82% des 185 millions d'euros de primes versées au titre de la garantie cyber en 2021. Face à la réduction de la couverture proposée par l'industrie de l'assurance, les entreprises explorent des solutions complémentaires telles que l'auto-assurance pour gérer le risque de fréquence, la mutualisation inter-entreprises pour accroître les capacités de couverture, et la mise en place d'un régime de "catastrophe cyber" pour réduire la volatilité des sinistres de grande ampleur.

Le risque cyber est un phénomène nouveau et en constante évolution, avec une capacité d'adaptation rapide des attaquants. Cependant, il est important de noter que les données disponibles pour évaluer ce risque sont très limitées. Les comportements de déclaration des incidents cyber évoluent au fil du temps, sous l'influence de la réglementation et des changements de perception du risque. Les événements extrêmes liés aux cyberattaques peuvent causer des pertes considérables. Un autre défi est le risque d'accumulation, qui se produit lorsque de nombreux incidents cyber se concentrent dans une même période ou région, entraînant une perte de mutualisation des risques. En outre, les portefeuilles d'assurance cyber sont souvent de taille insuffisante pour absorber pleinement le risque, ce qui ajoute une pression supplémentaire sur les assureurs et les assurés.

Il est essentiel de tenir compte de ces facteurs lors de l'évaluation et de la gestion du risque cyber afin de développer des stratégies et des solutions adaptées. Dans la section suivante, nous explorerons les différents mécanismes d'assurance envisageables pour y faire face, ainsi que les mécanismes déjà en place pour gérer les catastrophes naturelles.

1.3 Les mécanismes d'assurance disponibles

1.3.1 Assurance Classique

En premier lieu il est possible d'assurer les risques extrêmes avec des contrats d'assurance classiques et en ayant recours à la réassurance de marché. La réassurance est utilisée par les compagnies d'assurance pour transférer une partie du risque qu'elles ont assumé sur d'autres assureurs, appelés réassureurs. Les réassureurs assument ainsi une part du risque en échange d'une prime versée par l'assureur cédant. Pour les risques extrêmes, les compagnies d'assurance peuvent s'appuyer sur des contrats de réassurance spécifiques qui couvrent ces situations exceptionnelles. Ces contrats de réassurance permettent aux assureurs de transférer une partie du risque lié à ces événements majeurs à d'autres acteurs spécialisés dans la gestion de ces risques. Ainsi, en combinant des contrats d'assurance classiques et une réassurance de marché, les assureurs peuvent offrir une protection adéquate contre les risques extrêmes, en limitant leur exposition financière et en garantissant une capacité suffisante pour indemniser les assurés en cas de sinistre majeur. Cette approche permet également de diversifier le risque en le répartissant entre plusieurs acteurs du marché de l'assurance et de la réassurance.

C'est ce qui est utilisé, en France, pour les dommages causés par la tempête, la grêle ou le poids de la neige, qui sont considérés comme des risques assurables et sont couverts donc par une garantie d'assurance "classique". Cette garantie est proposée par les compagnies d'assurance sur le marché et est soutenue par une réassurance privée. Ces garanties peuvent être contractuelles, facultatives ou obligatoires. La loi du 25 juin 1990 a rendu obligatoire la couverture des dommages causés par la tempête pour toute personne ou entreprise détenant un contrat d'assurance garantissant les dommages

incendie. La garantie TGN (Tempête, Grêle, Neige) offre une protection contre les dommages causés par des événements climatiques violents tels que les tempêtes, la grêle ou le poids de la neige. Elle peut couvrir les dommages aux bâtiments, aux véhicules et à d'autres biens matériels. Le montant de la garantie TGN pourrait être ajusté en fonction du niveau de risque. Cependant, il est intéressant de noter que les territoires les plus exposés ne sont pas systématiquement pénalisés par des primes d'assurance significativement plus élevées, comme le montre l'étude, Chneiweiss & Bardaji (2020). En ce qui concerne les inondations, on peut également trouver des garanties contractuelles, notamment dans d'autres pays. Par exemple, aux États-Unis et en Allemagne, des systèmes de marché sont plutôt privilégiés pour assurer ce risque, telle que décrit par Klein & Wang (2009), Surminski & Thieken (2017). Aux États-Unis, le marché de l'assurance contre les inondations est géré par le National Flood Insurance Program (NFIP), qui propose des polices d'assurance inondation aux propriétaires de biens immobiliers dans les zones à haut risque d'inondation. Les assureurs privés peuvent également proposer des polices d'assurance contre les inondations, mais elles sont souvent réassurées par le NFIP. En Allemagne, la couverture des inondations est également principalement assurée par des assureurs privés. Les compagnies d'assurance proposent des polices spécifiques qui couvrent les dommages causés par les inondations, et ces polices peuvent être souscrites volontairement par les propriétaires de biens immobiliers.

Comme nous l'avons précédemment mentionné pour le risque cyber, il est également possible de souscrire des garanties d'assurance contractuelle classique. Cette approche est étroitement liée à la notion d'assurabilité, qui est un concept essentiel dans le domaine de l'assurance. Elle prend en compte différents aspects tels que la prévisibilité du risque, la disponibilité de données statistiques fiables, la capacité d'évaluation des pertes potentielles, et la faisabilité économique de la couverture. Dans le contexte des risques extrêmes, le débat autour de l'assurabilité se concentre souvent sur la capacité des assureurs à évaluer et à couvrir des événements rares et de grande ampleur, tels que les catastrophes naturelles majeures. Les problématiques associées à l'assurabilité des risques extrêmes sont donc complexes. L'étude, Lopez (2023), offre une bonne description des problématiques liées aux risques extrêmes. Elle examine les enjeux actuels et les questions clés entourant la couverture de ces risques par les assureurs.

Dans le contexte des risques extrêmes, l'intervention de l'État peut être envisagée comme une solution pour soutenir le secteur et faire face aux pertes importantes qui pourraient dépasser les capacités des assureurs. La cas le plus connu est le régime CatNat, qui sera présenté dans la section suivante.

1.3.2 Public privé : Régime CatNat

En France, la gestion des catastrophes naturelles repose sur un partenariat public-privé appelé le régime CatNat. Il repose sur une garantie, ce n'est pas une assurance obligatoire mais une extension de garantie obligatoire à tout contrat d'assurance dommages couvrant un bien. Cette spécificité française a une influence considérable sur la gestion des sinistres liés aux catastrophes naturelles. Le régime d'indemnisation des catastrophes naturelles a été établi par la loi du 13 juillet 1982, *LOI numéro 82-600 du 13 juillet 1982 relative à l'indemnisation des victimes de catastrophes naturelles* (1982) et repose sur le principe de solidarité : pour chaque contrat, un taux de surprime d'assurance identique, fixé par le gouvernement, est utilisé pour compenser les pertes dues aux catastrophes naturelles. Il s'agit d'un partenariat public-privé entre l'État et les compagnies d'assurance, où les assureurs sont chargés de la gestion des sinistres tandis que l'État réglemente les caractéristiques clés du contrat, tel que :

— la définition des risques couverts,

— la tarification,
— la franchise,
— le processus de reconnaissance de l'état de catastrophe naturelle.

L'État apporte sa garantie illimitée au régime CatNat par le biais d'une entreprise détenue à 100% par l'Etat, la Caisse Centrale de Réassurance (CCR). En pratique, la grande majorité des assureurs se réassurent auprès de la CCR, ce qui en fait un acteur clé et incontournable dans la gestion des risques naturels. Dans le cadre de ce régime, la cotisation correspond à un taux uniforme de surprime sur tout le territoire, établi par les pouvoirs publics. Ce taux s'élève à 12% pour les assurances de dommages aux biens des particuliers et des professionnels, et à 6% pour les garanties vol et incendie d'un véhicule terrestre à moteur (ou 0,5% pour la garantie dommages en cas d'absence de vol et incendie). Pour une habitation, la prime CatNat moyenne s'élève à 26 euros, Chneiweiss & Bardaji (2020). Il convient de souligner que cette prime est basée sur le principe de solidarité nationale, chaque individu paie le même taux, quel que soit son niveau d'exposition aux risques. Ainsi, il n'y a pas de modulation de la prime en fonction de l'exposition individuelle, du type de bien ou de sa construction.

De plus, une retenue de 12% est prélevée sur cette prime pour alimenter le Fonds national de prévention des risques naturels majeurs (FPRNM), également connu sous le nom de "fonds Barnier". Ce fonds joue un rôle dans la prévention en finançant des études relatives aux Plans de Prévention des Risques (PPR) et aux Programmes d'Actions de Prévention des Inondations (PAPI), ainsi qu'en soutenant des activités d'information préventive. En outre, il peut être utilisé pour financer des mesures d'aménagement de certains quartiers ou le relogement des populations les plus exposées. L'État a aussi établi une franchise obligatoire et non rachetable pour les sinistres relevant du régime CatNat. Pour les habitations et les véhicules, cette franchise est fixée à 380 euros pour tous les types de sinistres, à l'exception de la sécheresse où elle s'élève à 1520 euros. Pour les biens à usage professionnel, la franchise correspond à 10% des dommages directs, avec un montant minimum de 1140 euros. En ce qui concerne les pertes d'exploitation, la franchise est de 3 jours ouvrés. Cette franchise permet d'éviter un certain nombre de sinistres dit de "fréquence" et de renforcer la prévention des petits dommages facilement évitables.

Pour être indemnisé en cas de sinistre relevant du régime CatNat, il est nécessaire que la commune soit reconnue en état de catastrophe naturelle par un arrêté du gouvernement publié au Journal Officiel. L'assuré doit signaler le sinistre à sa mairie, qui transmettra ensuite une demande à la préfecture. Les demandes sont ensuite centralisées par la préfecture et transmises à la Direction Générale de la Sécurité Civile et de la Gestion des Crises (DGSCGC) pour instruction. Une commission interministérielle analyse ensuite les demandes et rend un avis. Ce processus est complexe comme illustré dans la figure 1.5.

La décision de reconnaître une commune en état de catastrophe naturelle est prise par une commission interministérielle qui évalue le caractère exceptionnel de l'événement naturel au niveau de la commune. Pour les inondations, la décision est basée sur la période de retour de l'événement, c'est-à-dire la fréquence à laquelle un événement similaire peut se produire.

Le 28 décembre 2021, une loi a été promulguée afin d'améliorer l'indemnisation des catastrophes naturelles. Cette loi a pour objectif d'accroître la transparence du processus décisionnel de reconnaissance de l'état de catastrophe naturelle à l'égard des maires et des sinistrés. La commission interministérielle chargée de cette reconnaissance, mentionnée précédemment, est désormais officiellement établie par la loi. De plus, les délais pour déclarer un sinistre et obtenir une réparation ont été modifiés. Enfin, les frais de relogement d'urgence des sinistrés de catastrophes naturelles seront désormais inclus dans l'indemnisation prévue.

Le régime CatNat en France offre un niveau d'assurance satisfaisant contre les catastrophes

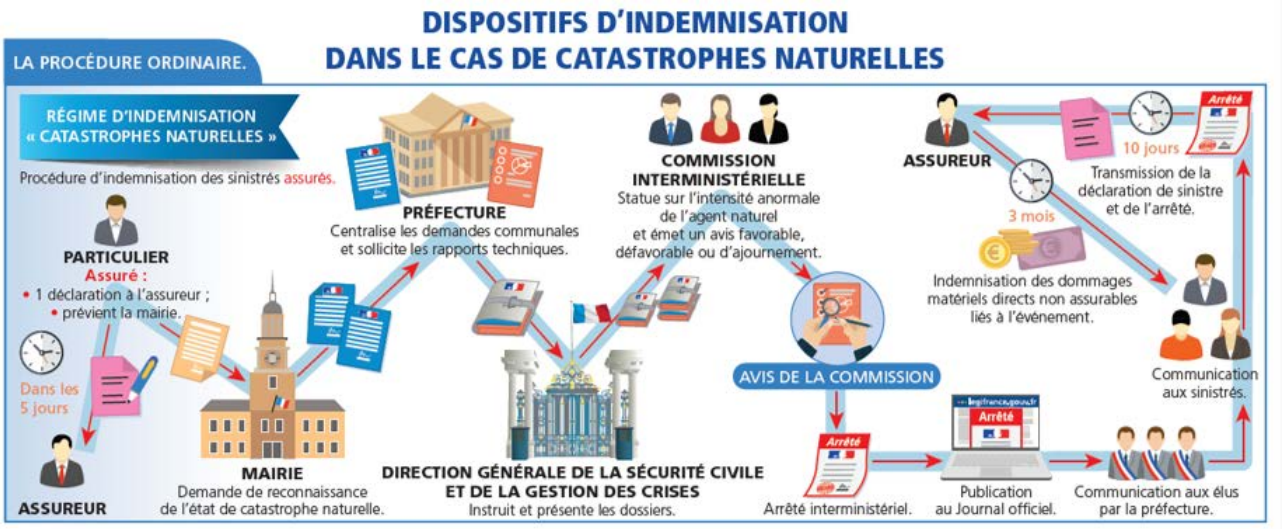


Figure 1.5 – Processus d'indemnisation dans le cadre du régime CatNat (Source: DGSCGC)

naturelles, même s'il peut être critiqué pour son caractère déresponsabilisant, comme mentionné par, Gérin (2011), Latruffe & Picard (2005). En effet, dans ce système basé sur la solidarité, les assurés ne paient pas le prix réel de leurs risques et sont peu incités à prendre des mesures de prévention. La récente réforme ne propose pas de solution à ce problème et envisage de l'aborder dans un rapport distinct, notamment pour la question de la sécheresse. La problématique de la prévention et de l'assurance des catastrophes naturelles est complexe, comme l'explique la thèse de Goussebaile (2016). Ces travaux recommandent que les politiques publiques s'emploient à sensibiliser les individus aux risques et à les rendre responsables de leurs choix en matière d'exposition. Cependant, dans le cadre du régime CatNat et de la solidarité nationale, cette approche n'est pas privilégiée.

Dans le rapport sur le risque cyber intitulé "Lucy" *"LUCY : LUmière sur la CYberassuranc* (2022), l'idée d'explorer un système similaire au régime CatNat est évoquée afin de mieux couvrir ce risque et combler les lacunes de l'assurance traditionnelle. Toutefois, selon nos connaissances actuelles, un tel système n'est pas à l'étude.

1.3.3 Atténuation par la prévention

La prévention joue un rôle clé dans l'assurance des risques extrêmes. En atténuant les conséquences des événements catastrophiques, la prévention permet de minimiser les pertes potentielles et d'améliorer la résilience face à ces risques. Les mesures de prévention peuvent inclure la mise en place de mesures de sécurité, l'adoption de bonnes pratiques, la sensibilisation du public, la mise en œuvre de réglementations et de normes de construction, ainsi que la planification et la gestion des risques. Pour les assureurs, la prévention est essentielle car elle contribue à maintenir la viabilité financière du système d'assurance. En réduisant la gravité des sinistres, la prévention permet de limiter les coûts de l'indemnisation et de maintenir les primes d'assurance à des niveaux plus abordables. Elle favorise également une meilleure gestion des risques à long terme. De plus, la prévention est bénéfique pour les assurés, car elle réduit leur exposition aux risques et les protège des pertes financières potentielles. Elle encourage également la responsabilisation des individus et des entreprises en les incitant à prendre des mesures proactives pour se protéger contre les risques.

Dans le cadre du régime CatNat, la prévention est mise en place à travers le FPRNM (Fonds de

Prévention des Risques Naturels majeurs). Selon le rapport, *Rapport au Ministre de l'économie, des finances et de la souveraineté industrielle et numérique sur le régime d'indemnisation des catastrophes naturelles* (2022), au cours de la dernière décennie, le FPRNM a financé en moyenne chaque année 770 opérations pour un montant total de 175 millions d'euros sur l'ensemble du territoire français et une augmentation du budget du FPRNM de 10 millions d'euros dès aujourd'hui se traduirait, toutes choses étant égales par ailleurs, par une baisse de la sinistralité moyenne annuelle de 13 millions d'euros d'ici 2050. Pour que la prévention soit pleinement efficace et puisse contenir l'augmentation prévisible de la sinistralité d'ici 2050, il serait ainsi nécessaire, tout en maintenant la réglementation en vigueur :

— d'augmenter de manière significative l'enveloppe annuelle du FPRNM ;
— de renforcer la prévention des risques insuffisamment traités jusqu'à présent, notamment les inondations par ruissellement, la sécheresse et les vents cycloniques ;
— de poursuivre et d'intensifier la réflexion sur l'intégration progressive du changement climatique dans les politiques publiques de prévention des risques naturels, afin de prendre en compte l'évolution des aléas en termes de fréquence et d'intensité.

Ce rapport confirme l'importance de la prévention et la nécessité de renforcer les moyens alloués, notamment à travers une augmentation significative du budget du FPRNM pour le régime CatNat, afin de faire face aux enjeux futurs et de mieux anticiper les conséquences du changement climatique.

La prévention joue aussi un rôle crucial dans l'atténuation des risques cyber. Il est essentiel de mettre en place des mesures préventives pour réduire les vulnérabilités et renforcer la résilience des systèmes informatiques. Cela comprend l'adoption de bonnes pratiques de sécurité telles que l'utilisation de mots de passe forts, la mise à jour régulière des logiciels, l'installation de pare-feu et de solutions de sécurité informatique, ainsi que la sensibilisation des utilisateurs aux risques cyber. Ces mesures préventives sont essentielles pour protéger les systèmes et les données sensibles contre les attaques. De plus, les entreprises peuvent mettre en place des stratégies de gestion des risques en matière de cybersécurité. Cela peut inclure la formation et la sensibilisation du personnel à la sécurité informatique, l'adoption de politiques de sécurité strictes et la mise en place de plans de réponse aux incidents. Une approche proactive de la sécurité permet de prévenir les attaques avant qu'elles ne se produisent et de réagir rapidement en cas d'incident. Il est important de souligner que de nombreux contrats d'assurance exigent déjà la mise en place de certaines de ces mesures. Les entreprises peuvent être tenues de respecter certaines normes de sécurité et de se conformer à des exigences spécifiques pour bénéficier d'une couverture d'assurance. Pour soutenir les investissements en matière de prévention, notamment pour les petites et moyennes entreprises (PME), des dispositifs tels que le suramortissement comptable ou les crédits d'impôt pourraient être envisagés, comme préconisé dans *Climat, cyber, pandémie : le modèle assurantiel mis au défi des risques systémiques* (2022). Ces incitations financières encourageraient les entreprises à investir dans des mesures de sécurité informatique et à renforcer leur résilience face aux cyberattaques.

1.3.4 Paramétrique

L'assurance paramétrique est une approche qui permet de simplifier la gestion des risques dans des situations où l'évaluation précise des pertes peut être complexe, Lin & Kwon (2020). Cette approche a été principalement développée pour les catastrophes naturelles, comme en témoignent les recherches menées par Van Nostrand & Nevius (2011) et Horton (2018). De nombreux mémoires d'actuariat ont également été rédigés sur ce sujet. Le concept de l'assurance paramétrique repose sur l'utilisation d'un "paramètre" qui est facilement mesurable et qui est lié aux pertes subies. Par exemple, lorsqu'une région est touchée par un ouragan, il peut être difficile d'évaluer précisément les dommages causés, ce qui entraîne des coûts et des retards liés à l'expertise nécessaire pour déterminer les indemnités. L'assurance paramétrique propose une approche alternative en utilisant des paramètres tels que le type

de cyclone, la vitesse du vent, le niveau des précipitations ou d'autres indices basés sur des variables physiques pertinentes.

Les caractéristiques attrayantes de l'assurance paramétrique sont :

— Simplification de la gestion des sinistres: Le paramètre utilisé étant facilement disponible peu de temps après l'événement, le montant de l'indemnisation peut être déterminé rapidement, évitant ainsi les délais liés à des expertises supplémentaires. Cela permet à la victime de recevoir rapidement une compensation qu'elle peut utiliser pour commencer les réparations après avoir déposé sa demande.

— Simplification du modèle économique de l'assureur : En se concentrant sur la modélisation de l'évolution du paramètre choisi plutôt que sur les conséquences directes du risque, l'assureur peut mieux évaluer et contrôler les exigences de solvabilité. Cela facilite la gestion des risques et la prévision des ressources nécessaires pour faire face aux sinistres.

Cependant, il est important de souligner que ces simplifications ont leurs limites. Il n'est pas réaliste de s'attendre à ce que le paramètre utilisé reflète exactement le risque sous-jacent, surtout dans le cas de risques volatils tels que ceux qui nous intéressent ici. Les solutions paramétriques sont prometteuses, mais elles ne peuvent pas à elles seules garantir une couverture complète du risque. Il est de plus essentiel de trouver un paramètre approprié, ce qui est souvent une tâche très complexe. La relation entre le paramètre choisi et le risque doit être établie scientifiquement et surveillée dans le temps, tout en permettant de prédire avec précision les coûts associés au risque. Une sélection rigoureuse du paramètre est donc nécessaire pour assurer la pertinence et l'efficacité de l'assurance paramétrique.

Par exemple dans le cadre de notre étude, après une phase vérifiant la pertinence, nous pourrions considérer les paramètres suivants :

— Pour le risque cyber : le nombre de données exposés par l'incident.

— Pour le risque d'inondation : hauteur d'eau lors de l'événement ou cumul des précipitations.

Dans le cas des risques extrêmes, l'assurance paramétrique présente un autre inconvénient majeur. Les produits d'assurance paramétrique ont été conçus pour offrir une couverture adéquate en moyenne, mais ils ne prennent pas nécessairement en compte les situations extrêmes à l'échelle individuelle. On utilise généralement le terme de "reste à charge" pour décrire la différence entre les pertes réelles causées par le risque de base et l'indemnisation perçue par l'assuré. Dans les cas typiques correspondant au scénario moyen, le reste à charge qui peut être positif ou négatif, tend à être relativement faible par rapport à la perte réelle. Cela signifie que l'assuré peut être soit sur-indemnisé, soit déçu, mais dans une mesure relativement limitée car la majeure partie de la perte est couverte. Cependant, selon Lopez & Thomas (2023), lorsque les conséquences d'un sinistre sont particulièrement graves, le reste à charge a tendance à être systématiquement inférieur à la perte réelle. Cette caractéristique structurelle est due à une volatilité plus faible du paramètre par rapport au risque de base. Cette volatilité réduite signifie qu'il n'y a pas de provision prévue pour les sinistres exceptionnels, ce qui entraîne une couverture insuffisante dans ces situations de risques extrêmes.

L'assurance paramétrique peut donc être un outil utiles pour la gestion des risques naturels et cyber. Cependant, il est important de retenir qu'elle peut entraîner une sous-évaluation des pertes dans le cas d'événements extrêmes. La tarification précise de ces risques constitue un défi majeur. Il est essentiel de trouver un équilibre entre la simplification apportée par l'assurance paramétrique et la nécessité d'une évaluation adéquate des pertes réelles. Une tarification correcte des risques extrêmes permettrait d'assurer une couverture appropriée et de réduire les écarts entre les pertes subies et les indemnités versées. En mettant en place des méthodes rigoureuses de tarification, il est possible de maximiser les avantages de l'assurance paramétrique tout en minimisant les risques de sous-évaluation.

1.4 La difficile tarification des risques extrêmes

1.4.1 Introduction

Comme mentionné à plusieurs reprises dans ce chapitre, la tarification constitue un enjeu majeur dans le cas des risques extrêmes. Le régime CatNat, illustre parfaitement cette problématique. Il a été créé en raison de l'incapacité à évaluer le coût des catastrophes naturelles. En 1982, les catastrophes naturelles n'étaient pas clairement définis et leurs dommages potentiels n'étaient pas quantifiés. Ils étaient considérés comme "non assurables", relevant de l'incertitude non mesurable plutôt que de risques mesurables. Le régime CatNat a donc été instauré afin de pallier cette absence de couverture d'assurance, comme expliqué par Bidan & Cohignac (2017). Une modification apportée en 1992 confirme que ce régime couvre les événements considérés comme "non assurables", tels que les dommages statistiquement inconnus ou systémiques, qui ne peuvent être couverts de manière temporaire ou durable par les produits d'assurance traditionnels. La liste implicite des événements couverts par ce régime a évolué au fil du temps, en fonction des progrès réalisés dans le domaine de l'assurance et des avancées technologiques. Par exemple, une couverture contre les tempêtes a été développée et est devenue obligatoire dans les polices d'assurance habitation courantes à partir de 1990, ce qui a exclut ce risque de la couverture du régime CatNat. En revanche, les dommages causés par la sécheresse ont été ajoutés à la liste informelle des risques couverts en 1989. Ce lien entre l'équilibre entre l'équité actuarielle et la solidarité est étudié dans l'article de Charpentier et al. (2021). Cet article montre que les progrès actuels permettent une tarification plus précise des risques, notamment des inondations, et ouvrent la possibilité de segmentation. Cependant, cette approche n'est pas mise en œuvre car elle remet en cause le principe de solidarité du régime CatNat. Le même problème se pose pour le risque cyber. Pour assurer ce risque, il est nécessaire de le tarifer de manière adéquate. Cependant, comme nous le verrons dans la section suivante, cette tâche est complexe et difficile

1.4.2 Les limites des modèles classiques

Comme évoqué, dans notre démarche de tarification, nous nous concentrons sur l'étude de événements extrêmes, donc sur l'étude de la queue de distribution, en particulier des queues lourdes. Notre objectif est de comprendre les facteurs qui influencent cette partie de la distribution et d'identifier des classes de risque permettant une gestion différenciée. Cependant, cette tâche présente deux problèmes majeurs :

- Complexité de l'étude des valeurs extrêmes : L'analyse des événements extrêmes est souvent complexe en raison de la nature volatile des coûts associés. Les risques extrêmes peuvent entraîner des pertes considérables et imprévisibles, ce qui rend la modélisation et la tarification de ces événements difficiles.
- Manque de données fiables : Les événements extrêmes sont par nature rares, ce qui limite la disponibilité de données historique. Cette rareté rend l'estimation des probabilités et des coûts associés encore plus difficile. Il est donc crucial de trouver des méthodes alternatives et innovantes pour combler ce manque de données et améliorer la tarification des risques extrêmes.

Les modèles de tarification classiques, tel que le modèle linéaire généralisé, présentent donc des limites lorsqu'il s'agit de tarifer les risques extrêmes. Dans ce modèle, la différenciation des prix est principalement basée sur le scénario central, ce qui n'est pas adapté pour l'étude des distribution à queue lourde. De plus, le tarif d'expérience, qui consiste à examiner l'historique des sinistres d'un client et à lui attribuer une prime en fonction de ses sinistres passés et lui aussi inadapté. En effet, cette technique ne peut être utilisée que si le client présente une sinistralité stable permettant une tarification individuelle précise, ce qui n'est la cas ici. Dans le cas des risques extrêmes, où les

événements sont rares et les coûts associés sont souvent volatils, il devient difficile d'estimer les primes individuelles en se basant uniquement sur l'historique des sinistres passés. La tarification des risques extrêmes demande donc une adaptation des modèles de tarification classiques afin de tenir compte de la spécificité de ces risques et de développer des méthodologies plus robustes pour évaluer les primes.

C'est tout l'intérêt de notre méthode qui lie la théorie des valeurs extrêmes, qui est un cadre de référence pour l'étude des événements extrêmes, à la théorie de la crédibilité afin de pallier le manque de données. L'inconvénient majeur des méthodes bayésiennes est qu'il est souvent difficile de trouver le bon prior. C'est pourquoi notre méthode s'appuie également sur l'utilisation d'arbres de régression pour estimer le meilleur prior en fonction de covariables, parmi des distributions à queue lourde.

Chapter 2

La tarification basée sur la théorie des valeurs extrêmes et la crédibilité bayésienne

2.1 Théorie des valeurs extrêmes

2.1.1 Introduction

La théorie des valeurs extrêmes cherche à analyser la queue de distribution de variables aléatoires, l'objectif est de quantifier les scénarios extrêmes, pour lesquels la valeur de ces variables aléatoires est élevée par rapport aux valeurs typiques. D'un point de vue statistique, les événements climatiques, notamment les catastrophes naturelles, sont souvent associés à des événements de ce genre. Lorsqu'ils surviennent, ces événements peuvent prendre des valeurs très faibles ou très élevées et entraîner des conséquences considérables. De même, cela s'applique à certains événements liés au Cyber, comme cela a été mentionné précédemment dans l'introduction. L'étude de ces événements est donc centrale pour la gestion des risques et vise à résoudre des problèmes d'inférence en dehors de l'échantillon observé. Comment pouvons-nous estimer la probabilité d'occurrence ou l'ampleur d'un événement lorsque celui-ci n'a pas été observé ? La théorie des valeurs extrêmes offre un cadre statistique permettant de résoudre ces problèmes. Historiquement, sa création remonte aux travaux de Fréchet (1927), Fisher & Tippett (1928), Gnedenko (1943), Gumbel (1958). Ces chercheurs ont identifié les lois limites qui décrivent le comportement des données extrêmes, c'est-à-dire celles qui dépassent un certain seuil, sous certaines hypothèses. Aujourd'hui, les domaines d'application sont nombreux et notamment pour l'étude des événements naturels Bousquet & Bernardara (2021), elle est aussi particulièrement utilisée en hydrologie, Katz et al. (2002), Guillou & Willems (2006), Smith (1987), ou en actuariat Brodin & Rootzén (2009), Embrechts et al. (2013), Resnick (1997), Rootzén & Tajvidi (1997), Farkas, Lopez & Thomas (2021).

La théorie des valeurs extrêmes s'avère donc particulièrement appropriée pour la tarification des risques extrêmes. Dans la section suivante, nous présenterons les notations et les outils de cette théorie qui seront utilisés tout au long de ce mémoire.

2.1.2 Méthode Peaks over threshold

Dans le cadre de cette méthode, nous adoptons l'approche des dépassements de seuil, également connue sous le nom de Peaks-Over-Threshold (PoT), dont le résultat fondamental a été établi par Balkema & De Haan (1974). Cette approche se base sur l'utilisation des observations qui ont dépassé un seuil spécifique.

Considérons des variables aléatoires Y_1, Y_2, ..., Y_n, indépendantes et identiquement distribuées (i.i.d), avec une fonction de répartition inconnue F. Dans la suite, nous noterons $\overline{F}$ la fonction de survie associée, définie comme $\overline{F}(y) = \mathbb{P}(Y_i > y)$ pour tout y.

Dans l'approche PoT (Peaks-Over-Threshold), une observation est considérée comme extrême si elle dépasse un seuil préalablement choisi, noté u (le choix de ce seuil sera discuté ultérieurement). Lorsqu'une observation est jugée extrême, on définit l'excès correspondant comme la différence entre cette observation et le seuil u. La figure 2.1 illustre cette méthode. La ligne rouge représente le seuil u, tandis que les points rouges correspondent aux observations extrêmes, c'est-à-dire celles qui ont dépassé le seuil u.

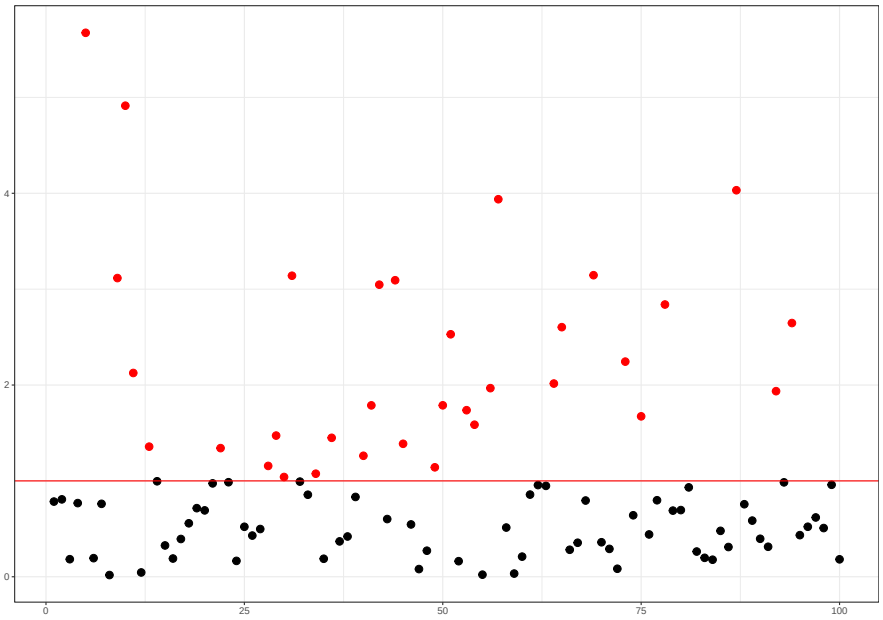


Figure 2.1 – Illustration de la méthode « Peaks-over-Threshold ». La ligne rouge représente le seuil u et les points rouges les observations extrêmes.

La loi des excès s'obtient facilement à partir de la fonction de répartition F :

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

En 1975, Pickands III démontre que si $\overline{F}$ satisfait la propriété suivante :

$$\lim_{t \to +\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0}, \forall y > 0,$$

avec $\gamma_0 > 0$, alors

$$\lim_{u \to +\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}_{\sigma_{0u},\gamma_0}(z)| = 0,$$

où $\sigma_{0u} > 0$ et $\overline{H}_{\sigma_{0u},\gamma_0}$ est la fonction de survie d'une loi non dégénérée qui appartient nécessairement à la famille des lois de Pareto généralisée (GP) avec

$$\overline{H}_{\sigma_{0u},\gamma_0}(z) = \left(1 + \gamma_0 \frac{z}{\sigma_{0u}}\right)^{-1/\gamma_0}, z > 0,$$

σ_{0u} est un paramètre d'échelle et $\gamma_0 > 0$ un paramètre de forme, appelé indice de queue, reflétant l'épaisseur de la queue de F. Plus γ_0 est grand, plus la queue de distribution est lourde. On peut noter que si $\gamma_0 \in]0\,;1[$ alors l'espérance est finie, alors que si $\gamma_0 > 1$ l'espérance de Y_i est infinie, Coles et al. (2001).

En pratique, une fois le seuil u choisi, une loi GP est ajustée aux excès associés au seuil. L'estimation des paramètres σ_{0u} et γ_0 peut être faite par maximum de vraisemblance.

Le choix du seuil u est une tâche complexe qui implique un compromis entre le biais et la variance. Un seuil trop bas affaiblirait les estimations asymptotiques et introduirait des biais dans l'analyse, tandis qu'un seuil trop élevé limiterait le nombre de données au-dessus du seuil, entraînant une variance élevée. Le choix optimal de ce seuil dépend de paramètres inconnus dans la pratique. Actuellement, les méthodes disponibles sont principalement graphiques Davison & Smith (1990), Coles et al. (2001). À notre connaissance, il n'existe pas de méthode automatique largement acceptée. Il est donc possible de se baser sur l'expertise métier pour sélectionner un seuil, tout en veillant à ce qu'il soit raisonnablement adapté à une loi de valeurs extrêmes pour les excès observés.

L'un des principaux avantages de la méthode PoT réside dans sa capacité à analyser le comportement des événements situés dans la queue de la distribution, c'est-à-dire ceux qui dépassent un seuil spécifique. Dans le domaine des risques naturels, il est souvent constaté que la grande majorité des dommages enregistrés provient d'un petit nombre d'événements extrêmement intenses. Ainsi, la théorie des valeurs extrêmes, offre une approche permettant d'étudier ces événements et d'en comprendre les caractéristiques. En se concentrant sur les observations qui dépassent le seuil prédéfini, cette méthode permet d'obtenir des informations précieuses sur le comportement de ces événements. Cela s'avère essentiel pour évaluer les risques associés aux catastrophes naturelles et pour prendre des décisions éclairées en matière de gestion des risques. Cela s'applique également au risque cyber, qui est susceptible de présenter des événements extrêmes.

Nous introduisons dans la partie suivante les notations de la théorie de la crédibilité, qui permettent de tenir compte du faible historique de données sur les événements extrêmes.

2.2 Théorie de la crédibilité bayésienne

La théorie de la crédibilité est une méthode largement utilisés en assurance pour la tarification des primes. Elle permet de combiner les informations provenant de deux sources principales :

— l'expérience passée de l'assuré et,

— l'information collective.

L'objectif est d'estimer de manière optimale la prime d'assurance en prenant en compte à la fois les données individuelles de l'assuré et les informations générales disponibles. La crédibilité est particulièrement indiquée lorsque les données disponibles sur un individu sont insuffisantes pour obtenir une estimation précise de son risque. Dans le cas des événements extrêmes, les données

peuvent être rares ou inexistantes, rendant difficile l'estimation du niveau de risque associé. Dans un tel cas, la crédibilité permet de combler ce manque en utilisant à la fois l'expérience individuelle et l'information agrégée sur le risque. La crédibilité repose sur le principe que l'information individuelle est essentiel et doit être prise en compte dans l'estimation de la prime, quand elle est disponible, mais elle est aussi pondérée par l'information agrégée, plus robuste. Cette pondération est déterminée par un paramètre appelé le coefficient de crédibilité, qui mesure l'influence relative de l'expérience individuelle et de l'information collective.

Une description détaillée de la théorie de la crédibilité peut être trouvée dans le cours de référence Bühlmann & Gisler (2005) et dans l'article de Heilmann Heilmann (1989). Cette méthode tire ses origines des travaux Mowbray (1914). Aujourd'hui, elle est largement reconnue comme une méthode de référence pour la tarification des primes, permettant de prendre en compte à la fois l'expérience individuelle et le profil de risque. Nous présentons ici une brève introduction des résultats et notations utilisés dans la théorie de la crédibilité, qui sont essentiels pour notre analyse.

Notons :
— I, classiquement le nombre d'assurés ;
— n_i, le nombre d'années d'observations ;
— $Y_{i,j}$, le montant des sinistres pour l'événement j pour l'assuré i ;
— $\mathbb{Y} = (Y_{i,j})_{j=1,....n_i,i=1.....I}$;
— $\theta_i \in \Theta$, le profil de risque de l'assuré.

En adoptant la terminologie bayésienne, nous considérons que θ_i, qui représente le profil de risque individuel, est une variable aléatoire distribuée selon une loi a priori notée $\mathbb{T}$. Les variables aléatoires $Y_{i,j}$, conditionnellement à θ_i, sont des observations i.i.d. suivant une loi F_{θ_i}. La loi conditionnelle de θ_i sachant les observations $Y_{i,j}$ est appelée la loi a posteriori. La prime individuelle π_i est déterminée comme l'espérance de la loi conditionnelle des observations $Y_{i,j}$ sachant θ_i. La prime collective, quant à elle, correspond à l'espérance des observations $Y_{i,j}$ sur l'ensemble des assurés. Elle représente une estimation globale des primes d'assurance pour la population assurée dans son ensemble. Enfin, la prime de crédibilité est calculée comme l'espérance conditionnelle de la variable aléatoire $Y_{i,n+1}$ sachant les observations $Y_{i,1}, \ldots, Y_{i,n}$. Cette prime tient compte des observations passées d'un individu pour estimer sa prime future, en accordant plus de poids à ses propres observations tout en considérant les informations agrégées de la population.

$$\pi_{ind,i} = \mathbb{E}[Y_{i,n_i+1} \mid \theta_i]$$
$$\pi_{col} = \mathbb{E}[\mathbb{Y}]$$
$$\pi_{cred,i} = \mathbb{E}[Y_{i,n_i+1} \mid Y_{i,1}, ...Y_{i,n_i}]$$

Considérons maintenant un assuré i avec $(Y_{i,1}, ..., Y_{i,n_i})$ le montant des sinistres pour l'événement. Soit $g = g_t$ la densité d'une loi de paramètre t. On a $\theta_i \sim \mathbb{T}$ avec $\mathbb{T}$ la loi a piori, une loi sur le paramètre de la loi de g. Les $(Y_{i,j})_j$ sont conditionnellement à $\theta_i = t$, i.i.d de loi g_t.

Pour trouver $\pi_{cred,i}$ la prime de crédibilité, on doit connaître alors la loi a posteriori de θ_i soit la

loi de θ_i sachant les Y_{ij}.

$$\begin{aligned}\pi_{cred,i} &= \mathbb{E}[Y_{i,n_{i+1}} \mid Y_{i,1}, ...Y_{i,n_i}] \\ &= \int_y y f_{Y_{i,n_i+1}|Y_{i,1},....Y_{i,n_i}}(y)\mathrm{d}y_{n+1} \\ &= \int_y y \int_t f_{Y_{i,n_i+1},\theta_i|Y_{i,1},....Y_{i,n_i}}(y,t)\mathrm{d}t\mathrm{d}y \\ &= \int_y \int_t y g_t(y) f_{\theta_i|Y_{i,1}...Y_{i,n_i}}\mathrm{d}t\mathrm{d}y.\end{aligned}$$

On peut remarquer que

$$\pi_{cred,i} = \int_t \mathbb{E}[Y_{i,n_{i+1}}|\theta_i = t] f_{\theta_i|Y_{i,1}...Y_{i,n_i}}(t)\mathrm{d}t.$$

Par définition des densités conditionnelles, on a

$$f_{\theta_i|Y_{i,1}=y_1,...,Y_{i,n_i}=y_{n_i}}(t) = \frac{f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1,....,y_{n_i},t)}{f_{Y_{i,1},....,Y_{i,n_i}}(y_1,....,y_{n_i})}.$$

On a pour le numérateur

$$\begin{aligned}f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1,....,y_{n_i},t) &= f_{Y_{i,1},....,Y_{i,n_i}|\theta_i=t}(y_1,....,y_{n_i}) f_{\theta_i}(t) \\ &= g_t(y_1)....g_t(y_{n_i}) f_{\theta_i(t)},\end{aligned}$$

car les $(Y_{i,j})_j$ sont i.i.d de loi g_t conditionnellement à $\theta_i = t$ et pour le dénominateur,

$$\begin{aligned}f_{Y_{i,1},....,Y_{i,n_i}}(y_1,....,y_{n_i}) &= \int_t f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1,....,y_{n_i},t)\mathrm{d}t \\ &= \int_t g_t(y_1)....g_t(y_{n_i}) f_{\theta_i(t)}\mathrm{d}t,\end{aligned}$$

d'où

$$f_{\theta_i|Y_{i,1}=y_1,...Y_{i,n_i}=y_{n_i}}(t) = \frac{g_t(y_1)....g_t(y_{ni}) f_{\theta_i}(t)}{\int_s g_s(y_1)....g_s(y_{n_i}) f_{\theta_i}(s)\mathrm{d}s}.$$

Ici, $\int_s g_s(y_1)\dots g_s(y_{n_i}) f_{\theta_i}(s)\mathrm{d}s$ joue le rôle d'une constante de normalisation. Cela garantit que $g_t(y_1)\dots g_t(y_{n_i}) f_{\theta_i}(t)$ constitue une densité de probabilité. En pratique, nous calculons le numérateur et tentons de reconnaître une loi usuelle afin d'éviter le calcul de l'intégrale.

L'estimation de la loi a priori constitue l'une des difficultés majeures. La méthode que nous proposons permet justement d'estimer cette loi en utilisant des arbres de régression. Nous présentons tout d'abord dans la partie suivante comment la théorie de la crédibilité et les valeurs extrêmes peuvent se lier pour faire de la tarification.

2.3 Loi de Pareto généralisée comme un mélange de variables aléatoires exponentielles

Comme nous l'avons vu précédemment, dans la théorie de la crédibilité, un assuré est associé à un facteur de risque θ qui suit une distribution a priori que nous appellerons p. Dans cette formulation,

nous nous concentrons sur un assuré spécifique, mais nous verrons par la suite comment utiliser des variables propres à chaque événement pour un assuré donné. Dans le cadre le plus simple, les pertes individuelles subies par cet assuré sont supposées être indépendantes et identiquement distribuées, notées $(Y_1, ..., Y_n)$ conditionnellement à $\theta = t$, avec g_t représentant leur densité. Sur la base des résultats mentionnés précédemment, une approche de crédibilité adaptée au contexte des risques extrêmes doit satisfaire la condition selon laquelle $\int g_t(y)p_i(t)dt$ correspond à la densité d'une distribution de Pareto généralisée. Cette condition garantit que la prime de crédibilité reflète adéquatement les caractéristiques des événements extrêmes et leur distribution. Ainsi, en choisissant correctement la distribution a priori, il est possible de construire une approche de crédibilité qui est cohérente avec la théorie des valeurs extrêmes.

Considérons que le facteur de risque θ est distribué selon la loi Gamma, c'est-à-dire que la distribution prior p est :

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

où Γ est la fonction Gamma, avec $r > 1$ et $\lambda > 0$. La distribution de Pareto généralisée peut être considérée comme un mélange Gamma de variables aléatoires exponentiellement distribuées, dans le sens où, si l'on suppose que $Y|\theta = t$ est exponentiellement distribué avec une moyenne de $1/t$, alors

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

On peut voir que cela correspond à une distribution de Pareto generalisé de paramètres :

$$\gamma = \frac{1}{r},$$

$$\sigma = \frac{\lambda}{r}.$$

En reprenant les calculs de la première section, nous pouvons montrer que la distribution postérieure de θ est une distribution Gamma avec des paramètres $(r + n, \lambda + \sum_{i=1}^n Y_i)$. Ainsi, si l'espérance $E[Y]$ est finie, nous pouvons calculer la prime de crédibilité (pure) à partir de cette distribution postérieure. Il est important de noter que dans le cas où Y est distribué selon une loi de Pareto généralisée, la condition $E[Y] < \infty$ est équivalente à $1/\gamma = r > 1$. Dans ce cas, la prime de crédibilité est donnée par:

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

Ce qui peut être réecrie en :

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

En introduisant le facteur de crédibilité $c_n(r) = \frac{n}{r+n-1}$, on peut observer que si $r \leq 1$, la prime de crédibilité n'est pas définie car l'espérance serait infinie. Cependant, la distribution a posteriori reste valide et peut être utilisée pour d'autres analyses. Par exemple, on peut estimer des quantiles à partir de cette distribution, ce qui peut fournir des indications utiles pour la tarification.

Dans le cas $r > 1$, la prime de crédibilité est linéaire et peut être calculée à partir d'une formule donnée. Le facteur de crédibilité $c_n(r)$ permet en quelque sorte de déterminer si l'on peut faire

confiance aux données historiques de l'assuré pour évaluer correctement le risque. Si n tend à être grand, c'est-à-dire si l'on dispose d'un long historique des sinistres pour un assuré, ce facteur est proche de 1. En revanche, en l'absence d'historique, la prime s'avère être $\frac{\lambda}{r-1}$, qui est l'espérance de $\frac{1}{\theta}$ à partir de la distribution a priori. Cette approche est particulièrement utile dans notre cas, car elle permet de tirer pleinement parti des données individuelles de sinistralité tout en reconnaissant qu'elles peuvent être incomplètes. Dans ces situations, il est important de prendre en compte l'expérience collective pour compléter l'évaluation du risque.

Le choix du modèle bayésien exponentiel/gamma est cohérent avec la nature de la distribution de Y qui est caractérisée par une queue lourde et correspond à une distribution de Pareto généralisée. Ce choix de modèle permet non seulement de prendre en compte cette caractéristique spécifique de la distribution, mais il permet également d'obtenir une formule simple et calculable pour la prime de crédibilité.

Réécrit en termes de γ et σ, le facteur de crédibilité c_n devient $c_n(\gamma) = \frac{n}{\frac{1}{\gamma}+n-1}$. Lorsque γ se rapproche de 1, nous observons que le facteur de crédibilité tend vers 1. Cela signifie que dans cette situation, l'information apportée par la distribution a priori est négligeable et il est plus efficace de se fier à la moyenne empirique des observations. En revanche, lorsque γ se rapproche de zéro, une quantité plus importante d'observations est nécessaire pour accorder une confiance suffisante à la moyenne empirique.

L'un des défis majeurs est d'estimer les paramètres γ et σ qui décrivent au mieux l'expérience collective. Pour cela, nous proposons une méthode basée sur des arbres de régression. Cette méthode nous permet d'utiliser des covariables et de conserver une excellente interprétabilité des résultats. En utilisant les arbres de régression, nous pouvons estimer les valeurs optimales de γ et σ qui correspondent le mieux aux données observées et qui permettent d'obtenir des primes de crédibilité précises et fiables.

Chapter 3

L'estimation du prior avec des covariables

3.1 Introduction

Pour tenir compte des caractéristiques de l'assuré $\mathbf{X} \in \mathbb{R}^d$, qui peuvent influencer son groupe de risque spécifique, nous souhaitons que $\mathbf{X}$ ait un impact sur la distribution a priori utilisée pour déterminer la prime de crédibilité. Nous supposons l'existence de fonctions $\mathbf{x} \to r(\mathbf{x})$ et $\mathbf{x} \to \lambda(\mathbf{x})$ (et par conséquent, de fonctions $\mathbf{x} \to \gamma(\mathbf{x})$ et $\mathbf{x} \to \sigma(\mathbf{x})$) qui décrivent l'hétérogénéité entre les classes d'assurés. Ces fonctions permettent de modéliser la relation entre les caractéristiques de l'assuré $\mathbf{X}$ et les paramètres γ et σ. En utilisant ces fonctions, nous pouvons estimer les paramètres γ et σ en fonction des caractéristiques de chaque assuré. Cela nous permet d'ajuster la prime de crédibilité en tenant compte de l'hétérogénéité entre les classes d'assurés et de mieux refléter les risques individuels.

Pour calibrer l'a priori, nous considérons alors que nous disposons de $(Z_1, \mathbf{X}_1, \cdots, Z_N, \mathbf{X}_N)$, i.i.d. réplications de $(Z_1, \mathbf{X}_1)$. La calibration de l'a priori consiste ainsi à estimer les fonctions de régression $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$, en supposant que $Z_1|\mathbf{X}_1 = \mathbf{x}_1$ est distribué selon une distribution de Pareto généralisée avec des paramètres $(\gamma(\mathbf{x}_1), \sigma(\mathbf{x}_1))$.

Dans ce mémoire, nous présentons une méthode développée par Farkas, Heranval, Lopez & Thomas (2021) qui utilise des arbres de régression pour créer des classes homogènes en termes de distributions extrêmes. Une caractéristique intéressante de cette méthodologie est la création d'un nombre fini de classes de risque, pour lesquelles les valeurs $(\gamma(\mathbf{x}), \sigma(\mathbf{x}))$ sont constantes. Cette approche permet de regrouper les assurés en fonction de leurs caractéristiques communes, facilitant ainsi la tarification et la gestion des risques. Le résultat de cette procédure est qu'en utilisant $\hat{\gamma}(\mathbf{x}), \hat{\sigma}(\mathbf{x})$ pour désigner les estimateurs obtenus,

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

où la multiplication $\times$ s'applique aux deux composantes d'un vecteur (γ, σ), et où $(r_j)_{1 \le j \le K}$ sont les "règles" utilisées pour attribuer un individu à l'une des K classes de risque déterminées par la procédure d'ajustement de l'arbre de régression. Plus précisément, ces fonctions sont telles que $r_j(\mathbf{x}) \in \{0, 1\}$ pour tous j, avec $r_j(\mathbf{x}) r_{j'}(\mathbf{x}) = 0$ pour tous (j, j') avec $j \neq j'$, et $\sum_{j=1}^{K} r_j(\mathbf{x}) = 1$. Cela signifie qu'un assuré ayant des caractéristiques $\mathbf{x}$ est affecté à exactement une classe de risque (et pas plus), sur la seule base de la valeur de ses caractéristiques $\mathbf{x}$. Le nombre de classes K adaptées à l'ensemble de

données est déterminé par la procédure d'estimation elle-même, et n'a pas besoin d'être spécifié.

En fin de compte, cela permet d'obtenir une procédure de tarification plus intelligible, par rapport à une situation où deux individus ayant des caractéristiques différentes $\mathbf{x}$ et $\mathbf{x}'$ seraient affectés à deux valeurs distinctes de la fonction de régression. Nous allons maintenant décrire l'algorithme CART qui est au cœur de notre procédure et expliquer comment il peut être utilisé pour la régression des valeurs extrêmes.

3.2 Arbres de régression GPD

Les arbres de régression, introduits par Breiman et al. (1984), font partie des outils simples et interprétables largement utilisés dans le secteur de l'assurance. Ils présentent de nombreuses propriétés intéressantes, telles que la possibilité d'introduire des non-linéarités tout en produisant un modèle facilement compréhensible. Ces modèles visent à constituer des classes d'observations qui ont un comportement relativement similaire en termes de variable réponse Y. Ces classes sont définies par des "règles" qui affectent une observation à l'une de ces classes en fonction des valeurs de ses covariables $\mathbf{X}$. Ces règles sont obtenues à partir des données grâce à l'algorithme CART (Classification And Regression Tree), et la non-linéarité de la procédure permet de s'adapter à de nombreuses fonctions de régression. Dans le processus d'ajustement des arbres de régression, la phase de croissance est utilisée pour déterminer les règles de division qui permettent de constituer les classes d'observations. Cette étape est décrite en détail dans la section 3.2.1 de notre travail. À partir de l'arbre construit, un estimateur de la fonction de régression θ_0 peut être obtenu, comme expliqué dans la section 3.2.2. Par la suite, l'étape d'élagage, qui agit comme une procédure de sélection de modèle, est présentée dans la section 3.2.3.

3.2.1 Construction de l'arbre maximal

L'algorithme CART consiste à déterminer itérativement un ensemble de règles $\mathbf{x} = (x^{(1)}, \dots, x^{(d)}) \to R_\ell(\mathbf{x})$ pour diviser les données, dans le but d'optimiser une certaine fonction $\theta^(\mathbf{x})$ (également appelée critère de segmentation). Cette fonction $\theta^(\mathbf{x})$ peut être considérée comme l'optimum d'une certaine fonction de risque sur une classe de fonctions cibles, à savoir

$$\theta^*(\mathbf{x}) = \arg\min_{\theta\in\Theta}[\phi(Y,\theta) \mid \mathbf{X} = \mathbf{x}],$$

où $\Theta \subset \mathbb{R}^d$ représente le paramètre spatial et ϕ est une fonction de perte dont le choix dépend de la quantité à estimer. Par exemple, si ϕ est la perte quadratique, alors θ^* correspond à la moyenne conditionnelle de Y étant donné $\mathbf{X}$. Dans notre cas, nous choisissons ϕ comme étant la perte quadratique négative généralisée, c'est-à-dire

$$\phi(z,\theta) = \log(\sigma) + \left(\frac{1}{\gamma} + 1\right)\log\left(1 + \frac{\gamma z}{\sigma}\right), \quad z > 0$$

où $\theta = (\sigma, \gamma)^t \in \Theta$. Ainsi, dans notre cas

$$\theta^*(\mathbf{x}) = \arg\min_{\theta\in\Theta}[\phi(Y-u,\theta)\mathbf{1}_{Y>u} \mid \mathbf{X} = \mathbf{x}].$$

Notons qu'ici, l'algorithme CART ne s'applique qu'aux observations Y_i telles que $Y_i > u$.

Un ensemble de rêgles $(R_\ell)_\ell$ est un ensemble de l'espace telles que, pour tout $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, $R_\ell(\mathbf{x}) = 1$ ou 0 selon certaines conditions, qui sont remplies ou non par $\mathbf{x}$, avec $R_\ell(\mathbf{x})R_{\ell'}(\mathbf{x}) = 0$ pour

$\ell \neq \ell'$ et $\sum_\ell R_\ell(\mathbf{x}) = 1$. Dans le cas des arbres de régression, ces règles de partition ont une structure particulière, puisqu'elles peuvent être écrites, pour les covariables quantitatives, sous la forme suivante $R_\ell(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_1 \leq \mathbf{x} < \mathbf{x}_2}$ pour un $\mathbf{x}_1 \in \mathcal{X}$ et $\mathbf{x}_2 \in \mathcal{X}$, avec des symboles de comparaison à comprendre comme des comparaisons entre composants. En d'autres termes, si $d = 1$, les règles peuvent être identifiées comme des segments de partition, si $d = 2$ ce sont des rectangles (hyper-rectangles dans le cas général). La détermination de ces règles d'une étape à l'autre peut être représentée sous la forme d'un arbre binaire, puisque chaque règle R_ℓ à l'étape k génère deux règles $R_{\ell 1}$ and $R_{\ell 2}$ (avec $R_{\ell 1}(\mathbf{x}) + R_{\ell 2}(\mathbf{x}) = 0$ if $R_\ell(\mathbf{x}) = 0$) à l'étape $k+1$. L'algorithme peut être décrit comme suit :

Étape 1: $R_1(\mathbf{X}_i) = 1$ pour tout $i = 1, \ldots, n$ et $n_1 = 1$, c'est la racine de l'arbre.

Étape k+1: Soit $(R_1, ...R_{n_k})$ les règles obtenues à l'étape k. For $\ell = 1, \ldots, n_k$,

— si toutes les observations i telles que $R_\ell(\mathbf{X}_i) = 1$ ont les mêmes caractéristiques, alors on garde la règle ℓ car il n'est plus possible de faire des partitions des données;

— sinon, la règle R_ℓ est remplacée par deux nouvelles règles $R_{\ell 1}$ et $R_{\ell 2}$ déterminées de la manière suivante : pour chaque composant $X^{(j)}$ de $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, définit le meilleur seuil $x_{\ell\star}^{(j)}$ permettant de séparer les données, de telle sorte que

$$x_{\ell\star}^{(j)} = \arg\min_{x^{(j)}} \left\{ \sum_{i=1}^n \phi(Y_i, \widehat{\theta}_{j-}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i) + \sum_{i=1}^n \phi(Y_i, \widehat{\theta}_{j+}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i) \right\},$$

où

$$\begin{cases} \widehat{\theta}_{j-}(x^{(j)}, R_\ell) & = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \phi(Y_i, \theta) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i), \\ \widehat{\theta}_{j+}(x^{(j)}, R_\ell) & = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \phi(Y_i, \theta) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i). \end{cases}$$

On sélectionne ensuite le meilleur indice pour faire une partition :

$$j_\star = \arg\min_{j} \left\{ \sum_{i=1}^n \phi(Y_i, \widehat{\theta}_{j-}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) + \sum_{i=1}^n \phi(Y_i, \widehat{\theta}_{j+}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \right\}$$

On définit deux nouvelles règles: $R_{\ell 1}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} \leq x_{\ell\star}^{(j_\star)}}$, et $R_{\ell 2}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} > x_{\ell\star}^{(j_\star)}}$.

— Soit $n_{k+1} = n_k + 2$ le nouveau nombre de règles.

Condition d'arrêt : on s'arrête si $n_{k+1} = n_k$.

Cet algorithme a une structure d'arbre binaire. La liste des règles (R_ℓ) est associée aux feuilles de l'arbre à l'étape k, et le nombre de feuilles de l'arbre augmente de l'étape k à l'étape $k+1$. L'algorithme s'arrête lorsque chaque feuille ne contient qu'une seule observation ou lorsque les observations dans la même feuille ont les mêmes caractéristiques. La règle d'arrêt peut également être légèrement modifiée pour s'assurer qu'il y a un nombre minimal de points des données originales dans chaque feuille de l'arbre à chaque étape.

Dans cette version, toutes les covariables sont continues ou binaires $0, 1$. Les variables catégorielles doivent être encodées en variables binaires préalablement. Il est également possible de modifier l'algorithme de manière à ce que les critères de séparation de chaque R_ℓ identifient la modalité de variable catégorielle minimisant la fonction de perte. De plus, la condition d'arrêt peut être modifiée pour garantir un nombre minimal d'observations dans chaque feuille de l'arbre.

3.2.2 De l'arbre à l'estimation des paramètres

A partir d'un ensemble donné de K règles, $\mathcal{R} = (R_\ell)_{\ell=1,\dots,K}$, soit $\mathcal{T}_\ell = \{\mathbf{x} : R_\ell(\mathbf{x}) = 1\}$, la ℓ–ème feuille de l'arbre correspondant. L'estimateur $\widehat{\theta}^K(\mathbf{x})$ associé à l'ensemble des feuilles $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ est obtenue par

$$\widehat{\theta}^K(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}^K(R_\ell) R_\ell(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}_\ell^K \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} = \sum_{\ell=1}^{K} \begin{pmatrix} \widehat{\sigma}_\ell^K \\ \widehat{\gamma}_\ell^K \end{pmatrix} \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

L'arbre obtenu lorsque l'algorithme précédent s'arrête est appelé arbre maximal et est noté $\widehat{T}_{\max}(u)$ avec l'ensemble de feuilles $(\mathcal{T}_\ell)_{\ell=1,\dots,K_{\max}}$, où $K_{\max}$ indique son nombre de feuilles. Il correspond à un estimateur trivial de la fonction objective $\theta^*(\mathbf{x})$ puisque soit le nombre d'observations dans une feuille est égal à un, soit toutes les observations dans cette feuille partagent les mêmes caractéristiques $\mathbf{x}$.

L'étape d'élagage, présentée dans la section suivante, consiste à extraire de l'arbre maximal $\widehat{T}_{\max}(u)$ un sous-arbre qui réalise un compromis entre la simplicité et ajustement.

3.2.3 Sélection d'un sous-arbre : étape d'élagage

Pour l'étape d'élagage, une façon standard de procéder est d'utiliser un critère pénalisé pour sélectionner le sous-arbre approprié de $\widehat{T}_{\max}(u)$, Breiman et al. (1984), Gey & Nedelec (2005). Pour déterminer ce sous-arbre, il n'est pas nécessaire de calculer tous les sous-arbres de $\widehat{T}_{\max}(u)$. Il suffit de déterminer, parmi tous les sous-arbres ayant K feuilles pour $K \leq K_{\max}$, le sous-arbre $\widehat{T}_K(u)$ qui minimise le critère suivant

$$\frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\theta}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i>u} \mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell} + \lambda K, \tag{3.1}$$

où $\lambda > 0$ désigne une constante de pénalisation, qui peut être choisie par validation croisée, Allen (1974), Stone (1974). Rappelons que k_n est le nombre moyen d'observations telles que $Y_i > u$, c'est-à-dire le nombre d'observations sur lesquelles l'algorithme CART est exécutée. Il ne reste plus qu'à déterminer l'arbre final parmi la liste obtenue des $K_{\max}$ sous-arbres admissibles. Les arbres $\widehat{T}_K(u)$, $K = 1, \dots, K_{\max}$, sont faciles à déterminer, puisque $\widehat{T}_K(u)$ est obtenu en enlevant une feuille de l'arbre $\widehat{T}_{K+1}(u)$ Breiman et al. (1984).

Le nombre de feuilles de l'arbre sélectionné est donc obtenu comme l'optimum du critère pénalisé (3.1), c'est-à-dire

$$\widehat{K}(u) = \arg \min_{K=1,\dots,K_{\max}} \left\{ \frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\theta}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i>u} \mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell} + \lambda K \right\},$$

et l'arbre sélectionné est désigné par $\widehat{T}_K(u) = \widehat{T}_{\widehat{K}(u)}(u)$.

À la fin du processus, chaque feuille de l'arbre contient des classes d'événements présentant un comportement de distribution de queue homogène, c'est-à-dire avec les mêmes paramètres $(\sigma_0(x), \gamma_0(x))$ au sein de chaque feuille. Ces paramètres peuvent être utilisés comme a priori dans notre approche de tarification basée sur la crédibilité, ou pour caractériser les distributions extrêmes des événements.

Cette méthode est également accompagnée de résultats théoriques, comme décrit dans Farkas, Heranval, Lopez & Thomas (2021). Nous pouvons démontrer la consistance de cette procédure. Dans un premier temps, nous obtenons des résultats pour la consistance d'un arbre avec K feuilles, puis nous pouvons démontrer la consistance de la procédure d'élagage de l'arbre. Ces résultats reposent sur les inégalités de concentration.

En comparaison avec d'autres méthodes de régression pour les valeurs extrêmes, cette procédure présente l'avantage de pouvoir introduire des discontinuités dans la fonction de régression. Contrairement aux approches paramétriques qui supposent la linéarité, comme décrit dans Beirlant & Goegebeur (2003). Les méthodes les plus flexibles, comme celle présentée dans Beirlant & Goegebeur (2004), reposent sur un lissage des données qui suppose une continuité des covariables. Chavez-Demoulin et al. (2016) propose une approche semi-paramétrique pour séparer les variables continues des variables discrètes.

En synthèse, l'algorithme CART offre une approche robuste et interprétable pour la régression des valeurs extrêmes. En utilisant des arbres de régression, nous pouvons identifier des règles de partition qui regroupent les observations ayant des comportements similaires en termes de variable réponse. Cela permet d'obtenir des classes d'événements avec des propriétés de distribution de queue homogènes, caractérisées par des paramètres constants dans chaque feuille de l'arbre. Ces paramètres peuvent être utilisés dans des problèmes de tarification pour fournir un distribution a priori ou pour analyser les distributions extrêmes des événements. De plus, des résultats théoriques ont été établis pour démontrer la consistance de cette procédure qui comparée à d'autres méthodes de régression des valeurs extrêmes, a l'avantage de pouvoir capturer des discontinuités dans la fonction de régression.

Chapter 4

Données utilisées pour les applications

Nous présentons dans ce chapitre les données utilisées pour appliquer notre méthode. Ce mémoire a été réalisé en grande partie à la Mission Risques Naturels, celle-ci occupe donc naturellement une place plus importante ici. Cependant, en fin de chapitre, nous présentons également une base de données qui peut être utilisée pour appliquer notre méthode au risque cyber.

4.1 Risques Naturels

4.1.1 La Mission Risques Naturels

L'association, Mission des sociétés d'assurances pour la connaissance et la prévention des Risques Naturels, abrégé en Mission Risques Naturels, a été créée en mars 2000, entre la Fédération Française des Sociétés d'Assurances (FFSA) et le Groupement des Entreprises Mutuelles d'Assurance (GEMA), aujourd'hui regroupés dans France Assureurs. Sa création fait suite aux tempêtes dévastatrices Lothar et Martin de 1999, qui sont encore à ce jour les événements climatiques les plus coûteux en France avec 13,9 Md € constants 2020, *Etude : Changement climatique et assurance à l'horizon 2040* (2021). Il s'agit pour la profession de l'assurance de contribuer à une meilleure connaissance des risques naturels et d'apporter une contribution technique aux politiques de prévention.

La gouvernance de l'association est assurée par un conseil d'administration composé de représentants des principaux groupes d'assurance en France. Cependant, toutes les sociétés adhérentes à France Assureurs, qui opèrent sur le marché français dans le domaine des "dommages aux biens des particuliers et des professionnels", participent au financement de ses activités. L'association est affiliée au GIE, "Gestion Professionnelle des Services de l'Assurance" (GPSA) en tant que groupement technique adhérent. Son budget annuel est voté par ses membres.

Ses activités s'organisent, selon quatre dimensions interdépendantes, comme illustrées en 4.1 :
— Connaissance;
— Co-construction;
— Innovation;
— Prévention.

Connaissance Pour informer et sensibiliser, la Mission Risques Naturels a mis en place plusieurs outils visant à améliorer la connaissance des risques naturels en France.

Figure 4.1 – Activités de la Mission Risques Naturels (Source : MRN)

Le premier outil développé par la MRN est le SIG MRN. Cet outil, créé initialement par Chemitte (2008) et enrichi depuis, permet d'analyser l'exposition des biens assurés aux aléas naturels et climatiques. Comme son nom l'indique, il repose sur les Systèmes d'Informations Géographiques (SIG), qui jouent un rôle essentiel dans l'étude des risques naturels. L'indice d'exposition, élaboré en collaboration avec les assureurs, permet d'identifier les zones vulnérables pour chaque type d'aléa.

Un autre outil essentiel est la base de données SILECC (Sinistres Liés aux Catastrophes Climatiques et Naturelles en France), présentée dans les travaux de Bourguignon (2014). Cette base de données est au cœur de nos travaux et sera donc examinée en détail dans la suite de ce chapitre. Son objectif est de fournir une meilleure connaissance du coût des événements naturels, ce qui est nécessaire pour améliorer la gestion et la prévention des risques. La base SILECC vise à être aussi exhaustive que possible et ne se limite pas aux événements de grande ampleur, mais inclut également les événements de fréquence plus élevée. Cette base de données est étroitement liée à une autre base de données développée par la MRN, la base de données des événements CatNat (Catastrophes Naturelles) et climatiques, qui sera également détaillée par la suite. Cette base de données repose sur la caractérisation des événements naturels en fonction de leurs paramètres spatio-temporels, telle qu'introduite par Bourguignon (2014), et a été ensuite étendue aux autres aléas par la MRN. La base de données SILECC offre la possibilité d'agréger les sinistres par événement, ce qui facilite leur étude et leur analyse. Elle constitue également un outil précieux pour l'étude des territoires impactés par les événements naturels. Les enseignements tirés de l'exploitation conjointe de ces bases de données sont extrêmement précieux, tant pour la profession que pour l'intérêt général. Comme nous le verrons par la suite, ces enseignements ont notamment permis d'améliorer les cartographies de l'exposition aux risques naturels. Grâce à ces bases de données et outils, la MRN surveille de près les événements naturels et fournit des rapports détaillés à ses adhérents. Le sujet de ce mémoire concerne l'estimation en temps réel des événements naturels, réalisée au nom de la profession et plus spécifiquement de la fédération. Cette activité s'inscrit parfaitement dans le cadre des efforts de la MRN visant à améliorer la connaissance des événements naturels, et elle mobilise toute son expertise métier.

Co-construction La MRN collabore étroitement avec les autres acteurs de la gestion des risques naturels en France afin de co-construire des actions de prévention. Un exemple concret de cette collaboration est la mise à jour de la carte de susceptibilité au retrait gonflement des argiles, réalisée en col-

laboration avec le Bureau de recherches géologiques et minières (BRGM) en 2020 *Lettre d'information de la Mission Risques Naturels 30* (2019). Cette nouvelle carte prend en compte les sinistres passés et permet ainsi de mieux évaluer l'exposition à ce risque. Cette cartographie revêt une importance cruciale pour la reconnaissance en tant que catastrophe naturelle. Il est donc essentiel que cette carte soit aussi précise que possible. De plus, la loi pour l'évolution du logement, de l'aménagement et du numérique (ELAN) de 2018, à travers son article 68, établit un dispositif visant à favoriser la prévention dans les maisons individuelles construites dans les zones exposées au phénomène de retrait-gonflement des argiles (RGA), identifiées comme présentant un niveau moyen ou élevé d'exposition. Cela confère à cette carte une importance réglementaire significative.

Pour mieux identifier et promouvoir les référentiels techniques de conception du bâti, la MRN anime un groupe de travail qui publie un répertoire de référentiels de résilience du bâti *Référentiels de résilience du bâti aux aléas naturels* (2022). L'objectif de ce répertoire est d'améliorer, à long terme, la résilience des bâtiments face aux aléas naturels, ce qui constitue un enjeu majeur pour maîtriser les coûts associés. En favorisant l'utilisation de référentiels de qualité et en diffusant les bonnes pratiques en matière de construction résiliente, la MRN contribue ainsi à renforcer la capacité des bâtiments à résister aux événements naturels et à limiter les dommages.

Innovation L'innovation est également au cœur des préoccupations de la Mission Risques Naturels, et elle entretient un lien étroit avec la recherche. À ce titre, six thèses CIFRE ont été réalisées ou sont en cours de réalisation en collaboration avec la MRN. Les sujets abordés sont variés, et deux de ces thèses ont déjà abouti à des outils innovants pour la MRN et la profession, comme mentionné précédemment Chemitte (2008), Bourguignon (2014). Deux autres thèses ont contribué à améliorer l'évaluation des mesures de prévention en France Gérin (2011), Guillier (2017). C'est aussi dans le cadre d'une thèse CIFRE, Heranval (2022), que certains des travaux de ce mémoire ont été développés.

Prévention Avec l'augmentation prévue du coût des catastrophes naturelles, la prévention devient un levier d'action majeur. C'est la mission principale de la MRN, et toutes les activités précédemment mentionnées contribuent directement ou indirectement à une meilleure prévention. La MRN participe également à la sensibilisation et à l'information du public, d'où la publication d'études accessibles au grand public basées sur ses nombreux outils. Parmi celles-ci, on peut citer par exemple *Lettre d'information de la Mission Risques Naturels 34* (2020), *Lettre d'information de la Mission Risques Naturels 36* (2021), *Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti* (2018).

La contribution de la MRN à la prévention passe également par l'évaluation de l'efficacité des mesures nationales. Dans Gérin (2011), une évaluation de la pertinence de la couverture des Plans de Prévention des Risques est réalisée. Une approche expérimentale de l'efficacité des Programmes d'Action de Prévention des Inondations (PAPI) est développée dans Guillier (2017). Ces travaux de la MRN abordent les deux mesures phares de prévention des inondations au niveau national.

La structure de la MRN présente une particularité intéressante : elle allie la réactivité et la polyvalence d'une petite équipe tout en bénéficiant des ressources d'un GIE d'envergure. De plus, elle peut compter sur le soutien de la fédération et des sociétés d'assurances. Cette structure active dispose d'un large éventail de compétences et peut mener des activités uniques et pionnières en France. Son réseau et ses ressources lui permettent d'exploiter pleinement des bases de données riches, qui seront présentées dans les sections suivantes.

4.1.2 Base de données événements

Dans l'étude des risques naturels, la première étape consiste à regrouper les événements par catégorie. Cette approche est basée sur la définition initiale proposée par Bourguignon (2014), mais elle a depuis été enrichie et élargie. Les données de sinistralité sont collectées au niveau communal et pour une date donnée, mais il est pertinent de regrouper ces données en événements afin de faciliter leur analyse. Dans notre cas, le regroupement par événement permet de constituer une base d'apprentissage qui sera utilisée pour estimer les coûts lorsqu'un événement se produit. De plus, ce regroupement permet de fournir des indicateurs pertinents sur les territoires touchés. Actuellement, il existe des bases de données spécifiques pour les inondations, la grêle et les tempêtes. Dans cette étude, nous nous concentrerons principalement sur la base de données relative aux inondations, car elle est au cœur de notre application.

Dans notre étude, un événement est défini par une date de début, une date de fin et un ensemble de communes impactées. Les territoires touchés sont identifiés a posteriori en fonction des demandes de reconnaissance en tant que catastrophe naturelle (CatNat) et peuvent également être complétés par des données de sinistralité. Les demandes de reconnaissance CatNat sont regroupées pour définir les événements selon un périmètre spatio-temporel cohérent. Pour ce faire, un arbre de décision est utilisé, comme illustré dans la Figure 4.2. L'objectif est de créer un événement à partir d'une liste de communes en vérifiant si ces communes ont des dates cohérentes et si elles se trouvent dans le même secteur hydrographique ou dans des secteurs hydrographiques adjacents. Si ces conditions sont remplies, un événement unique est formé. Sinon, plusieurs événements distincts sont créés. Cet approche permet de regrouper les communes touchées par des événements similaires, ce qui facilite l'analyse et l'estimation des coûts associés à chaque événement.

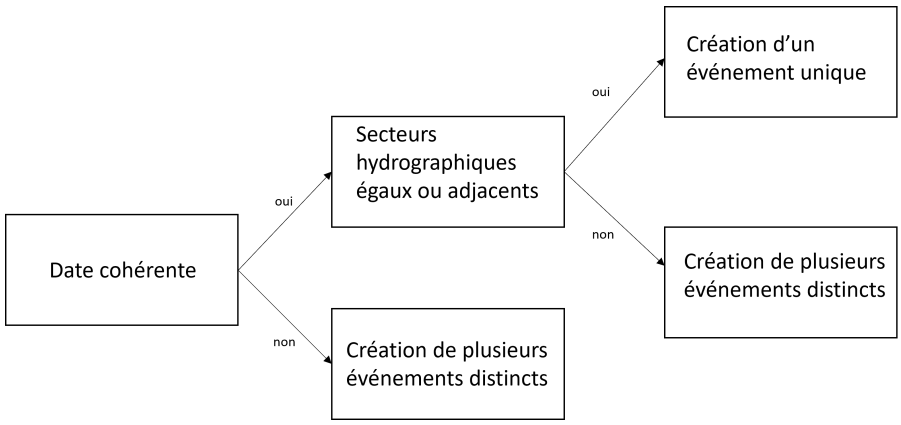


Figure 4.2 – Construction de la Base de données événements

L'arbre de décision utilisé pour regrouper les événements repose sur des critères liés à deux variables :

— les secteurs hydrographiques touchés, ils proviennent de la BD Carthage de l'IGN. Ce découpage est défini par la *Circulaire numéro 91-50 du 12 février 1991* (1991). Les limites s'appuient sur celles des bassins-versants topographiques, ces secteurs permettent de séparer et d'identifier les différents territoires hydrographiques.
— Les dates, elles sont spécifiées dans l'arrêté CatNat et correspondent aux dates de début des événements. Pour regrouper les événements, il est nécessaire de déterminer si les dates sont suffisamment proches, et le seuil dépend du régime de crue. En effet, lors d'une crue lente, un écart de trois jours peut être considéré comme appartenant à un même événement, tandis que

dans le cas d'une crue rapide, cela peut ne pas être le cas. Une table est établie préalablement pour définir les limites spécifiques à chaque régime de crue.

l

Ces critères de secteurs hydrographiques et de dates permettent de regrouper les communes touchées par des événements similaires, en créant des événements cohérents sur le plan spatio-temporel. Ainsi, chaque événement correspond à un périmètre de dommages indemnisés au titre du régime CatNat, dans une période de temps restreinte, à l'échelle communale. L'ensemble des arrêtés CatNat concernant les inondations depuis 1982 est exploité et regroupé pour construire cette base de données.

La base de données événement est constituée de près de 140 000 arrêtés CatNat inondation regroupés en plus de 4 300 événements distincts entre 1982 et 2021. Cette base est très déséquilibrée, comme beaucoup de jeu de données en assurance, il y a des événements majeurs qui concentrent une grande partie des communes. Les 10 événements de plus grande ampleur représentent 35% de la base. Il existe aussi des événements extrêmes en terme de nombre de communes comme l'atteste la figure 4.3.

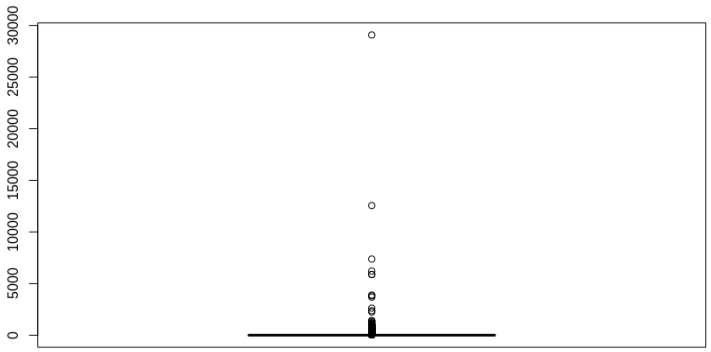


Figure 4.3 – Nombre de communes par événement inondation

Cette base de données couvre l'ensemble du territoire français, avec 99% des communes touchées par au moins un événement d'inondation. Elle comprend un total de 225 000 entrées, chaque entrée représentant une paire commune-événement. Cette base prend en compte tous les périls liés aux inondations, et le tableau 4.1, présente les différents libellés des arrêtés CatNat pris en compte. Il convient également de noter qu'il existe un nombre non négligeable de communes en dehors du régime CatNat qui sont incluses dans la base en fonction de la sinistralité constatée.

Péril	Part de communes rapportées
Chocs Mécaniques lies a l'action des Vagues	3%
Coulée de Boue	0%
Inondations et/ou Coulées de Boue	61 %
Inondations Remontée Nappe	1%
Lave Torrentielle	0%
Raz de Marée	0%
Hors CatNat	35 %

Table 4.1 – Nombre de communes par péril CatNat

4.1.3 Base de données des sinistres

Pour étudier la sinistralité la MRN récolte les sinistres de tous les périls CatNat et climatique auprès de 12 grandes compagnies d'assurance françaises. Cela représente 70% du marché dommages aux biens en France. Ces sinistres sont ensuite harmonisés et agrégés pour former une base cohérente. La localisation des sinistres est également enregistrée avec différents niveaux de précision afin de les croiser avec les événements et les cartes d'exposition. Certains sinistres sont localisés au niveau communal, tandis que d'autres sont localisés au niveau du bâtiment assuré. Un travail similaire est réalisé pour les portefeuilles, ce qui permet d'obtenir des informations sur la répartition géographique des primes d'assurance. Ces données seront utiles dans la suite de notre étude. Le coût des sinistres est actualisé, selon l'indice de la Fédération Française du Bâtiment (FFB), calculé par rapport au coût de la construction d'un immeuble en France. Cela permet d'être sur le même référentiel de coût lorsque l'on parle de sinistres éloignés dans le temps en prenant en compte l'augmentation des coûts liés à la construction. Après la phase préliminaire de mise en forme, nous obtenons une base de données qui enregistre, pour chaque sinistre, la date de survenance, le type de péril, le segment de risque, la localisation et le coût. Ces informations permettent d'analyser et d'étudier la sinistralité liée aux différents événements et aux différents secteurs géographiques et de risque.

En effet, en raison de sa profondeur temporelle, il peut y avoir des incertitudes sur le coût réel des sinistres. Ce problème est bien fréquent dans la littérature actuarielle du provisionnement ("micro-level reserving"), voir par exemple Norberg (1993, 1999), Antonio & Plat (2014), Pigeon et al. (2014) : pour certains types de dommages, plusieurs années peuvent s'écouler entre l'instant de la survenance et celui de la clôture, où le montant du sinistre devient connu, ce qui relie l'étude de ces problématiques à certains développements d'analyse de survie. Il peut également y avoir des incertitudes liées à l'adresse : elle peut ne pas correspondre à l'adresse exacte du sinistre ou être mal géolocalisée. Ces incertitudes sont inhérentes à une étude de la sinistralité à une échelle si fine. Étant donné notre approche de retour d'expérience, nous n'appliquons aucun correctif et analysons la base telle quelle. Nous essayons de réduire au maximum les sources d'incertitude en amont grâce à un nettoyage et un traitement minutieux des données, mais certaines incertitudes font partie intégrante de la base et se répercutent sur nos études.

Grâce à la collecte et à l'agrégation des données de sinistralité, nous avons pu constituer une base de données volumineuse qui atteint un bon niveau de représentativité de la sinistralité en France. Cette base couvre un large éventail de périls et de segments de risque, ce qui permet d'avoir une vision globale de l'impact des sinistres sur le territoire. Grâce à cette représentativité, nous sommes en mesure d'effectuer des analyses approfondies et d'obtenir des informations pertinentes pour la gestion des risques et la prévention des dommages.

4.1.4 La base de données SILECC

Ces deux bases de données sont ensuite fusionnées pour former la base de données des sinistres liés aux catastrophes climatiques et naturelles en France (SILECC).

Cette base permet de suivre la sinistralité en fonction des événements. On ajoute les informations relatives à chaque événement à la liste des informations de la base de données des sinistres. Par exemple, pour les inondations, on peut représenter la répartition des événements les plus coûteux, comme illustré dans la figure 4.5. Cette base est d'une grande valeur pour suivre la répartition des coûts liés aux événements naturels.

On peut aussi observer des différences dans les coûts en fonction du type d'événement et du segment

Figure 4.4 – Construction de la BD SILECC (Source : MRN)

de risque. On observe des disparités en fonction de la reconnaissance CatNat et du segment.

Sinistralité	Coût moyen particuliers	Coût moyen professionnels
Reconnue CatNat	8 000 €	28 000
Non reconnue CatNat	3 000 €	19 000
Hors CatNat	5 000 €	14 500

Table 4.2 – Répartition du coût moyen des sinistres en fonction du segment de risque et de la reconnaissance CatNat

Ce type d'étude et cette base de données sont très utiles pour la profession, notamment pour le positionnement des assureurs sur des sujets de place, tels que la réforme du régime CatNat ou l'étude sur l'impact du changement climatique précédemment mentionnée. La base de données SILECC a par exemple été utilisée pour rendre compte de la sinistralité passée dans cette étude. Elle peut également être utile à l'intérêt général, par exemple pour la mise à jour de la carte du retrait-gonflement des argiles du BRGM.

4.1.5 Carte exposition ruissellement

Plusieurs cartes existent pour mesurer l'exposition aux risques d'inondation en France. Cependant, ces cartes se concentrent principalement sur les inondations causées par des débordements de cours d'eau et ne prennent pas en compte les inondations par ruissellement. Grâce à la base de données SILECC, la MRN a constaté qu'une grande partie des sinistres inondations se produit en dehors de toute cartographie existante. En effet, 56% du nombre total de sinistres inondations se produisent en dehors des Enveloppes Approchées des Inondations Potentielles (EAIP), qui sont des cartographies des zones potentiellement inondables développées par le Ministère de la Transition écologique.

A partir de ce constat la MRN a entrepris le développement d'une cartographie d'exposition prenant aussi en compte l'accumulation des eaux de ruissellement. Pour cela elle se base sur la BD ALTI®de l'IGN. Dans cette base se trouve un modèle numérique de terrain (MNT) maillé qui décrit le relief du territoire français à moyenne échelle. En utilisant le MNT de 25 mètre préalablement corrigé, la MRN calcule ensuite le Compound Topographic Index, Moore et al. (1991). Cet indice combine des informations sur les zones de concentration des eaux et les pentes pour déterminer les

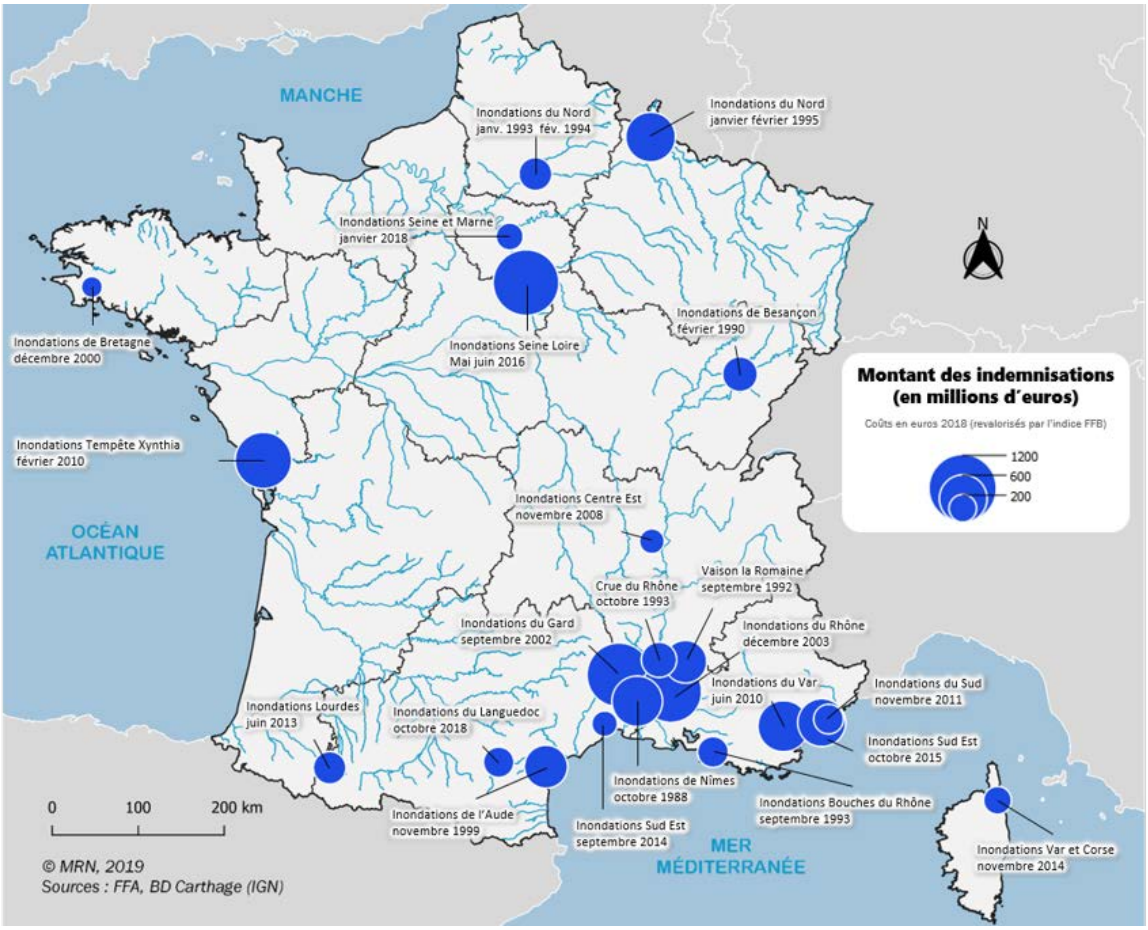


Figure 4.5 – Montants indemnisés pour les inondations CatNat les plus coûteuses selon la BD SILECC (Source : MRN)

zones où l'accumulation des eaux est élevée. Ensuite, cet indice est discrétisé en fonction des critères spécifiques à chaque hydro-écorégion, afin de prendre en compte les différences régionales. Le nouvel indice ainsi obtenu est ensuite agrégé par cercles de 75 mètres et classé en cinq niveaux de risque. Après intégration des zones EAIP pour tenir compte des débordements, une carte (voir Figure 4.6) est générée pour rendre compte des cinq niveaux d'exposition aux inondations. Cette carte permet de mieux prendre en compte la sinistralité dans son ensemble, car elle couvre l'ensemble du territoire. On constate que la majorité des sinistres se produit dans les zones de niveau moyen, fort et très fort. Ces trois zones concentrent 80% du nombre total de sinistres. Cela confirme la validité et l'utilité de cette cartographie, qui permet de rendre compte de manière plus précise de la sinistralité observée.

Zone d'exposition	en nombre	en charge
Très forte	38 %	47 %
Forte	27%	27%
Moyenne	15 %	13%
Faible	5 %	4 %
Très faible	15%	9%

Table 4.3 – Répartition des sinistres inondations par zone d'exposition

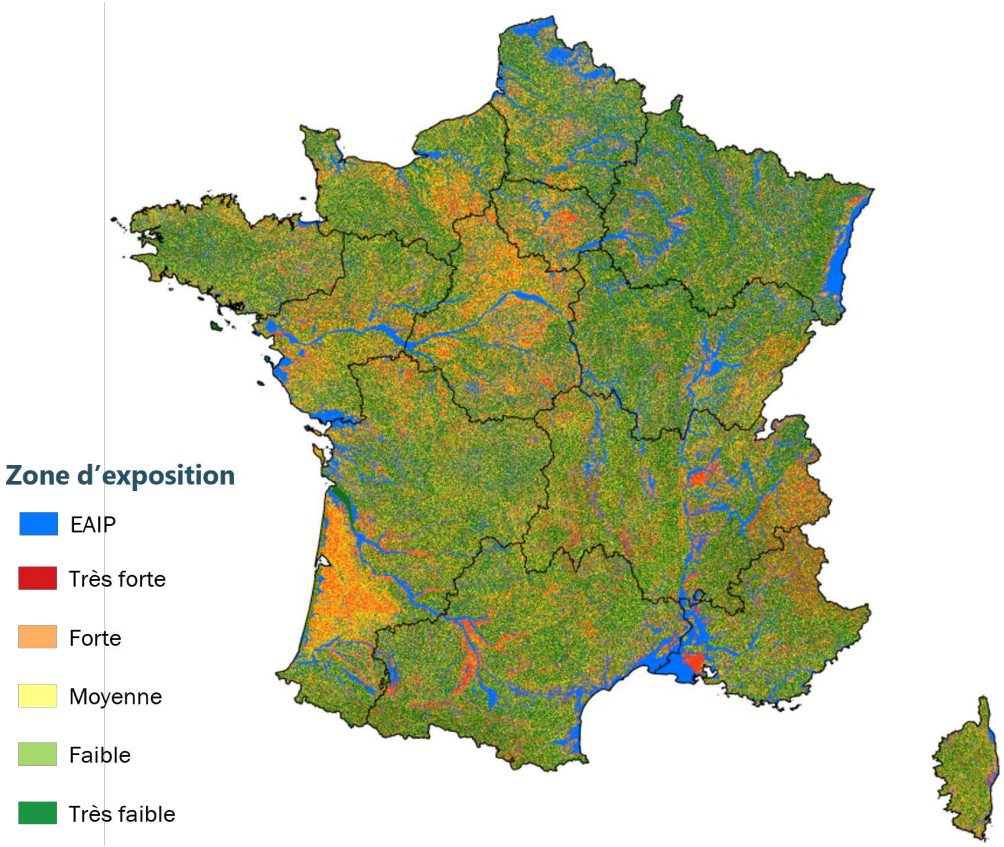


Figure 4.6 – Cartographie MRN d'exposition aux inondations (Source : MRN)

4.2 Risque Cyber

Pour étudier le risque cyber, nous n'avions pas accès à des données de sinistralité, nous avons donc utilisé une base de données publique, la Privacy Rights Clearinghouse (PRC). Cette base de données est l'une des rares sources disponibles sur les événements Cyber qui fournit une évaluation de la gravité des événements. Cette information est nécessaire pour notre application, notre méthode visant précisément à estimer cette gravité. La base de données du PRC ne fournit pas directement la perte associée à un événement, mais elle rapporte le nombre d'enregistrements, c'est-à-dire le nombre de comptes d'utilisateurs affectés par la violation. Ce nombre peut ensuite être utilisé pour déduire un coût, comme décrit dans Farkas, Lopez & Thomas (2021).

Le Privacy Rights Clearinghouse (PRC) est une organisation à but non lucratif fondée en 1992 dans le but de protéger la vie privée des citoyens américains. Depuis 2005, le PRC maintient une base de données recensant les entreprises impliquées dans des violations de données affectant les citoyens américains. Les données de cette base sont constituées d'informations publiquement disponibles sur les violations signalées et ne peuvent pas être considérées comme une représentation complète et précise de toutes les violations de données aux États-Unis. Elles reflètent les violations signalées aux États-Unis qui sont rendues publiques par des entités gouvernementales. Pour notre étude, nous utilisons la base de données datée du 29 mars 2023, mise à notre disposition par l'équipe du Privacy Rights Clearinghouse, pour nos travaux et que nous tenons encore à remercier ici.

Cette base de données contient 11 222 événements cyber touchant principalement des entreprises

américaines. Elle recueille des informations sur chaque événement cyber, telles que son type, le nombre d'enregistrements affectés par la violation et une description de l'événement, ainsi que des informations sur la victime, comme le nom de l'entreprise ciblée, ses activités et sa localisation. Différentes modalités sont observées pour les types de violation et les organisations touchées, comme décrit en détail ci-dessous.

Pour les types de violation :
— CARTE - Fraude aux cartes de débit et de crédit sans piratage (dispositifs d'écrémage dans les terminaux des points de service, etc.)
— HACK - Piratage par un tiers ou infection par un logiciel malveillant
— INSD - Insider (employé, contractant ou client)
— PHYS - Physical (documents papier perdus, jetés ou volés
— PORT - Dispositif portable (ordinateur portable, PDA, smartphone, clé USB, CD, disque dur, bande de données, etc. perdus, mis au rebut ou volés)
— STAT - Stationary Computer Loss (perte, accès inapproprié, mise au rebut ou vol d'un ordinateur ou d'un serveur non conçu pour la mobilité)
— DISC - Divulgation involontaire n'impliquant pas de piratage, de violation intentionnelle ou de perte physique (informations sensibles publiées publiquement, mal manipulées ou envoyées à la mauvaise personne par le biais d'une publication en ligne, d'un courriel, d'un courrier ou d'une télécopie).
— UNKN - Inconnu (pas assez d'informations sur la violation pour savoir comment exactement les informations ont été exposées)

Pour les types d'organisation :
— BSF - Entreprises (services financiers, banques, assurances)
— BSO - Entreprises (fabrication, technologie, communications, autres)
— BSR - Entreprises (commerce de détail/marchand, y compris les épiceries, les détaillants en ligne, les restaurants)
— EDU - Établissements d'enseignement (écoles, collèges, universités)
— GOV - Gouvernement et armée (gouvernements nationaux et locaux, agences fédérales)
— MED - Soins de santé et prestataires médicaux (hôpitaux, services d'assurance médicale)
— NGO - Organismes à but non lucratif (organisations caritatives et religieuses)
— UNKN - Inconnu

Chapter 5

Estimation du coût d'un événement Cyber

5.1 Introduction

Dans ce chapitre, nous présentons une première application de notre méthode pour estimer le coût d'un événement Cyber. Nous nous appuyons sur la base de données PRC décrite dans la section 4.2. C'est une base qui est largement utilisée pour l'étude du risque cyber. On peut citer, par exemple, les articles de Li & Mamon (2023) ou de Carfora & Orlando (2019) qui ont exploité cette base pour modéliser le risque cyber. En outre, plusieurs mémoires d'actuariat ont également porté leur intérêt sur ce sujet et sur cette base, tels que ceux de Anais (2019), Peyrat (2023) ou Bastard (2021). Cette base de données a été largement utilisée en partie en raison de son accès libre. Cependant, récemment, elle est devenue payante et l'organisme responsable, la "Privacy Rights Clearinghouse Team", a effectué un important travail de nettoyage et de consolidation des données. Pour notre étude, cette association a gracieusement accepté de nous mettre à disposition la base dans sa version la plus récente datant de mars 2023. Cette base de données est une des seules qui nous permet d'estimer l'ampleur d'un événement, grâce au nombre de perte de données que l'on peut ensuite utiliser pour estimer le coût. En effet, comme étudié dans l'article de Farkas, Lopez & Thomas (2021), il est possible de lier le coût au nombre de pertes de données. Une première formule introduite par Jacobs (2014) permet de relier le volume de pertes de données, noté Y, à une perte financière $L = f(Y)$ par l'équation $log(L) = \beta + \alpha log(Y)$. Avec α et β à estimer sur des données. L'article de Farkas, Lopez & Thomas (2021) propose une estimation de ces paramètres qui prend en compte les méga-événements et qui a été ajustée sur la base PRC, ce qui correspond bien à notre problème. Nous utiliserons donc ces paramètres et la formule devient :

$$log(L) = 9.59 + 0.57 log(Y).$$

Cette formule n'est cependant pas précise pour établir une correspondance exacte entre une perte financière et le nombre de perte. Notre objectif est simplement d'avoir une approximation grossière. Avec les données publiques dont nous disposons, il est impossible de prétendre que nous pouvons effectuer cette évaluation avec une bonne précision statistique. Afin de limiter les approximations, nous travaillerons autant que possible sur le nombre de perte de la base PRC, par exemple pour la partie de l'arbre de régression, et nous limiterons l'utilisation de cette formule à l'application numérique de la théorie de la crédibilité.

Nous pouvons ainsi utiliser cette base de données pour illustrer notre méthode. Dans un premier temps, nous allons appliquer la méthode CART GPD pour regrouper les événements en classes homogènes en termes de distribution extrême, ce qui nous fournira le prior pour la théorie de la crédibilité. Pour illustrer la tarification, nous utiliserons un exemple fictif, car malheureusement, nous n'avons pas accès à des données de sinistralité permettant de confronter cette méthode dans le cadre du risque Cyber. Cependant, dans le chapitre suivant sur les inondations, nous pourrons réaliser une application sur des données réelles.

5.2 Application

5.2.1 CART GPD

Nous commençons donc par appliquer la méthode CART GPD à la base PRC afin d'estimer le prior. Grâce à cette méthode, nous créons des classes qui sont homogènes dans leurs comportements extrêmes. On utilise dans la base de données PRC, décrite en 4.2, le nombre d'enregistrement, qui est la variable que nous cherchons à prédire, et pour chaque événement les informations suivantes :

— les types de violation,
— les types d'organisation,
— les sources.

Nous disposons de 11 000 événements enregistrés entre 2005 et 2022. La variable d'intérêt, à savoir le nombre de violation, présente une grande volatilité, allant de 0 à 250 000 000, avec une variance empirique de 1.67e+13. Nous pouvons également noter que les 10 événements les plus coûteux représentent 62% du coût total de la base de données, tandis que les 100 premiers événements représentent 98% du coût total. On est bien en présence d'événements extrêmes. L'étude de ces événements permettra donc d'expliquer une grande partie de la sinsitralité.

Nous cherchons à comprendre l'hétérogénéité de la variables nombres d'enregistrement, pour ces événements extrêmes. Comme décrit précédemment pour définir un événement extrême on choisit un seuil u qui correspond à un compromis biais variance. Nous choisissons ici un seuil $u = 500$ selon une analyse graphique de la distribution et du Hill Plot, comme présenté dans Resnick (2007), Boucheron & Thomas (2015). On a 6 600 événements au-dessus de notre seuil.

Ensuite, la régression CART GPD est appliquée à la sous-base des événements dont le nombre de perte est supérieur au seuil. Les variables de la base de données sont résumées dans les tables 5.1. On peut encore observer la grande volatilité de la variable à prédire, avec un premier quartile à 1 200, une médiane à 2 692, une moyenne à 298 722 et un troisième quartile à 14 000. La moyenne, bien supérieur à la médiane illustre encore présence d'événements extrêmes.

L'arbre obtenu avec cette méthode est présenté dans la figure 5.1. Les diagrammes quantile-quantile sont également disponibles dans la figure 5.2. Notre arbre comporte 8 feuilles, avec des séparations basées sur toutes les variables fournis. Cet arbre nous permet de déduire la distribution des évènements en fonction de certaines covariables. On peut observer que tous les paramètres de formes γ sont supérieurs à 1, ce qui illustre les fortes valeurs prisent par les nombres d'enregistrements. Toutes les classes ne sont pas semblables et la répartition est hétérogène comme on peut s'y attendre. On a deux classes "moins extrêmes" (même si le paramètre de forme reste supérieur à 1) qui regroupe 65% des évènements et d'autres classes qui au contraire sont plus extrêmes mais regroupent moins d'événement comme par exemple une classe avec un γ de 3.35 qui regroupe 2% des événements. On a une répartition qui permet de bien discriminer les événements, cependant il faut bien noter que cette réparation est très dépendante des données et en particulier de la qualité des données elle présente

donc ainsi forcement des incertitudes.

Table 5.1 – Liste des variables catégorielles dans la base de données et leurs caractéristiques. Le Tableau a) présente le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et le maximum de la variable cible, le nombre d'événements, et pour les variables catégorielles, le Tableau b) présente le nombre d'observations par catégorie.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Records.Affected	501	1 200	3 192	299 222	14 500	250 000 000

a)

Variable	Category	Number of observations
Type.of.Breach	CARD	84
	DISC	759
	HACK	1146
	INSD	280
	PHYS	1411
	PORT	700
	STAT	178
	UNKN	2049
Organization.Type	BSF	301
	BSO	359
	BSR	126
	EDU	524
	GOV	309
	MED	4415
	NGO	39
	UNKN	534
Source	Media	3
	Non profit organization	1560
	US GA: Federal - HIPAA	4044
	US GA: State	1000

b)

5.2.2 Application de la théorie de la crédibilité

Une fois que l'on a l'arbre de régression on peut appliquer la théorie de la crédibilité pour avoir un coût pour ces événements. Dans un premier temps il faut transformer le nombre de perte en coût. En effet les paramètres de la GPD correspondent au nombre d'enregistrement ce qui n'est pas l'objectif de notre méthode de tarification. Comme on l'a vu en introduction il est possible de lier ces deux variables, grâce à la formule :

$$log(L) = 9.59 + 0.57log(Y).$$

Cependant, le paramètre de forme de la distribution GPD de f(Y) peut être facilement déduit. Si $P(Y \geq y) \sim Cy^{-1/\gamma}$, où $\gamma > 0$ est le paramètre de forme de Y et C est une constante, en considérant $f(y) = exp(\alpha + \beta log(y))$, cela conduit à :

$$P(f(Y) \geq z) = P(Y \geq \exp(\frac{logz - \alpha}{\beta}))$$

$$\sim C \exp(\frac{-\alpha}{\beta\gamma}) z^{\frac{-1}{\beta\gamma}}.$$

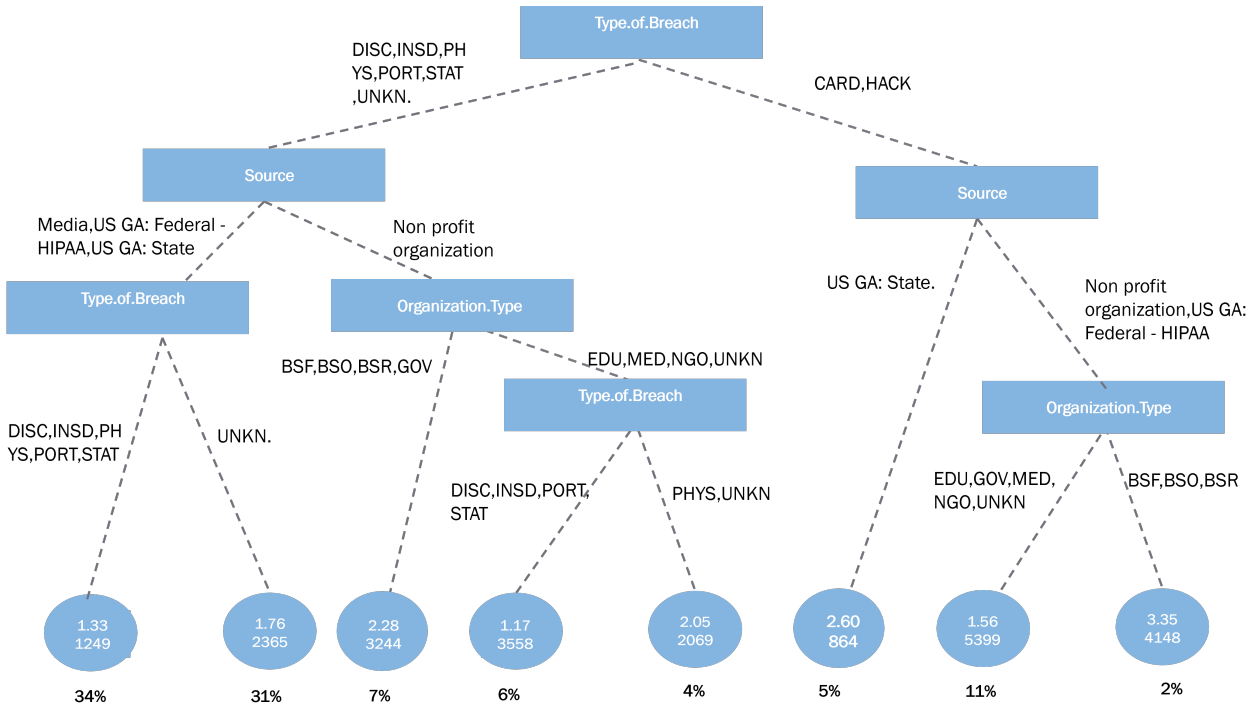


Figure 5.1 – Arbre de régression GP obtenu pour les événements cyber. Pour chaque feuille, la valeur du paramètre de forme γ (première ligne) et du paramètre d'échelle σ (deuxième ligne) sont indiquées. Le pourcentage d'observations attribué à chaque feuille est aussi indiqué.

On en déduit ensuite, par identification, que le nouveau paramètre de forme est $\beta\gamma$. Cependant on ne peut pas déterminer directement la valeur de σ par le calcul, et nous devons l'estimer. Pour cela, plusieurs méthodes sont possibles. On peut estimer les paramètres de la GPD dans chaque feuille par maximum de vraisemblance ou bien par la méthode des moments. Une autre approche que nous allons appliquer ici est d'utiliser la médiane empirique pour approximer σ. En effet, la formule de la médiane théorique est égale à $\sigma(2^{\gamma}-1)/\gamma$. Ainsi, en utilisant la valeur de γ que nous venons de déduire, et en approximant la médiane théorique par la médiane empirique, nous pouvons estimer sigma. Cette démarche est assez robuste, car, comme on peut le voir sur le nombre de pertes, les médianes théoriques et empiriques dans chaque feuille de l'arbre sont très proches.

Leaf	Empirical Median	Theoretical Median
1	1486	1421
2 3477	3207	
3	6500	5487
4	3800	3801
5	4000	4702
6	1800	1682
7	7000	6743
8	19500	11387

Table 5.2 – Empirical median and mean, and theoretical median and mean for each leaf.

Ensuite pour avoir le coût nous pouvons appliquer la formule démontrée dans les précédents chapitres, qui nous dit que la prime de crédibilité est donnée par:

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

avec

$$r = \frac{1}{\gamma},$$

$$\lambda = \frac{\sigma}{\gamma}.$$

$\sum_{i=1}^{n} Y_i$ et n correspondent quand à eux à l'historique des données.

On peut donc ensuite déduire un coût pour un événement, on verra dans le chapitre suivant une application détaillé mais pour se fixer les idées on peut prendre un exemple fictif.

Imaginons que l'entreprise, "Prudential life insurance company", soit victime d'une perte de donnée. Le jour J on a les informations relatives à cette perte de données, disons que c'est une fuite interne (Type of breach = INSD) que la source est un media (Source =Media), on sait aussi que c'est une compagnie d'assurance (Organisation Type = BSF). On peut grâce à ces données et l'arbre en déduire que $\gamma_{records} = 1.33$ et $\sigma_{records} = 1249$ et donc avec les méthodes d'approximations évoquées plus haut que $\gamma_{price} = 0.76$ et $\sigma_{price} = 115.98$. Ensuite en regardant dans l'historique de la base PRC on trouve que cette société a eu dans le passé $n = 6$ sinistre que la somme du coût de ces sinistres est de 1 206 682\$. On peut donc grâce à notre formule dire que le coût du sinistre rapporté sera de :

$$\frac{\lambda + \sum_{i=1}^{n} Y_i}{r + n - 1} = 191\ 257\$.$$

Prenons deux autres exemples intéressants. Supposons cette fois-ci que la "Bank of America" soit victime d'une perte de données, mais dans ce cas, il ne s'agit plus d'une fuite interne, mais d'un "hack" (Type of breach = HACK). La source de la violation est toujours un organisme sans but lucratif (Source = Non-profit organization), et l'on sait également que l'Organisation Type est BSF. Ici, nous sommes dans le cas où $\gamma_{records} = 3.35$ et $\sigma_{records} = 4148$, ce qui implique que $\gamma_{price} = 1.90$ et $\sigma_{price} = 502.98$. Comme γ_{price} est supérieur à 1, la formule de la crédibilité n'est pas applicable car le calcul de l'espérance est infini. Cependant, nous avons toujours accès à un résultat général : nous pouvons examiner les quantiles de la distribution. En effet, nous disposons des paramètres de la GPD, ce qui nous permet, par exemple dans ce cas, de calculer que le quantile 99 est de 1 736 073. Cela nous fournit des informations utiles et précieuses même sans le calcul de la prime.

Enfin en dernier exemple prenons le cas d'une société qui n'apparaît pas dans la base PRC car elle n'a jamais eu de perte de données (en tous cas rapportée dans la base PRC). C'est le cas qui sera sûrement le plus fréquent si jamais on utilise cette méthode sur des données réelles. En effet déjà dans la base PRC il n'y a que 17% des entreprises qui ont plus d'un événements. On peut donc logiquement penser que dans le cas réel ce pourcentage sera bien supérieur. Prenons :

— Type of breach = HACK,
— Source = Non profit organization,
— Organization Type = MED.

Alors on est dans le cas où :

$$\gamma_{records} = 1.56, \sigma_{records} = 5399$$

et donc

$$\gamma_{price} = 0.88, \sigma_{price} = 280.50$$

Mais ici $\sum_{i=1}^{n} Y_i$ et n sont tous les deux nuls car on a pas d'historique. Dans notre méthode on peut quand même à partir des paramètres de la GPD proposer une estimation du coût et dans notre cas fictif on pourrai dire que le coût de l'événement serai de

$$\frac{\lambda + \sum_{i=1}^{n} Y_i}{r + n - 1} = \frac{\lambda}{r - 1} = 1275\$$

Ces étapes peuvent bien sûr être automatisées, et en pratique, elles le seront. Cependant, nous avons souhaité décrire la procédure manuellement en détaillant ces étapes pour mettre en avant l'un des avantages majeurs de notre méthode de tarification : son caractère interprétable et la liberté qu'elle offre aux gestionnaires de risques. En effet, tous les paramètres sont accessibles et peuvent être ajustés si nécessaire en fonction de l'expérience et du type d'événement observé. Ceci s'avère particulièrement précieux pour ce type de risque.

5.3 Discussion

Nous avons donc dans ce chapitre monter une illustration de notre méthode de tarification pour le risque cyber à partir de la base PRC. Dans un premier temps nous avons appliqué la méthode CART GPD pour trouver le prior en fonction de trois caractéristiques :

— les types de violation,
— les types d'organisation,
— les sources.

Ensuite, nous avons illustré notre approche avec trois exemples fictifs. Le premier permet d'obtenir un coût pour une société d'assurance en fonction des caractéristiques et de l'historique des données. Dans le deuxième exemple, nous nous sommes intéressés au cas où la théorie de la crédibilité ne peut pas s'appliquer car le paramètre γ est supérieur à 1. Enfin, dans le dernier exemple, nous nous sommes penchés sur la situation où il n'y a pas d'historique disponible. Dans tous les cas, nous pouvons réellement suivre l'évolution du coût ainsi que les paramètres qui vont l'influencer et dans quelle mesure. C'est très précieux pour la gestion d'un risque émergent comme le risque Cyber.

Ces résultats concernant le risque cyber sont prometteurs, mais une analyse plus robuste pourrait être réalisée en se basant sur des données réelles. En effet, compte tenu de la base de données utilisée, plusieurs incertitudes subsistent, notamment en ce qui concerne les coûts. Il convient donc d'interpréter ces résultats avec prudence. Notre méthode pourrait se révéler intéressante pour les compagnies d'assurance, car elle prend en compte l'hétérogénéité présente dans les données. À cet égard, utiliser cette méthode pour élaborer une police d'assurance cyber en se fondant sur les sinistres passés pourrait être particulièrement pertinent.

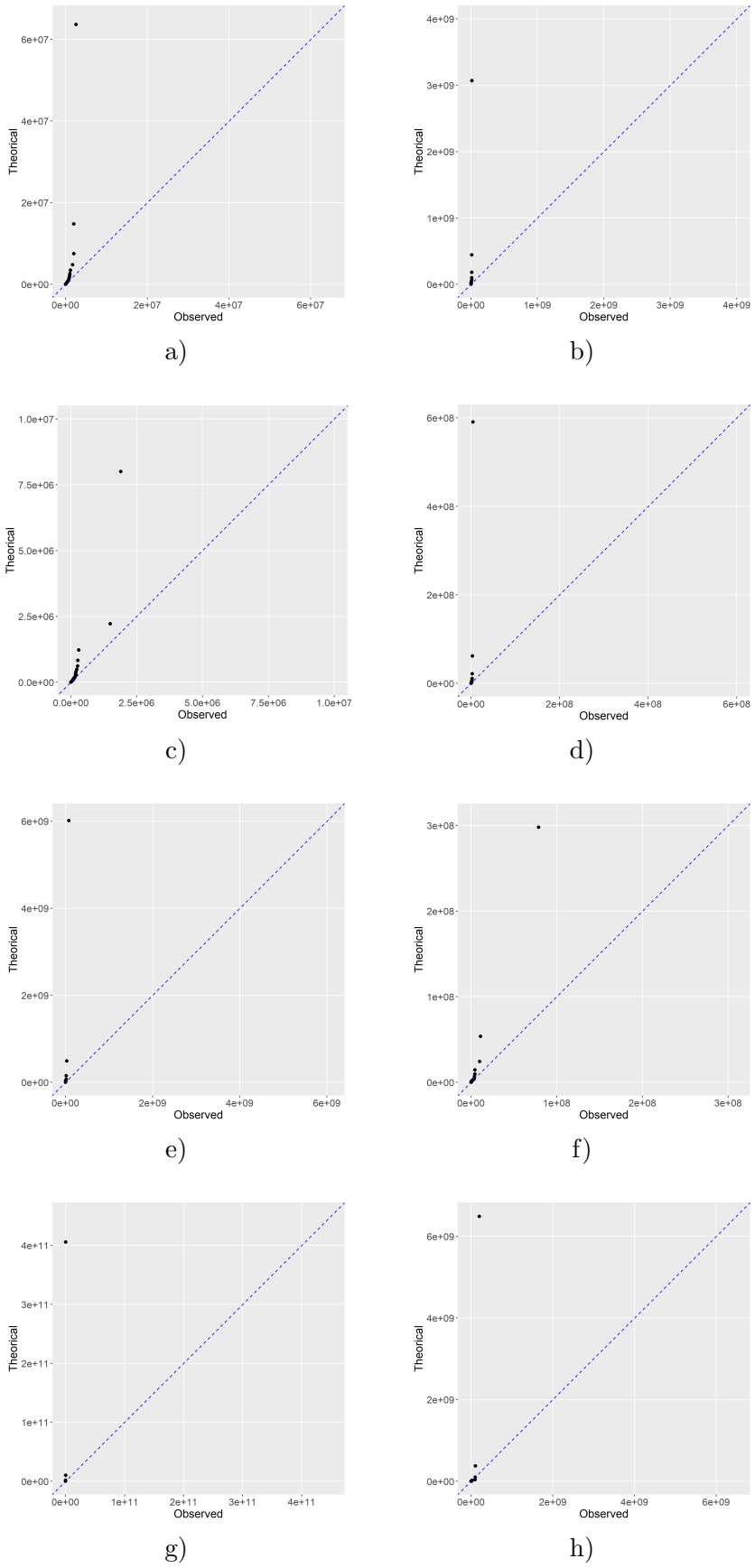


Figure 5.2 – Diagrammes quantile-quantile pour chaque feuille de l'arbre

Chapter 6

Estimation du coût des inondations rapidement après leurs occurence

6.1 Introduction

6.1.1 Contexte

L'estimation du coût des inondations est enjeu majeur pour le secteur de l'assurance notamment pour évaluer leurs expositions. Néanmoins c'est un exercice difficile qui comporte de nombreuses incertitudes Hall & Solomatine (2008), Eleutério (2012). Dans ce chapitre, nous appliquons notre méthode pour estimer le coût des événements d'inondation, pour avoir un point de comparaison nous introduisons aussi une méthode Fréquence/séverité. Notre objectif se concentre sur l'estimation des conséquences des inondations, sans prendre en compte la modélisation du phénomène lui-même. Notre démarche s'inscrit dans le cadre d'une mission d'appui à France Assureurs afin de dimensionner les réponses en cas de gestion de crise liée aux événements naturels. Ainsi, nous cherchons à estimer le coût d'un événement d'inondation après sa survenue. Notre attention se porte uniquement sur les événements "majeurs", définis par la combinaison des critères suivants :

— Une étendue spatiale et temporelle importante de l'événement ;
— Un retentissement médiatique significatif ;
— Un nombre de décès provoqués par l'événement ;
— Une spécificité ou rareté de l'événement.

La principale contrainte de notre approche est d'être en mesure de fournir une estimation rapide du coût des événements inondations. Cependant, nous avons également accordé une attention particulière à développer une méthode compréhensible, facile à utiliser et dotée de paramètres contrôlables. Étant conscient de la difficulté de cet exercice et de l'existence d'incertitudes inhérentes à toute estimation, nous avons cherché à laisser une part de contrôle aux gestionnaires de risques. Nos méthodes reposent avant tout sur l'expertise métier et la comparaison avec les données historiques, robuste récoltées par la MRN. Ainsi, nous avons privilégié une approche qui intègre ces éléments pour fournir une estimation la plus précise possible.

En France, grâce au régime CatNat, la Caisse Centrale de Réassurance (CCR) est un acteur incontournable qui propose des estimations pour les catastrophes naturelles. Elle développe des outils de modélisation permettant d'estimer le coût d'un événement quelques jours après sa survenance, ainsi que de mesurer son exposition financière. Des descriptions de leurs méthodes pour les inondations peuvent être trouvées dans de nombreuses références, notamment Moncoulon et al. (2014), Moncoulon

(2014), Moncoulon & Quantin (2013). La thèse de Mao (2019) permet de comprendre en détails ces modèles. Ces travaux de recherche fournissent des informations précieuses sur les méthodes utilisées par la CCR pour estimer les coûts des événements et évaluer les risques financiers associés. La CCR a développé un modèle déterministe pour estimer le coût et un modèle probabiliste pour évaluer l'exposition aux catastrophes naturelles. Notre approche est similaire au modèle déterministe développé par la CCR, cependant, nous n'utilisons ni les mêmes données ni les mêmes méthodes. Le modèle de la CCR repose sur un modèle d'aléa, un modèle de vulnérabilité et un modèle de dommage.

- Le modèle d'aléa permet de simuler les écoulements d'eau en intégrant les différents processus hydrologiques conduisant à une crue. Il intègre la transformation de la pluie en débit (modèle pluie-débit), l'écoulement, l'infiltration et l'hydrologie des cours d'eau.
- Le modèle de vulnérabilité est construit à partir des données de police d'assurance du marché et fournit, pour chaque bien, la localisation et les caractéristiques du risque.
- Le modèle de dommage combine ces deux informations pour estimer les dommages au niveau de chaque bien assuré. Il prend en compte une probabilité de sinistre, un taux de destruction, la probabilité de reconnaissance CatNat et la valeur du bien. Pour déterminer le taux de destruction et la fréquence de sinistres, des distributions statistiques sont calibrées sur la base de l'intensité de l'aléa.

Notre approche est différente car nous ne modélisons pas l'aléa, mais nous déterminons, via la cartographie décrite en section 4.1.5, les zones les plus susceptibles d'être impactées. Nous mesurons la vulnérabilité en calculant le nombre de biens par zone, et les dommages sont estimés à partir de ces données. Pour cela, nous utilisons les méthodes décrites ci-dessous, et notre démarche est similaire dans l'idée à celle de la CCR. Nous tentons d'ajuster des distributions statistiques, mais avec des données et des échelles d'analyse différentes.

6.1.2 Mode opératoire

Notre objectif est d'estimer rapidement le coût d'un événement après son occurrence. Pour cela, France Assureurs mène une enquête statistique en envoyant un questionnaire directement aux sociétés d'assurance. Cette méthode donne des résultats très fiables car elle prend en compte directement le nombre de sinistres rapportés. Cependant, elle prend du temps et les résultats ne sont pas connus avant une à deux semaines. Notre approche vise à fournir une estimation dans un délai plus court, généralement un à deux jours après l'événement

Cela soulève un premier problème, qui est la définition de l'événement. Comme décrit dans la section 4.1.2, notre définition des événements repose sur les reconnaissances en état de catastrophe naturelle. Cependant, ces reconnaissances interviennent bien après les événements. Dans le cas d'une procédure accélérée, les reconnaissances peuvent prendre une à deux semaines, mais uniquement pour les communes les plus sévèrement touchées. Dans les cas usuels, le délai pour les inondations dépasse souvent deux mois. Par conséquent, nous ne pouvons pas nous appuyer sur les reconnaissances et nous devons nous tourner vers d'autres sources.

Dans un premier temps, nous examinons les communes situées le long des cours d'eau placées en vigilance orange et rouge selon le service d'information sur le risque de crues des principaux cours d'eau en France, Vigicrue. Cela nous fournit un premier périmètre très susceptible d'être impacté par les débordements. Ensuite, nous complétons ce périmètre avec les informations que nous trouvons dans la presse locale. La définition d'un nouvel événement repose sur une procédure préalablement définie, mais peut varier en fonction des événements. Le gestionnaire de risques intervient pour corriger et ajuster ce périmètre en fonction des informations disponibles et de son expertise. Ainsi, nous obtenons une liste de communes qui constitue, avec la date, notre événement.

Cette définition d'un événement, qui diffère de celle utilisée pour notre base historique, crée une première source d'incertitude car les périmètres peuvent varier en fonction des différentes définitions. Utiliser comme base d'apprentissage des événements définis selon la procédure appliquée pour les événements en temps réel pourrait permettre d'éviter cela. Cependant, nous ne disposons pas de l'historique nécessaire donc nous utilisons la base reposant sur les arrêtés CatNat.

Nous exploitons les informations disponibles à la MRN sur les événements passés pour aider à caractériser les événements en cours. Pour cela, nous utilisons deux méthodes : une méthode d'arbre de régression basée sur le comportement de la queue de distribution, décrite précédemment, et une méthode de type sévérité-fréquence qui repose sur la comparaison d'événements similaires, qui va nous permettre d'avoir un point de comparaison.

6.2 Application de la théorie de la crédibilité

6.2.1 CART GPD

Dans cette partie, nous appliquons la méthode CART GPD aux événements d'inondations afin de mieux comprendre leurs comportements extrêmes et d'estimer le prior. Nous observons une certaine hétérogénéité dans les événements en fonction de certaines caractéristiques telles que la région météorologique. Grâce à cette méthode, nous créons des classes qui sont homogènes dans leurs comportements extrêmes, ce qui est extrêmement précieux d'un point de vue opérationnel.

On utilise la base de données SILECC, décrite en 4.1.4, enrichie de plusieurs covariables. En plus du coût, chaque événement est accompagné des informations suivantes :

- la région météorologique (en cas de plusieurs régions pour un même événement, nous séparons l'événement par région),
- la saison,
- le nombre d'hydro-écorégions affectées,
- le nombre de logements individuels en zone de risque d'inondation,
- le nombre de professionnels en zone de risque d'inondation.

La zone de risque d'inondation est déterminée à partir de la carte de ruissellement réalisée en 4.1.5. Nous calculons le nombre de particuliers et de professionnels se trouvant dans les zones de risques moyennes et fortes. Ces informations, qu'elles soient déterministes ou caractéristiques de l'événement, sont disponibles rapidement après son occurrence. Nous disposons de 2 400 événements enregistrés entre 1999 et 2019. La variable d'intérêt, à savoir le coût des événements, présente une grande volatilité. Les coûts varient de 0 à 380 487 000 euros, avec une variance empirique de 3.16e+14. La figure 6.1 illustre la moyenne des coûts parmi les 10% des événements les plus coûteux pour chaque région météorologique. Cela met en évidence l'hétérogénéité en termes de gravité des événements les plus importants. Nous pouvons également noter que les 10 événements les plus coûteux représentent 49% du coût total de la base de données, tandis que les 100 premiers événements représentent 87% du coût total.

Nous cherchons à comprendre l'hétérogénéité du coût des événements inondations les plus sévères, les événements extrêmes. Comme décrit précédemment pour définir un événement extrême on choisit un seuil u qui correspond à un compromis biais variance. Nous choisissons ici un seuil $u = 100\ 000$ selon des considérations pratiques. Ce choix est validé par une analyse de la sensibilité, on a fait varier le seuil comme montré en figure 6.3. On a 820 événements au-dessus avec notre seuil $u = 100\ 000$.

Ensuite, la régression CART GPD est appliquée à la sous-base des événements dont le coût est

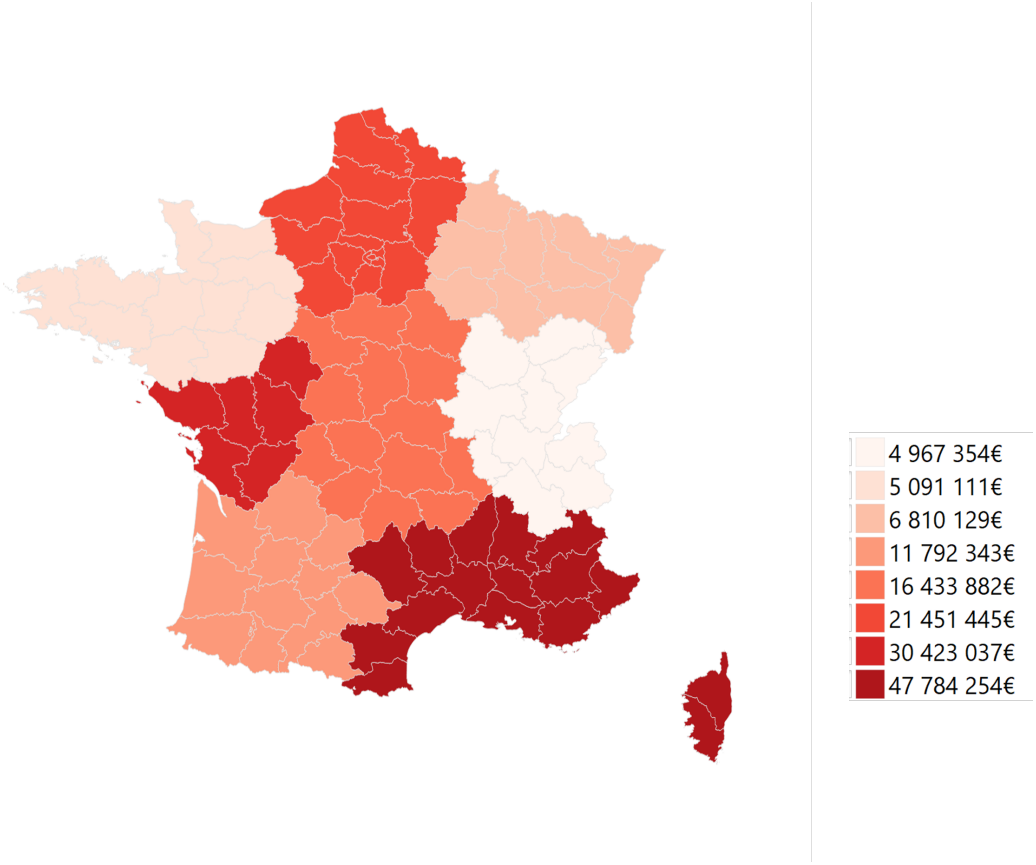


Figure 6.1 – Cartographie des coûts des événements inondations de 1999 à 2019. Pour chaque région météorologique, on montre la moyenne du coût des événements faisant partie des 10% les plus chers. Le rouge clair suggère un coût faible alors que le rouge foncé un coût plus important.

supérieur à 100 000 euros. Les variables de la base de données et leurs caractéristiques sont résumées dans les tables 6.1 et 6.2. On peut encore observer la grande volatilité de la variable de coût.

L'arbre obtenu avec cette méthode est présenté dans la figure 6.2. Les diagrammes quantile-quantile sont également disponibles dans la figure 6.4. Notre arbre comporte 6 feuilles, avec des séparations basées sur 3 critères :

— le nombre de logements individuels en zone de risque inondation,
— le nombre de professionnels en zone de risque inondation,
— le nombre d'hydro-écorégions affectées.

Cette répartition semble cohérente, les deux premières covariables reflètent l'exposition aux inondations ainsi que la densité de population de la zone touchée, tandis que la troisième covariable représente l'étendue de l'événement. Le cas le plus extrême correspond à la feuille la plus à droite, qui a un paramètre de forme de 0.92 et contient 7% des événements. Cette feuille correspond à une proportion importante de logements individuels touchés et à une zone étendue.

Dans le tableau 6.3, nous avons calculé les médianes et moyennes empiriques et théoriques pour chaque feuille. Il est important de rappeler que dans le cas d'une distribution GPD avec des paramètres (γ, σ), la médiane théorique est égale à $\sigma(2^\gamma - 1)/\gamma$ et la moyenne théorique est égale à $\sigma(1-\gamma)$ pour $\gamma < 1$, et à ∞ pour $\gamma \geq 1$. Pour chaque feuille, la médiane est bien inférieure à la moyenne, suggérant que nous sommes effectivement confrontés à des événements extrêmes.

Ensuite, nous observons un très bon ajustement avec des valeurs très proches entre les médianes théoriques et empiriques dans toutes les feuilles. De plus, pour la moyenne, les valeurs théoriques et empiriques sont également proches, sauf pour les feuilles 4 et 6 qui correspondent aux paramètres de forme les plus importants. Les paramètres semblent donc très bien ajustés à la distribution dans chaque feuille et la classification semble aussi pertinente.

Variable	Min	1^{er}Q	Médiane	Moyenne	$3^{ème}$Q	Max
Coût	100 093	199 287	477 943	6 066 835	1 941 047	380 487 161
Nombre d'hydro-écorégions affectées	1	1	2	4	4	35
Nombre de logements individuels en zone de risque inondations	0	5 874	20 692	92 477	71 094	4 097 075
Nombre de professionnels en zone de risque inondations	0	2 230	8 163	44 830	26 321	2 050 165

Table 6.1 – Liste et résumé des statistiques descriptives des variables quantitatives utilisées.

Variable	Catégories	Nombre d'observations
Régions météorologiques	Centre	60
	Nord-Ouest	85
	Nord	135
	Nord-Est	87
	Est	96
	Sud	209
	Ouest	30
	Sud-Ouest	121
Saisons	Printemps	272
	Eté	279
	Automne	187
	Hiver	85

Table 6.2 – Liste et nombre d'observations des variables qualitatives utilisées.

Feuille	Paramètre de forme	Médiane empirique	Médiane théorique	Moyenne empirique	Moyenne théorique
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 6.3 – Médiane et moyenne empirique et théorique pour chaque feuille de l'arbre

6.2.2 Application de la méthode

Nous obtenons donc, avec cette méthode, une classe avec une distribution pour chaque événement, qui est notre prior dans notre approche bayésienne. Nous nous plaçons à l'échelle de la commune et nous appliquons la théorie de la crédibilité bayésienne, comme décrite précédemment. L'association de ces deux méthodes constitue le caractère novateur de notre approche. La classification CART GPD est un préalable. Nous commençons par effectuer la classification CART GPD pour trouver les classes et les distributions, puis nous appliquons la crédibilité.

Nous cherchons $Y_{i,j}$ le coût total pour la commune i sachant qu'elle fait partie d'un événement de type j, j étant la classe du CART.

Nous connaissons l'historique $(Y_{i,j,1}, ...Y_{i,j,k}, ...Y_{i,j,n})$ des événements passés pour une commune i d'un type j. Cet historique correspond à la base d'apprentissage de l'arbre, mais en se plaçant à l'échelle de la commune.

Or grâce à la classification CART GPD nous connaissons la distribution du coût des événements de type j. On suppose ensuite que :

$$Y_{i,j} \sim \mathbb{GPD}(\gamma_j, s_j),$$

avec $s_j = p_i\sigma_j$, où p_i est la proportion des primes de la commune par rapport au total des primes de l'événement et σ_j et γ_j les paramètres de la GPD de l'événement. On suppose donc que le coût est distribué uniformément dans un événement selon la répartition des primes. Les primes provenant aussi de la BD SILECC, cette hypothèse paraît acceptable, même si cela apporte une source d'erreur potentielle supplémentaire.

Ensuite grâce aux résultats de l'introduction nous pouvons en déduire que :

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \frac{\sum_{k=1}^{n} y_{i,k} + (\frac{pi\sigma_j}{\gamma_j})}{n + \frac{1}{\gamma_j} - 1}.$$

Empiriquement, nous pouvons constater que les coûts des communes pour chaque classe j sont peu corrélés entre eux. Ainsi, l'utilisation des coûts passés par feuille semble pertinente en lien avec la classification, comme le montre le tableau 6.4.

Feuille	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 6.4 – Coefficient de corrélation de Pearson pour le coût empirique des communes dans chaque feuille. On compare les moyennes des coûts dans les mêmes communes mais dans des feuilles différentes

Sur la base de test de la section 6.4, nous constatons que près de la moitié des communes que nous essayons d'estimer n'ont pas d'historique de sinistres ($n = 0$), et seulement 20% des communes ont plus d'un événement passé ($n > 1$). Cette méthode est donc particulièrement adaptée, car l'expérience des événements passés n'est pas suffisante pour fournir une estimation fiable. L'information apportée par les classes du CART GPD permet d'enrichir l'estimation dans ces cas.

En résumé, le jour J, nous avons un événement qui correspond à une liste de communes impactées. Nous calculons les variables d'entrée pour cet événement, qui seront utilisées pour la classification CART GPD. Ensuite, pour chaque commune, nous pouvons estimer un coût total en fonction de la classe obtenue. Le coût total correspond à la somme des coûts des communes. Le coût total dépend donc de la liste de communes impactées, de l'historique des coûts de ces communes, ainsi que du type d'événement. Cette approche est précieuse pour notre étude, car nous prenons en compte les particularités locales des communes, ainsi que l'événement dans son ensemble. De plus, nous exploitons pleinement la méthode CART GPD en utilisant sa distribution de sortie comme a priori sur le profil de risque. En fin de compte, nous estimons le coût d'un événement en cours de la manière suivante :

$$\hat{C} = \sum_{i=1}^{M} \mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}].$$

6.3 Estimation fondée sur une approche type fréquence x sévérité

Nous avons également développé une autre méthode reposant sur des indicateurs à l'échelle de l'événement, qui est un peu plus facile à interpréter. Cette méthode nous permet également d'avoir un point de comparaison. Il s'agit d'une approche de type fréquence-sévérité, qui est classique en actuariat. Dans cette approche, nous estimons la fréquence et la sévérité d'un événement en les comparant à des événements passés, puis nous multiplions cette fréquence par l'exposition observée.

Pour mesurer l'exposition, nous utilisons la carte de ruissellement précédemment décrite. Pour mesurer la fréquence passée, nous calculons un taux de sinistralité en divisant le nombre de sinistres par le nombre de biens. La sévérité est mesurée par le coût moyen d'un sinistre à l'échelle communale.

En multipliant ces trois quantités, nous obtenons un coût. Afin d'introduire de la variabilité et d'éliminer une part de la sinistralité résiduelle, nous répartissons les mesures de sévérité et de fréquence sur des échantillons différents. À la fin, nous obtenons une table d'événements historiques de grande ampleur, comprenant plusieurs mesures rendant compte de la sévérité et de la fréquence des sinistres à l'échelle des communes pour un événement. Le jour J, nous disposons d'une liste de communes pour lesquelles nous pouvons calculer l'exposition. Ensuite, nous consultons la table de référence afin de trouver l'événement passé le plus similaire et l'échantillon qui a donné les meilleurs résultats. Nous utilisons ensuite les mesures de sévérité et de fréquence correspondantes que nous appliquons à l'exposition calculée. Pour améliorer la précision, cette analyse est réalisée pour chaque zone de risques inondations de la cartographie présentée à la Figure 4.6. La comparaison avec les événements passés est effectuée directement par le gestionnaire de risque en utilisant divers critères tels que la zone touchée, la saison, le type de crue et les communes importantes touchées. Cette approche laisse une certaine part de contrôle et permet une construction « à la main » du coût. Cependant, cette méthode est vulnérable à différents biais et il est également difficile de l'évaluer préalablement.

Nous allons détailler cette méthode, tout d'abord, nous cherchons à mesurer l'exposition pour chaque zone de la cartographie des inondations.

Exposition Le jour J, lorsqu'un événement touche N communes, nous calculons l'exposition pour chaque zone de risques aux inondations (représentées par l'indice i) ainsi que pour chaque commune (représentées par l'indice j). Pour rappel, nous avions défini 5 zones qui définissent un facteur de risque. Dans chaque zone, nous comptons le nombre de professionnels, que nous notons N_pro_i. Cependant, nous n'avons pas le détail du nombre de logements individuels par zone, nous devons donc le calculer en utilisant les données de la zone urbaine de l'INSEE. En croisant les informations de la zone urbaine avec la cartographie des zones de risques, nous pouvons déterminer la proportion de zone urbaine dans chaque zone de risques, que nous appelons P_ZU_i. Grâce à cette proportion, nous pouvons estimer le nombre de logements individuels par zone en multipliant le nombre total de logements par cette proportion, que nous notons N_ind. Ainsi, nous avons l'estimation suivante :

$$N_{ind_i} = P_{ZU_i} N_{ind}.$$

On peut ainsi calculer le nombre de biens exposé par zone de risques dans une commune j:

$$N_{tot_{i_j}} = N_{pro_{i_j}} + N_{ind_{i_j}} = N_{pro_{i_j}} + P_{ZU_{i_j}} N_{ind_j}.$$

Ensuite en sommant sur toutes les communes N on obtient une mesure du nombre de biens exposés à l'échelle de l'événement pour chaque zone i:

$$N_{tot_i} = \sum_{j=1}^{N} N_{tot_{i_j}}.$$

Sévérité Pour mesurer la sévérité, nous examinons les événements passés de la base de données SILECC. Étant donné que la liste des communes touchées peut différer, supposons que nous ayons maintenant N' communes pour l'événement étudié. Nous allons également mesurer la sévérité pour chaque zone de risque i et chaque commune j'. Pour ce faire, nous calculons la moyenne arithmétique des coûts des sinistres dans chaque zone et chaque commune, de la manière suivante :

$$M_{i'_j} = \overline{C_{loc_{i_{j'}}}},$$

avec $C_{loc_{i_{j'}}}$ le coût d'un sinistre rapporté et localisé dans la BD SILECC pour la commune j' et le facteur de risque i. Pour déterminer la zone de risque, nous avons préalablement géolocalisé les sinistres afin de les croiser avec les données correspondantes. Cependant, cela entraîne des pertes d'informations, car les adresses des sinistres ne sont pas toujours renseignées et la précision de localisation peut être insuffisante. Ainsi, nous utilisons un sous-échantillon des sinistres reçus pour lesquels nous avons réussi à obtenir une localisation précise, que nous appelons C_{loc}. Ensuite, pour obtenir une mesure de la sévérité à l'échelle de l'événement par zone de risque, nous calculons la moyenne des coûts moyens.

$$S_i = \overline{M_{i_{j'}}} = \frac{1}{N'} \sum_{j'=1}^{N'} M_{i_{j'}}.$$

Fréquence Pour mesurer la fréquence, nous restons dans le même périmètre que pour la sévérité, c'est-à-dire les N' communes et chaque facteur de risque i. Nous cherchons à calculer un taux de sinistralité pour chaque commune. Le nombre de sinistres est donc déterminant, et par conséquent, nous prenons également en compte les sinistres que nous n'avons pas réussi à localiser. En effet, ces sinistres ne sont pas localisés avec suffisamment de précision mais dans la plupart des cas, nous avons l'information sur la commune où ils se sont produits. Nous pouvons donc, de la même manière que pour la population, les attribuer à une zone de risque en utilisant la zone urbaine, avec N_{nloc} représentant le nombre de sinistres non localisés :

$$N_{nloc_i} = P_{ZU_i} N_{nloc}.$$

On peut ainsi calculer le nombre de sinistres, avec $N_{loc_{i_{j'}}}$ le nombre de sinistres localisés en zone de risque i pour une commune j' :

$$N_{sin_{i_{j'}}} = N_{loc_{i_{j'}}} + N_{nloc_{i_{j'}}}.$$

On peut ensuite calculer un taux de sinistralité T en divisant par le nombre de biens exposés, calculé comme décrit précédemment, dans chaque commune :

$$T_{i_{j'}} = \frac{N_{sin_{i_{j'}}}}{N_{tot_{i_{j'}}}}.$$

Pour en faire une mesure de la fréquence à l'échelle de l'événement on prend la moyenne géométrique :

$$F_i = \overline{T_{i_{j'}}} = \prod_{j'=1}^{N'} T_{i'_j}^{\frac{1}{N'}}.$$

Échantillons différents Nous obtenons donc deux mesures pour chaque zone de risque à l'échelle de l'événement : une pour la sévérité et une pour la fréquence. Pour faire varier les coûts, nous pouvons calculer S_i et F_i sur des échantillons différents. En effet, afin d'éliminer la sinistralité "résiduelle" et de nous concentrer sur les communes d'intérêt, nous appliquons un filtre. Nous identifions les communes qui concentrent 90%, 92,5%, 95%, 97,5%, 99% et 100% de la sinistralité de l'événement, puis nous calculons nos indicateurs sur ces sous-échantillons. Cela permet d'exclure un certain nombre de communes ayant peu de sinistres, ce qui peut affaiblir la robustesse de nos indicateurs car ils sont fortement influencés par le nombre de sinistres.

On cherche donc

$$N'_{x\%} \text{ tel que } \sum_{j'=1}^{N'_{x\%}} CT_{j'} = 0.xCT,$$

avec $CT_{j'}$ représentant le coût total des sinistres pour la commune j' et CT le coût total de l'événement, nous classons les communes par ordre décroissant en fonction du coût pour privilégier celles qui contribuent le plus. Cela nous permet d'obtenir plusieurs sous-échantillons de communes pour chaque événement, sur lesquels nous calculons nos indicateurs. Ainsi, nous obtenons 6 échantillons différents : $N'90\%$, $N'92,5\%$, $N'95\%$, $N'97,5\%$, $N'99\%$, $N'100\%$. Chaque échantillon nous fournit des indicateurs différents, ce qui nous permet de choisir la méthode la plus adaptée en fonction de la nature de l'événement.

Table de référence Nous calculons ces indicateurs uniquement pour certains événements de grande ampleur, et non pour l'ensemble des événements de la base SILECC. L'objectif est d'avoir un nombre limité d'événements le jour J, afin de pouvoir rapidement identifier celui qui est similaire à l'événement en cours. Pour cela, nous utilisons les événements de grande ampleur rapportés par CCR dans son espace professionnels, ce qui nous permet d'avoir une base de test, comme décrit dans la section suivante. Nous faisons correspondre ces 56 événements avec 56 événements de notre base SILECC. Les événements sont similaires, mais avec un périmètre différent. En effet, notre méthode diffère de celle de CCR, ce qui entraîne des listes de communes différentes pour des événements "identiques". Pour ces 56 événements en se basant sur les communes SILECC nous construisons les indicateurs de sévérité et de fréquence selon les 6 sous-échantillons. Pour chacun de ces événements nous allons ensuite calculer le coût avec chaque sous-échantillon. Nous commençons par calculer l'exposition selon les communes de l'événement CCR, identique pour tous les sous-échantillons. Nous prenons la méthode décrite au début de cette section en utilisant les communes rapportées par CCR, on obtient un N_{tot_i}. On calcule ensuite en prenant les communes rapportées dans SILECC, $S_{i_{90\%}}$, ...$S_{i_{100\%}}$ et

$F_{i_{90\%}}$, ...$F_{i_{100\%}}$. Finalement, le coût pour un sous-échantillon et dans une zone de risque est :

$$C_{i_{x\%}} = N_{tot_i} \times S_{i_{x\%}} \times F_{i_{x\%}}.$$

et l'on peut ensuite sommer sur toutes les zones de risques pour avoir le coût total par échantillon :

$$C_{x\%} = \sum_{i=1}^{5} C_{i_{x\%}}.$$

Le jour J, un événement survient dans N communes. Nous calculons le nombre total de biens exposés, noté N_{toti}, dans ces N communes. Ensuite, nous cherchons l'événement de référence le plus similaire à cet événement en utilisant les critères choisis. Une fois que nous avons identifié l'événement le plus proche, nous examinons l'échantillon x qui fournit un coût $Cx\%$ le plus proche du coût réel de cet événement. Étant donné que tous ces événements sont passés, nous disposons d'un coût réel rapporté pour l'ensemble du marché, que nous avons préalablement actualisé en fonction de l'indice de la FFB du coût de la construction (de la même façcon que pour les coûts des sinistres).

Ensuite, nous utilisons les indicateurs $S_{i_{x\%}}$ et $F_{i_{x\%}}$ associés à cet événement de référence, c'est-à-dire les indicateurs qui ont permis d'obtenir le coût $C_{x\%}$ dans la base de référence, et nous les appliquons à l'événement en cours. Ainsi, si nous appelons S_{ref} et F_{ref} les indicateurs de l'événement de référence le plus proche de l'événement en cours, et avec l'échantillon fournissant le meilleur ajustement du coût total, nous pouvons alors utiliser ces indicateurs pour estimer :

$$C = \sum_{i=1}^{5} N_{tot_i} \times S_{ref_i} \times F_{ref_i}.$$

6.4 Discussion des résultats

Pour évaluer et comparer nos méthodes, nous les avons utilisées pour estimer des événements en utilisant la base des événements rapportés par CCR. Cette base contient les informations sur 56 événements, y compris les communes touchées et les coûts totaux pour l'ensemble du marché. Ces événements sont déjà présents dans notre base de données, mais avec des communes différentes. En effet, la définition d'un événement peut varier et plusieurs méthodes peuvent conduire à des événements différents. Ces événements sont de grande ampleur, ce qui se traduit par des coûts significatifs. Un résumé des variables d'intérêt est présenté dans le tableau 6.5. Nous avons sélectionné uniquement les événements postérieurs à 2008 afin de garantir que notre base de données soit aussi représentative que possible. Contrairement à la sécheresse, où nous travaillons sur des coûts à l'échelle de la commune et des événements locaux, nous sommes ici plus sensibles à la répartition des sinistres et des portefeuilles.

Variable	Min	1$^{\text{er}}$Q	Médiane	Moyenne	3$^{\text{ème}}$Q	Max
Coût	10 710 000	16 110 000	35 680 000	116 900 000	98 270 000	1 056 000 000

Table 6.5 – Résumé du coût des événements étudiés

Nous disposons d'une liste de communes sur laquelle nous pouvons calculer nos paramètres d'entrée pour la classification, ainsi qu'une liste d'événements de référence pour les indicateurs de la deuxième méthode.

MODELE	MAE
$C_{90\%}$	162 462 546
$C_{92.6\%}$	128 057 050
$C_{95\%}$	102 094 036
$C_{97.5\%}$	74 005 296
$C_{99\%}$	56 513 672
$C_{100\%}$	88 538 340
Méthode crédibilité GPD	80 996 208

Table 6.6 – Comparaison de la MAE des différentes méthodes

Nous avons calculé l'erreur absolue moyenne (MAE) pour chaque méthode. On peut déjà remarquer que les MAE sont assez élevées. Cela s'explique par le fait que les méthodes ont du mal à estimer l'événement d'une valeur d'un milliard, ce qui augmente considérablement la moyenne des erreurs. Étant donné que nous disposons seulement de 56 événements, la MAE est très sensible aux variations individuelles, en particulier celles de grande ampleur.

Ensuite, on peut constater que la méthode basée sur la crédibilité donne des résultats comparables aux autres méthodes.

Les meilleures méthodes sont $C_{99\%}$ et $C_{97.5\%}$, ce qui confirme l'idée que les sinistres de faible importance peuvent compromettre la qualité des indicateurs considérés. On remarque également que les prédictions les plus éloignées le sont pour toutes les méthodes. Cela pourrait être lié à un problème de représentativité globale de notre base de données. En effet, notre base peut ne pas prendre en compte certaines spécificités qui, à l'échelle d'un événement, peuvent avoir un impact significatif sur la qualité des prédictions. Ici nous comparons la MAE de chaque méthode mais dans l'utilisation prévue, nous utiliserons que la méthode qui correspond le mieux. C'est un avantage car cela laisse une part d'interprétation et permet de profiter de l'expertise métier de la MRN sur les événements en cours. Cependant cela limite aussi car si l'événement diffère trop de la base de référence alors nous ne pourrons pas l'estimer. Dans ce cas la méthode se basant sur le CART GPD est plus adaptée.

Nous obtenons des résultats encourageants avec les deux méthodes, qui seront utiles à la fédération dans son processus de gestion. C'est en effet dans une logique d'aide à la décision que ces méthodes ont été créées. La principale incertitude de ces méthodes et l'une des difficultés réside dans la détermination de la liste des communes impactées le jour J. Dans ce contexte, l'expertise métier de la MRN est essentielle car tout repose sur cette liste. Enfin, ces méthodes pourraient certainement être améliorées en utilisant des informations sur l'intensité de l'aléa, telles que des variables météorologiques. En effet, dans notre approche actuelle, nous nous limitons à des variables qui décrivent l'exposition et l'ampleur des événements, sans prendre en compte l'intensité de ces événements. L'intégration de cette information représente une perspective prometteuse pour affiner encore les estimations.

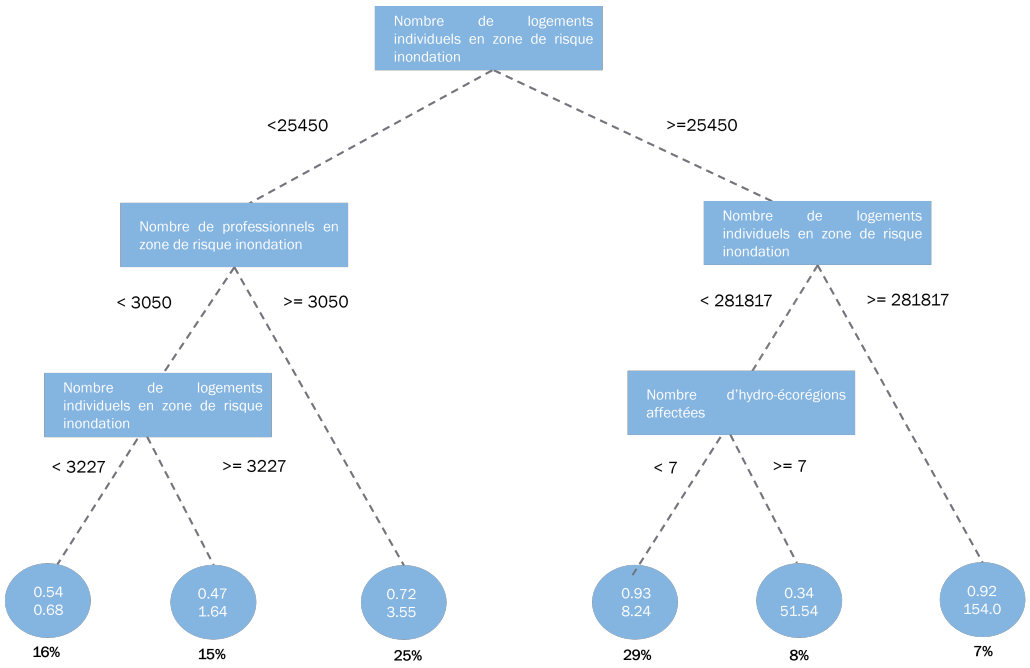
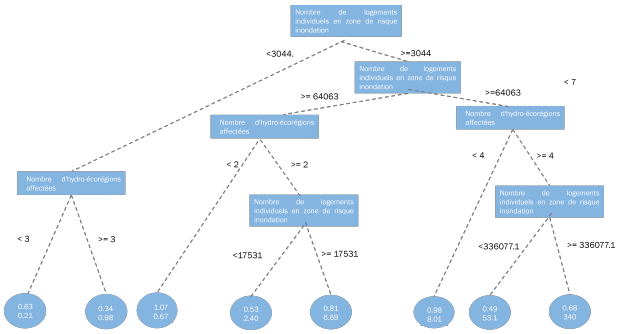
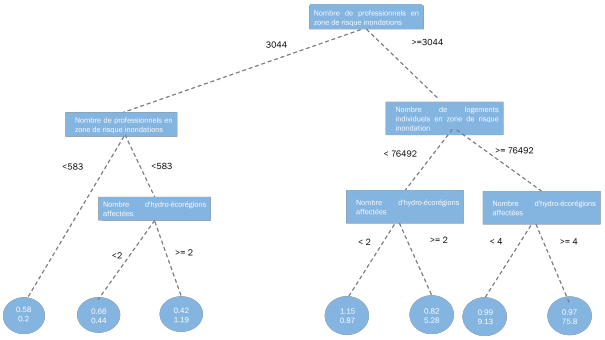


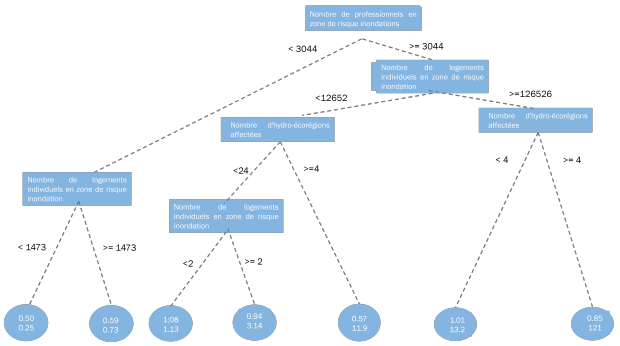
Figure 6.2 – Arbre de régression GPD obtenue pour les événements inondations. Pour chaque feuille on indique le paramètre de forme γ (première ligne), le paramètre d'échelle σ à 10^{-5} (deuxième ligne). Les pourcentages d'observations dans chaque feuille sont aussi présentés.



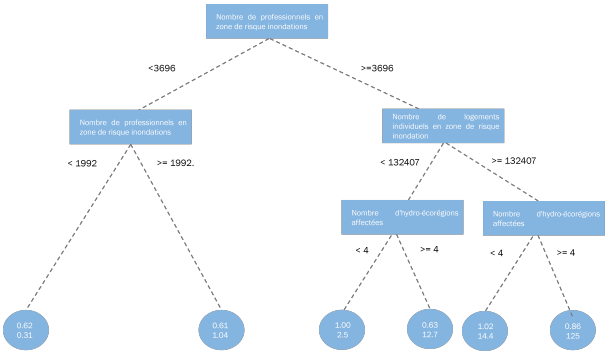
a) 2 000 plus grandes observations



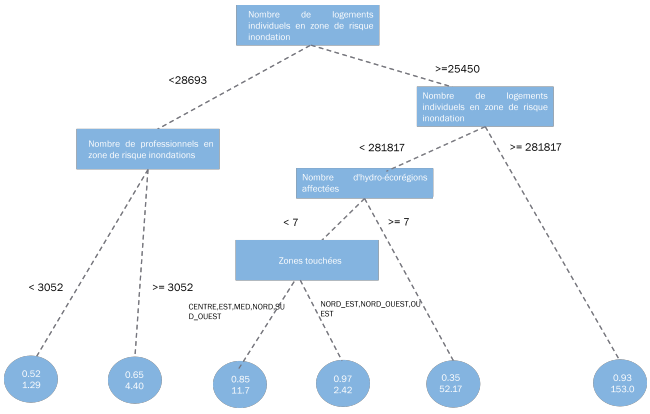
b) 1 750 plus grandes observations



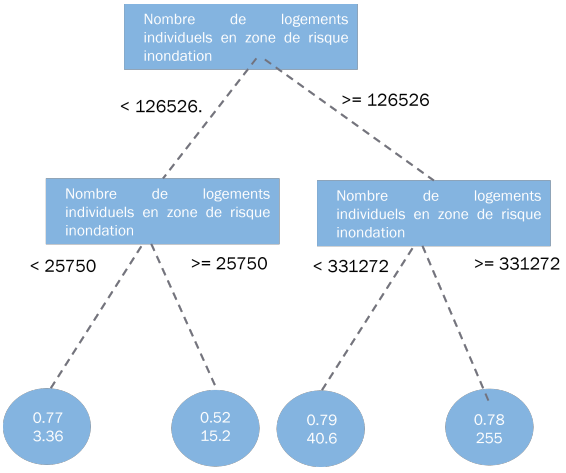
c) 1 500 plus grandes observations



d) 1 250 plus grandes observations



e) 750 plus grandes observations



f) 500 plus grandes observations

Figure 6.3 – Arbres obtenus pour le CART GPD en utilisant les plus grandes observations, on fait varier le nombre de 2 000 à 500 avec un pas 250 pour illustrer la sensibilité. Pour chaque feuille on donne une estimation du γ.

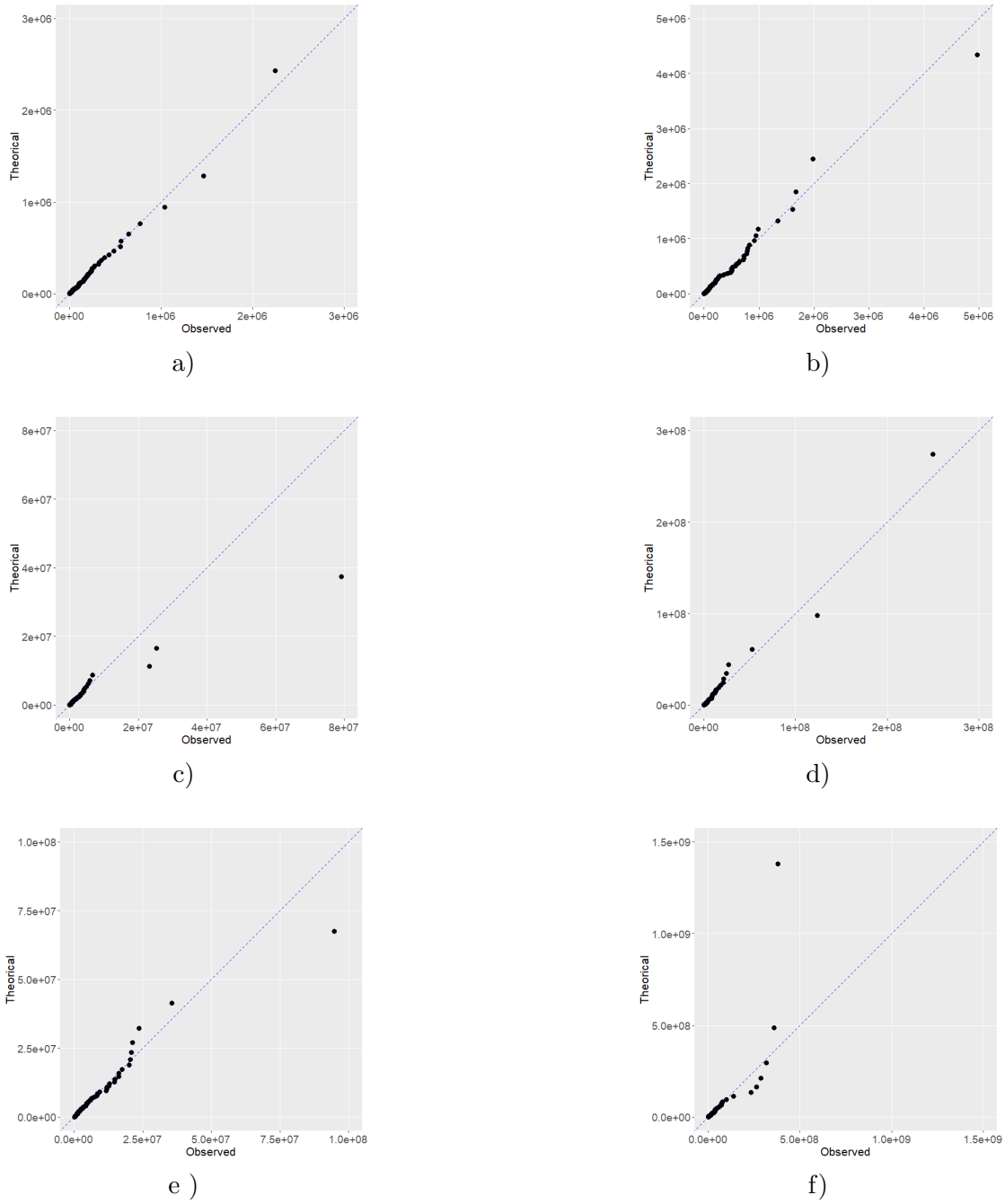


Figure 6.4 – Diagrammes quantile-quantile pour chaque feuille de l'arbre

Conclusion

Dans ce mémoire, nous avons introduit une méthode permettant d'estimer les coûts liés aux événements extrêmes en tenant compte à la fois de l'historique des données, même partiel, et de diverses covariables associées. Cette approche est particulièrement adaptée au contexte de la gestion des risques, car elle permet d'élaborer des estimations de coûts tout en mettant à disposition des informations cruciales à chaque étape du processus. Au-delà de la simple prévision du coût, notre méthode propose une analyse détaillée de la distribution des événements extrêmes.

Pour ce faire, nous avons mobilisé deux théories mathématiques : la théorie des valeurs extrêmes et la théorie de la crédibilité bayésienne. Nous avons montré comment ces deux approches peuvent être conjuguées pour la tarification des risques. Pour estimer le "prior", c'est-à-dire la connaissance préalable des risques, nous avons également présenté une procédure s'appuyant sur l'utilisation de covariables, laquelle offre des résultats théoriques. En effet, il est possible de démontrer la consistance de cette démarche qui, en comparaison avec d'autres méthodes de régression des valeurs extrêmes, présente l'avantage de pouvoir capturer des discontinuités dans la fonction de régression.

Tout au long de ce mémoire, nous avons exploré les multiples facettes de notre méthodologie, depuis ses fondements théoriques jusqu'à ses applications concrètes.

En effet nous avons proposé deux applications de cette méthode de tarification. La première application concerne le risque cyber tandis que la seconde application porte sur l'estimation rapide du coût des inondations rapidement après leur occurrence. La partie consacrée aux catastrophes naturelles a été plus développée, étant donné notre collaboration avec la Mission Risques Naturels, un groupement technique de France Assureurs. Néanmoins l'application au risque cyber démontre la capacité de généralisation de notre méthode. Nous avons illustré comment, même avec des données limitées, il est possible de proposer un tarif cohérent pour évaluer un risque.

Les résultats pour le risque cyber sont intéressants, mais une analyse plus solide pourrait être réalisée en se basant sur des données réelles. En effet, avec la base de données utilisée, plusieurs incertitudes subsistent, notamment en ce qui concerne les coûts. Les résultats sont donc à prendre avec précautions. Notre méthode pourrait se révéler utile pour les compagnies d'assurance, car elle tient compte de l'hétérogénéité présente dans les données. Une application visant à élaborer une police d'assurance cyber en se basant sur les sinistres passés pourrait à ce titre être très intéressante.

Comme exposé dans l'introduction, une augmentation des coûts liés aux événements naturels est à prévoir dans les années à venir, en raison de l'accroissement des biens et des effets du changement climatique. Afin de maintenir un niveau élevé de couverture des dommages par le biais de l'assurance, la réduction des coûts associés aux catastrophes naturelles représente un enjeu crucial. Atteindre cet objectif passe par une amélioration de la compréhension et de la prévention des risques. Dans cette optique, nous avons cherché à y contribuer en approfondissant la connaissance des événements naturels, plus particulièrement de leurs coûts. En effet, notre approche concernant les inondations

nous offre une classification des événements avec une focus sur les extrêmes. Cette catégorisation permet de repérer les territoires potentiellement vulnérables.

Pour les inondations une piste d'amélioration majeure réside dans l'incorporation d'informations relatives à l'intensité de l'aléa, en particulier, en utilisant des variables météorologiques. À l'heure actuelle, notre étude repose exclusivement sur des variables dénotant l'exposition et l'ampleur des événements, mais sans informations sur leurs intensité. Bien que l'application de telles informations à l'échelle de chaque événement puisse présenter une certaine complexité, elles constituent l'une des principales perspectives pour affiner nos estimations. Une application pertinente consisterait à étudier l'influence des variables météorologiques sur le coût en exploitant cette méthode. Cela nous permettrait, par exemple, d'évaluer de manière approfondie l'évolution potentielle du coût en fonction de divers scénarios établis par le GIEC.

D'un point de vue académique, les travaux sur les arbres de régression ouvrent des perspectives intéressantes pour l'étude des événements extrêmes. Cette méthode, tout en préservant une bonne interprétabilité, peut être appliquée à une multitude de contextes. Cependant, il est à noter que les arbres de régression peuvent parfois présenter une certaine instabilité. Dans cette optique, il pourrait être pertinent d'explorer comment d'autres approches, telles que les forêts aléatoires, pourraient être employées pour aborder la régression des valeurs extrêmes. L'application de la théorie de la crédibilité à notre problématique offre également un champ d'investigation novateur en permettant de modéliser le coût au niveau de chaque commune en fonction d'un profil de risque extrême lié à l'événement qui l'affecte. Cette approche pourrait être étendue à d'autres types de risques et à des ensembles de données différents. L'association de ces deux méthodes, pour l'analyse de la régression, pourrait donc être sujette à des approfondissements.

En outre, il reste toujours pertinent de confronter les modèles mathématiques et les algorithmes d'apprentissage statistique aux données du monde réel. Dans ce mémoire, nous avons appliqué ces méthodes aux données provenant du domaine de l'assurance, évaluant ainsi les prédictions sur des scénarios concrets.

D'un point de vue plus global, notre méthode de tarification offre une perspective exhaustive et novatrice pour l'évaluation des risques extrêmes. Cette approche tient compte des défis liés à la disponibilité limitée des données, partiellement adressés par notre approche bayésienne, tout en s'adaptant à la nature exceptionnellement rare de ces événements. Cette méthodologie pourrait jouer un rôle dans la tarification des risques émergents. Par conséquent, explorer son application pour évaluer les risques liés à la cyber, ainsi que pour les événements climatiques, au sein du secteur de l'assurance et dans d'autres domaines pertinents, semble être une démarche intéressante, avec un potentiel réel pour générer des avantages significatifs.

Bibliography

Allen, D. M. (1974), 'The relationship between variable selection and data agumentation and a method for prediction', *Technometrics* **16**(1), 125–127.

Anais, M. (2019), 'Titre: Modélisation assurantielle du risque cyberx'.

Antonio, K. & Plat, R. (2014), 'Micro-level stochastic loss reserving for general insurance', *Scandinavian Actuarial Journal* **2014**(7), 649–669.

Balkema, A. A. & De Haan, L. (1974), 'Residual life time at great age', *The Annals of probability* **2**(5), 792–804.

Bastard, T. (2021), 'Titre: Modélisation du risque cyber de perte de données à caractère personnel, modèle de tarification, inclusion dans le bgs et proposition de scénarios de stress pour l'orsa.'.

Beirlant, J. & Goegebeur, Y. (2003), 'Regression with response distributions of pareto-type', *Computational statistics & data analysis* **42**(4), 595–619.

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

Bidan, P. & Cohignac, T. (2017), 'Le régime francais des catastrophes naturelles: historique du régime, variances, 11'.

Boucheron, S. & Thomas, M. (2015), 'Tail index estimation, concentration and adaptivity'.

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

Bousquet, N. & Bernardara, P. (2021), *Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice*, Springer Nature.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Brodin, E. & Rootzén, H. (2009), 'Univariate and bivariate gpd methods for predicting extreme wind storm losses', *Insurance: Mathematics and Economics* **44**(3), 345–356.

Bühlmann, H. & Gisler, A. (2005), *A course in credibility theory and its applications*, Vol. 317, Springer.

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

Charpentier, A., Barry, L. & James, M. R. (2021), 'Insurance against natural catastrophes: balancing actuarial fairness and social solidarity', *The Geneva Papers on Risk and Insurance - Issues and Practice* .

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

Chemitte, J. (2008), Adoption des technologies de l'information géographique et gestion des connaissances dans les organisations. Application à l'industrie de l'assurance pour la gestion des risques naturels, PhD thesis, École Nationale Supérieure des Mines de Paris.

Chneiweiss, A. & Bardaji, J. (2020), *Les assureurs face au défi climatique*, Fondapol, Fondation pour l'innovation politique.

Circulaire numéro 91-50 du 12 février 1991 (1991), Technical report, Ministère de la Transition écologique.

Climat, cyber, pandémie : le modèle assurantiel mis au défi des risques systémiques (2022), Technical report, Conseil Économique, Social et Environnemental.

Coles, S., Bawa, J., Trenner, L. & Dorazio, P. (2001), *An introduction to statistical modeling of extreme values*, Vol. 208, Springer.

Cost of Cyber Incidents Study (2020), Technical report, Cybersecurity and Infrastructure Security Agency's.

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds', *Journal of the Royal Statistical Society: Series B (Methodological)* **52**(3), 393–425.

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

Embrechts, P., Klüppelberg, C. & Mikosch, T. (2013), *Modelling extremal events: for insurance and finance*, Vol. 33, Springer Science & Business Media.

Etat de la menace rançongiciel (2021), Technical report, Agence Nationale de la Sécurité des Systèmes d'Information.

Etude : Changement climatique et assurance à l'horizon 2040 (2021), Technical report, France assureurs.

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

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

Fisher, R. A. & Tippett, L. H. C. (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 24, Cambridge University Press, pp. 180–190.

Fréchet, M. (1927), 'Sur la loi de probabilité de l'écart maximum', *Ann. Soc. Math. Polon.* **6**, 93–116.

Gérin, S. (2011), Une démarche évaluative des Plans de Prévention des Risques dans le contexte de l'assurance des catastrophes naturelles: Contribution au changement de l'action publique de prévention, PhD thesis, Université Paris-Diderot-Paris VII.

Gey, S. & Nedelec, E. (2005), 'Model selection for cart regression trees', *IEEE Transactions on Information Theory* **51**(2), 658–670.

Gnedenko, B. (1943), 'Sur la distribution limite du terme maximum d'une série aléatoire', *Annals of mathematics* pp. 423–453.

Goussebaile, A. (2016), Prevention and insurance of natural disasters, PhD thesis, Université Paris-Saclay (ComUE).

Guillier, F. (2017), Evaluation de la vulnérabilité aux inondations: Méthode expérimentale appliquée aux Programmes d'Action de Prévention des Inondations, PhD thesis, Paris Est.

Guillou, A. & Willems, P. (2006), 'Application de la théorie des valeurs extrêmes en hydrologie', *Revue de statistique appliquée* **54**(2), 5–31.

Gumbel, E. J. (1958), *Statistics of extremes*, Columbia University Press.

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

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

Heranval, A. (2022), Contributions of insurance data to the study of natural disasters, PhD thesis, Sorbonne Université.

Hillairet, C. & Lopez, O. (2022), 'Cyber-assurance : enjeux, modélisations et leviers de mutualisation'.

Horton, J. B. (2018), 'Parametric insurance as an alternative to liability for compensating climate harms', *Carbon & Climate Law Review* **12**(4), 285–296.

Inondations, s'informer pour mieux se protéger (2019), Technical report.

Insurance 2020 beyond: Reaping the dividends of cyber resilience (2020), Technical report, PwC.

Internet Organized Crime Threat Assessment (2018), Technical report, EUROPOL.

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

Katz, R. W., Parlange, M. B. & Naveau, P. (2002), 'Statistics of extremes in hydrology', *Advances in water resources* **25**(8-12), 1287–1304.

Klein, R. W. & Wang, S. (2009), 'Catastrophe Risk Financing in the United States and the European Union: A Comparative Analysis of Alternative Regulatory Approaches', *Journal of Risk and Insurance* **76**(3), 607–637.

Lallie, H. S., Shepherd, L. A., Nurse, J. R., Erola, A., Epiphaniou, G., Maple, C. & Bellekens, X. (2021), 'Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic', *Computers & security* **105**, 102248.

L'assurance des événements naturels en 2020 (2022), Technical report, France Assureurs.

Latruffe, L. & Picard, P. (2005), 'Assurance des catastrophes naturelles: faut-il choisir entre prévention et solidarité?', *Annales d'économie et de statistique* pp. 33–56.

Lelarge, M. & Bolot, J. (2009), Economic incentives to increase security in the internet: The case for insurance, *in* 'IEEE INFOCOM 2009', IEEE, pp. 1494–1502.

Lettre d'information de la Mission Risques Naturels 30 (2019), Technical report, Mission Risques Naturels.

Lettre d'information de la Mission Risques Naturels 34 (2020), Technical report, Mission Risques Naturels.

Lettre d'information de la Mission Risques Naturels 36 (2021), Technical report, Mission Risques Naturels.

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

Lin, X. & Kwon, W. J. (2020), 'Application of parametric insurance in principle-compliant and innovative ways', *Risk Management and Insurance Review* **23**(2), 121–150.

LOI numéro 82-600 du 13 juillet 1982 relative à l'indemnisation des victimes de catastrophes naturelles (1982).

Lopez, O. (2023), 'Insurability of large emerging risks'.

Lopez, O. & Thomas, M. (2023), 'Parametric insurance for extreme risks: the challenge of properly covering severe claims', *arXiv preprint arXiv:2301.07776* .

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

Moncoulon, D. (2014), 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.

Moncoulon, D., Labat, D., Ardon, J., Leblois, E., Onfroy, T., Poulard, C., Aji, S., Rémy, A. & Quantin, A. (2014), '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.

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

Moore, I. D., Grayson, R. & Ladson, A. (1991), 'Digital terrain modelling: a review of hydrological, geomorphological, and biological applications', *Hydrological processes* **5**(1), 3–30.

Mowbray, A. H. (1914), How extensive a payroll exposure is necessary to give a dependable pure premium, *in* 'Proceedings of the Casualty Actuarial society', Vol. 1, pp. 24–30.

Norberg, R. (1993), 'Prediction of outstanding liabilities in non-life insurance', *ASTIN Bulletin: The Journal of the IAA* **23**(1), 95–115.

Norberg, R. (1999), 'Prediction of outstanding liabilities ii. model variations and extensions', *ASTIN Bulletin: The Journal of the IAA* **29**(1), 5–25.

Peyrat (2023), 'Titre: Risque cyber, un modèle épidémiologique sur réseaux pour le risque d'accumulation du cyber silencieux'.

Pickands III, J. (1975), 'Statistical inference using extreme order statistics', *the Annals of Statistics* pp. 119–131.

Pigeon, M., Antonio, K. & Denuit, M. (2014), 'Individual loss reserving using paid–incurred data', *Insurance: Mathematics and Economics* **58**, 121–131.

Rapport au Ministre de l'économie, des finances et de la souveraineté industrielle et numérique sur le régime d'indemnisation des catastrophes naturelles (2022), Technical report, Caisse centrale de réassurance.

Référentiels de résilience du bâti aux aléas naturels (2022), Technical report, Mission Risques Naturels.

Resnick, S. I. (1997), 'Discussion of the danish data on large fire insurance losses', *ASTIN Bulletin: The Journal of the IAA* **27**(1), 139–151.

Resnick, S. I. (2007), *Heavy-tail phenomena: probabilistic and statistical modeling*, Springer Science & Business Media.

Rootzén, H. & Tajvidi, N. (1997), 'Extreme value statistics and wind storm losses: a case study', *Scandinavian Actuarial Journal* **1997**(1), 70–94.

Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti (2018), Technical report, Mission Risques Naturels.

Smith, J. A. (1987), 'Estimating the upper tail of flood frequency distributions', *Water Resources Research* **23**(8), 1657–1666.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the royal statistical society: Series B (Methodological)* **36**(2), 111–133.

Surminski, S. & Thieken, A. H. (2017), 'Promoting flood risk reduction: The role of insurance in germany and england', *Earth's Future* **5**(10), 979–1001.

Van Nostrand, J. M. & Nevius, J. G. (2011), 'Parametric insurance: using objective measures to address the impacts of natural disasters and climate change', *Environmental Claims Journal* **23**(3-4), 227–237.

"LUCY : LUmière sur la CYberassuranc (2022), Technical report, Association pour le Management des Risques et des Assurances de l'Entreprise.