Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 04/03/2024

Par : **Zineb Meskani**

Titre: **Modélisation de la charge du risque cyber des entreprises à partir des données externes**

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

Entreprise : Deloitte Conseil

Nom : CHALIN Cyrin

Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur de mémoire en entreprise :

Nom : ROBERT Simon

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 :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, France

Remerciements

Tout d’abord, je tiens à exprimer ma gratitude envers Madame Claude Chassain ainsi que Messieurs Baptiste Brechot et Cyril Chalin, associés Risk Advisory de Deloitte, qui m'ont offert l’opportunité d’effectuer mon stage de fin d’études au sein de leur équipe.

Mes remerciements les plus sincères vont à Monsieur Simon Robert, pour son encadrement et son suivi régulier malgré son emploi du temps chargé, ainsi que pour son soutien précieux tout au long de la rédaction de ce mémoire. J'étends mes remerciements à toute l'équipe Risk Advisory de Deloitte pour son accueil chaleureux et sa bienveillance.

Je tiens également à exprimer ma reconnaissance envers Madame Caroline Hillairet pour son aide précieuse dans l'obtention de la base de données utilisée dans ce mémoire.

Mes remerciements s'adressent également à Monsieur Antoine Heranval pour son implication en tant que tuteur académique, sa disponibilité et ses remarques pertinentes, ainsi qu'à l'ensemble des enseignants du mastère spécialisé Actuariat de l'ENSAE, particulièrement pour la qualité de la formation dispensée.

Enfin, je remercie l'ensemble de mon entourage pour leur soutien constant, du début de mes études jusqu’à la finalisation de ce mémoire.

Résumé

La numérisation de l'économie expose les entreprises à de nouveaux risques, en particulier le risque cyber, désormais considéré par de nombreuses organisations comme leur principal risque. Ce risque connaît une montée en puissance et une innovation constante. Ses conséquences peuvent être très lourdes sur toutes les dimensions de l'activité d'une entreprise, de sa réputation à ses capacités de production. La résilience face au risque cyber devient ainsi un enjeu majeur de souveraineté. Pour aider les entreprises à faire face à ce risque, le marché de la cyber-assurance se développe. Cependant, la tâche s'avère complexe en raison de la faible quantité des données disponibles et un risque en perpétuelle mutation, entravant une maîtrise et évaluation efficace de ce risque.

L'objet de ce mémoire est d'étudier la pertinence de la modélisation des coûts des sinistres cyber d'une entreprise, à partir d'une base de données SAS d'événements opérationnels. Après avoir développé une méthodologie de fouille de texte pour identifier les risques cyber dans la base de données, une analyse des charges correspondantes est réalisée. Souhaitant prendre en compte plus finement la disparité des charges des sinistres cyber, une modélisation distincte des sinistres extrêmes et attritionnels est proposée.

Les charges extrêmes sont modélisées par la GPD, tandis que des modèles prédictifs ont été préférés pour quantifier le risque cyber attritionnel d'une organisation en fonction de variables explicatives. Ces modèles comprennent des GLM et des arbres de régression CART. Le seuil de séparation entre les deux types de sinistres est déterminé en utilisant des méthodes issues de la TVE. Le risque de dépassement de ce seuil a également été modélisé via des modèles de régression logistique, et des arbres de classification CART.

Mots-clés : *Risque cyber, Risque opérationnel, Non-vie, Fouille de texte, Théorie des Valeurs Extrêmes (TVE), Peaks Over Threshold (PoT), Distribution de Pareto Généralisée (GPD), Modèle Linéaire Généralisé (GLM), Arbres de classification et de régression (CART).*

Abstract

The digital transformation of the economy exposes businesses to emerging risks, notably the escalating cyber risk, now regarded as the foremost concern by numerous organizations. This risk is dynamic and continuously evolving, potentially resulting in severe repercussions across various facets of a company's operations, ranging from its reputation to its production capabilities. The ability to withstand cyber risk has become a significant sovereignty challenge. In response to this risk, the cyber insurance market is expanding. However, the complexity of this task is heightened due to the limited availability of data for a comprehensive understanding and assessment of the risk.

Our purpose is to investigate the distribution of costs associated with cyber claims using a SAS database of operational events. After developing a text mining methodology to identify cyber risks in this database, an analysis of the corresponding costs is conducted. To address the nuanced nature of charges related to cyber claims, a distinct modeling of extreme and attritional charges has been adopted.
Tools from extreme value theory are employed to determine a threshold for separating claims. Extreme charges are modeled by the Generalized Pareto Distribution (GPD), while predictive models are developed to quantify an organization's attritional cyber risk based on explanatory variables. These models encompass Generalized Linear Models (GLM) and Classification and Regression Trees (CART). The risk of exceeding the threshold has also been modeled using logistic regression models and CART classification trees.

Keywords : *Cyber risk, Operational Risk, Non-life, Text mining, Extreme Value Theory, Peaks Over Threshold, Generalized Pareto Distribution, Generalized Linear Model, Classification and Regression Trees (CART).*

Note de synthèse

Le risque cyber englobe l'ensemble des risques liés à l'usage des technologies numériques. Aujourd'hui, il constitue l'une des menaces majeures pour les entreprises et les institutions. Les assureurs accompagnent les entreprises dans leur protection contre le risque cyber à travers des couvertures spécifiques. Cependant, le marché de la cyber-assurance peine à se développer. Le risque cyber étant récent, les assureurs ne disposent que d'un historique de sinistralité restreint. De plus, le risque cyber revêt de nombreuses formes, ce qui le rend encore mal connu et peu maîtrisé. Pour renforcer leur expertise et mieux appréhender ce risque, les assureurs pourraient vouloir exploiter des données externes.

Disposant de données SAS relatives aux pertes opérationnelles subies par des entreprises à l'international, notre objectif principal est d'évaluer la pertinence de l'utilisation de ces données externes dans la modélisation du risque cyber. Une modélisation en deux composantes majeures compose ce mémoire, la première se focalise sur la modélisation des pertes liées aux sinistres cyber dans cette base de données. La seconde partie se penche sur la modélisation de la probabilité qu'un sinistre cyber soit extrême ou non.

Identification des incidents cyber dans SAS OpRisk Global data

L'identification des sinistres cyber dans la base SAS OpRisk Global data est réalisée à l'aide d'un algorithme de fouille de texte appliqué aux descriptifs de chaque sinistre. Pour optimiser le temps de cet algorithme, nous avons préalablement traité les descriptifs. Avant d'entamer la recherche, nous avons établi une première liste étendue de mots-clés liés aux risques cyber, puis une deuxième liste plus restreinte, constituée uniquement des mots-clés les plus spécifiques au risque cyber. Un sinistre était alors identifié comme sinistre cyber s'il respectait l'une de ces 3 conditions :

1. Au moins un mot clé de la liste restreinte se trouvait dans le descriptif du sinistre
2. Au moins 10 mots de la liste globale s'y trouvaient
3. Le sinistre appartenait aux événements opérationnels pouvant être associés au risque cyber selon la classification bâloise.

Afin de valider la procédure d'identification, des échantillonnages aléatoires des sinistres ont été effectués pour confirmer leur classification réellement liée aux risques cyber. **Au total, nous avons identifié 1136 sinistres cyber.**

Modélisation de la charge des sinistres cyber

Pour les sinistres cyber ainsi identifiés, la variable cible de la modélisation est la « Current Value of Loss », qui représente le montant des pertes, en millions de dollars (M$), ajusté du CPI[1] pour prendre en compte l'inflation. L'analyse du résumé statistique des coûts des incidents cyber dans le tableau 1, et

[1] Aux États-Unis, l'indice des prix à la consommation (CPI) mesure les variations dans les prix payés par les consommateurs pour un panier de biens et de services.

des figures 1 et 2, révèle que la distribution des charges des sinistres cyber présente une queue épaisse à droite, indiquant la présence de valeurs extrêmes.

Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum	Ecart-type	Skewness	Kurtosis
0,11	0,53	2,16	31,65	10,41	3 505,35	149,07	14,14	279,76

Tableau 1 : Statistiques descriptives des coûts indexés des incidents cybers exprimés en M$

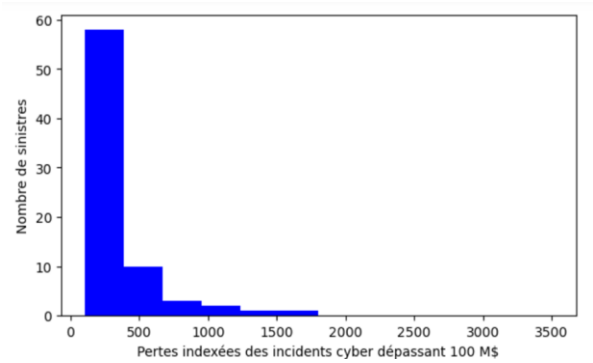

Figure 2: Histogramme des coûts indexées des incidents cyber dépassant 100 M$

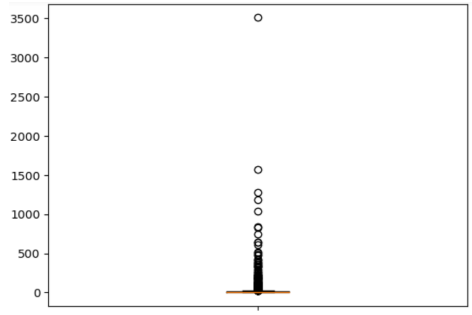

Figure 1: Boîte à moustaches des coûts indexées des incidents cyber en M$

La modélisation des sinistres peut être rendue complexe par les sinistres les plus extrêmes. C'est pourquoi une modélisation des montants de sinistres par deux modèles, raccordés à un certain seuil, est proposée.

Choix du seuil d'écrêtement des sinistres

Afin de déterminer le seuil, nous nous appuyons sur la théorie des valeurs extrêmes, en particulier sur le théorème de Pickands, qui démontre qu'il existe un seuil au-delà duquel les observations suivent une distribution de Pareto Généralisée.

La loi de Pareto Généralisée, notée GPD (Generalized Pareto Distribution) de paramètres d'échelle $\xi \in \mathbb{R}$ et de forme $\beta > 0$, se définit par sa fonction de répartition :

- $Cas\ \xi \neq 0 : G(x) = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$;
- $Cas\ \xi = 0 : G(x) = 1 - e^{-\frac{x}{\beta}}$; où : $y \geq 0\ pour\ \xi \geq 0\ et\ 0 \leq y \leq -\frac{\beta}{\xi}\ pour\ \xi < 0$

Le choix du seuil nécessite de trouver un équilibre approprié entre la variance et le biais. Une façon de déterminer un seuil est via l'approche graphique et d'utiliser pour cela, le Mean Residual Life Plot[2] (MRL). Celui-ci correspond au graphe des points $(u, e(u))$, *où* $e(u)$ est la moyenne des excès au-delà du seuil u, définie par :

$$e(u) = E(X - u < x / X > u)$$

Sous l'hypothèse d'une loi GPD de paramètre de forme $\xi < 1$, il existe un seuil $v > u$, où l'espérance de la moyenne des excès est linéaire en v. Les excès moyens peuvent ainsi être représentés

[2] Le MRL est introduit par Davison et Smith (1990).

graphiquement et lorsque le MRL commence à montrer un comportement linéaire, un seuil approprié peut être sélectionné.

Une deuxième méthode graphique est le graphe de stabilité des paramètres d'échelle et forme (Parameter stability plot). Cette méthode permet de déterminer un seuil optimal en ajustant les données à une distribution GPD et en utilisant différents seuils. Deux graphes de stabilité des paramètres peuvent ensuite être réalisés en traçant les valeurs estimées des paramètres ($\hat{\xi}\ et\ \hat{\beta}$) en fonction des différents seuils u. Au-delà d'un seuil $v > u$ et sous l'hypothèse d'un paramètre de forme constant, le paramètre d'échelle modifié est estimé par $\hat{\beta}_u - \hat{\xi}u, avec\ \beta_u = \xi u + \beta$. La stabilité des paramètres (forme et échelle) peut alors être observée et localisée.

L'utilisation de ces deux méthodes n'a pas été proprement concluante s'agissant de la détermination du seuil, mais a permis de donner un premier ordre de grandeur à celui-ci autour de 50 M$. Une analyse de sensibilité a été réalisée, afin de sélectionner le seuil avec la plus petite erreur standard du paramètre de forme estimé. Nous estimons les paramètres de forme et d'échelle et les écarts types correspondants pour les seuils allant de 50 jusqu'au $90^{ème}$percentile +0,5 M$, avec un pas de 0,5 M$. Seul un échantillon des résultats de l'analyse de sensibilité est ici présenté dans le tableau 2.

Comme nous nous intéressons à la queue de distribution des charges, le paramètre de forme attire particulièrement notre attention plus que le paramètre d'échelle. Nous choisissons donc le seuil présentant la plus petite erreur standard du paramètre de forme estimé. Il s'agit du **seuil de 50 M$** (en bleu dans le tableau).

Seuil	Nombre de dépassements	Estimateur du paramètre d'échelle	Ecart type du paramètre d'échelle	Estimateur du paramètre de forme	Ecart type du paramètre de forme
50	123	91,8877	14,9586	0,5839	0,145
52,5	122	87,5148	14,7041	0,618	0,1517
57	119	81,4836	14,552	0,6757	0,1649
60	114	85,519	15,5653	0,6629	0,1675

Tableau 2 : Echantillon des résultats de l'analyse de sensibilité de seuils compris entre 50 et le $90^{ème}$ percentile +0,5 M$, avec un pas de 0,5 M$.

Modélisation des charges des risques cyber extrêmes

Nous calibrons une loi GPD sur les excès des charges des sinistres cyber avec un seuil de dépassement de 50M$. Dans cette étude, nous utilisons la méthode classique du maximum de vraisemblance (ML) pour estimer les paramètres ξ et β, dans le tableau suivant :

Seuil	
Nombre de dépassement	123
Estimateur du paramètre d'échelle $\hat{\beta}$	91,89
Estimateur du paramètre de forme $\hat{\xi}$	0,58

Tableau 3 : Les estimateurs du ML des paramètres d'échelle et de forme de la GPD avec un seuil de dépassement de 50 M$

La fonction de densité empirique de la charge des sinistres au-delà du seuil et celle associée à la GPD calibrée semblent relativement proches dans la figure 3, ce qui semble indiquer un bon ajustement de la loi.

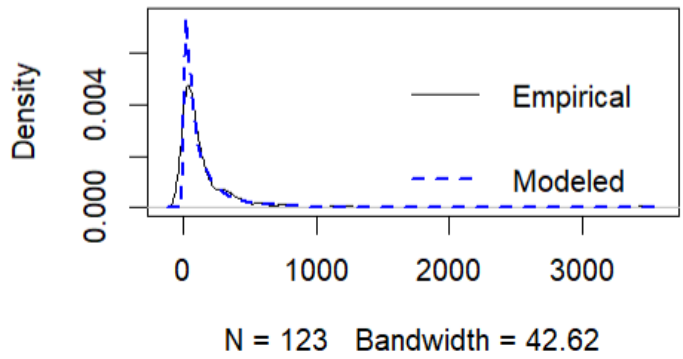

Figure 3: Fonctions de répartition des données empiriques et de la GPD ajustée

Modélisation des charges des risques cyber attritionnelles

Les charges des sinistres en dessous du seuil sont modélisées d'abord par des GLM et ensuite par des arbres de régressions CART. Une étape cruciale de ce processus est la division des données par un tirage aléatoire en base d'apprentissage (70 %) sur laquelle la définition de la modélisation s'appuiera et de test (30 %), utilisée pour déterminer la qualité prédictive du modèle.

Choix des lois de modélisation

L'analyse initiale porte sur l'évaluation de la pertinence de différentes distributions (Exponentielle, Gamma, Log-normale et Weibull) pour modéliser les charges des sinistres cyber attritionnels. Une proximité entre les trois lois Exponentielle, Gamma et Weibull, en termes de résultats, est remarquée par rapport à la loi log-normale. L'analyse de la log-vraisemblance ainsi que des critères AIC et BIC indique que la distribution de Weibull offre les meilleurs résultats, avec la plus grande log-vraisemblance et les critères AIC et BIC les plus bas, suivie par la distribution Gamma.

Modèle	Log-vraisemblance	AIC	BIC
Exponentielle	-1 863,93	3 729,86	3 734,42
Gamma	-1 755,47	3 514,94	3 524,07
Log-normale	-11 468,62	22 941,24	22 950,37
Weibull	**1 725,79**	**3 455,57**	**3 464,70**

Tableau 4 : Analyse de l'adéquation de l'ajustement des lois classiques aux charges des sinistres cyber attritionnels

Choix des variables explicatives

La base SAS OpRisk Global data comprend 49 variables. Afin de sélectionner les variables explicatives des charges des sinistres cyber, la revue de la littérature utilisant également cette base, ainsi que des tests statistiques (comme le test de Kruskal-Wallis et le coefficient V de Cramer) ont été utilisés. Le choix s'est finalement porté sur les variables suivantes :

Variables existantes dans la base :

- Basel Business Line - Level 2 : référant au 2ème niveau de la ligne métier (selon les normes BIS[3]),
- Event Risk Category : correspondant à la classification des événements opérationnels (selon le régime Bâle II),
- Region of Domicile : renseignant la région géographique du siège social de l'entreprise.

Variables construites :

- Region of Incident : renseignant la région de survenance du sinistre, à partir de la variable du pays de survenance du sinistre « Country of Incident »,
- Taille entreprise : donnant la classification de l'entreprise (GE, ETI, PME, TPE [4]) à partir du nombre de salariés « Number of Employees »,
- Secteur : correspondant au secteur de l'entreprise (financier ou non-financier) à partir du 1er niveau de la ligne métier « Basel Business Line - Level 1 ».

Pour simplifier la modélisation et compte tenu du nombre très important de modalités avec un nombre d'observations non significatifs, un regroupement par fréquence de certaines modalités des variables a été effectué.

Modèles GLM de Weibull et Gamma

Les modèles GLM mis en place pour la modélisation des pertes des sinistres cyber attritionnels sont ceux de Weibull et Gamma. Pour chaque GLM, l'approche commence par la construction d'un modèle complet, englobant toutes les variables explicatives présélectionnées, appelé le "Full model". Ensuite, la régression Stepwise, plus spécifiquement la méthode Forward, est utilisée pour sélectionner les variables les plus significatives à intégrer dans le modèle. À partir d'un modèle sans variables explicatives, appelé le "Null model", les variables sont rajoutées au fur et à mesure jusqu'à l'apport d'une

[3] BIS en anglais Bank for International Settlements est la Banque des règlements internationaux.
[4] GE : Grande Entreprise ; ETI : Entreprise à Taille Intermédiaire ; PME :Petite et Moyenne Entreprise ; TPE :Très Petite Entreprise.

nouvelle variable devienne marginal. Cette procédure se termine lorsque l'amélioration de la qualité du modèle n'est plus significative, indépendamment de la variable ajoutée.
Afin d'évaluer la qualité prédictive des modèles construits, nous les testons sur la base de test. La métrique utilisée est l'erreur quadratique moyenne (RMSE), et qui évalue la dispersion des écarts entre les coûts observés et les coûts prédits par le modèle.
$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$; où $\hat{y}_i$ sont les valeurs prédites par le modèle, y_i sont les valeurs réellement observées, et n est la taille du jeu de données. La RMSE s'exprime dans la même unité que la variable à prédire et est par conséquent facile à interpréter. Plus le RMSE est faible, plus la précision du modèle est bonne.

Modèles CART

Une autre approche pour modéliser les charges attritionnelles des sinistres cyber est d'utiliser les arbres de régression CART. Cet algorithme repose sur une partition binaire itérative de l'espace des observations, suivie de la détermination de sous-partitions optimales pour prédire les données. La construction d'un arbre CART se déroule en deux phases distinctes. Tout d'abord, la construction d'un arbre maximal, qui établit une gamme de modèles parmi lesquels le meilleur sera sélectionné. Ensuite, une seconde phase, de « pruning » ou d'élagage, se concentre sur la création de sous-arbres optimaux à partir de l'arbre maximal. Tout comme pour les GLM, les modèles CART sont construits sur la base d'apprentissage et validés sur la base de test.

Il est à noter que l'arbre élagué construit donne plus d'importance aux deux variables « Basel Business Line - Level 2 » et « Event Risk Category ». Il serait donc intéressant de construire des modèles GLM de Weibull et gamma uniquement avec ces deux variables. Ces modèles serviront de point d'appui aux comparaisons avec les autres modèles déjà construits.

Le tableau suivant récapitule les performances des différents modèles GLM et arbres CART, testés sur la base de test. Les niveaux des RMSE sont globalement proches et relativement faibles. Cependant, **c'est le modèle GLM gamma, utilisant les deux variables explicatives considérées comme les plus importantes par CART, qui a la plus petite RMSE de 9,2M$ soit 0,61% de la charge totale des sinistres cyber de la base de test. Ce modèle est donc le plus performant et sera retenu.** Ainsi, la combinaison des deux approches GLM et CART dans notre étude, améliore le pouvoir prédictif du modèle de la charge du sinistre cyber attritionnel.

	Weibull			**Gamma**			**CART**
Modèle	Full model	Model Stepwise Forward	Modèle avec choix des variables par CART	Full model	Model Stepwise Forward	Modèle avec choix des variables par CART	
RMSE base test	9,53	9,39	9,32	9,60	9,32	**9,20**	9,38

Tableau 5 : RMSE (en M$), des différents modèles GLM et CART appliqués à la base de test

Modélisation de la probabilité que le sinistre cyber soit grave ou non

Maintenant, que nous avons élaboré un modèle combiné pour représenter la charge des sinistres cyber, intégrant un modèle GPD pour les sinistres extrêmes et un modèle prédictif avec des variables explicatives pour les sinistres attritionnels, notre attention se tourne vers la modélisation de la probabilité qu'un sinistre cyber soit grave ou non.

Modèles de régression logistique

Nous commençons par mettre en œuvre une régression logistique pour modéliser la probabilité que le sinistre cyber soit grave ou non. Nous implémentons d'abord un modèle complet, le « Full Model », qui comprend les six variables explicatives. Ensuite, nous procédons à une régression Stepwise Forward. Pour évaluer la performance du modèle, nous utilisons des métriques basées sur la matrice de confusion suivantes :

Métrique	**L'exactitude (Accuracy)**	**Le rappel (Recall)**	**La précision (Precision)**	**F1-score**
Description	Elle mesure le taux de prédictions correctes sur l'ensemble des prédictions.	Il mesure la capacité d'un modèle à identifier tous les exemples positifs dans un ensemble de données.	Elle permet de mesurer le coût des faux positifs, c'est-à-dire ceux détectés par erreur.	Il est une moyenne harmonique de la précision et du rappel.
Formule	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

Tableau 6 : métriques de performance basées sur la matrice de confusion[5]

Sélection du seuil de classification optimal

Un bon moyen pour déterminer le seuil de classification optimal consiste à tracer la figure 4 de la métrique du F1-score en fonction de différentes valeurs de seuil. Nous choisissons de travailler avec le F1-score, car il est plus robuste en présence des données déséquilibrées. C'est le cas ici avec beaucoup plus de sinistres attritionnels que graves. Le seuil de classification optimal est celui ayant le plus grand F1-score. Il s'agit du seuil de **0,157.**

[5] TP : Vrais Positives, TN : Vrais Négatifs, FP : Faux Positifs, FN : Faux Négatifs

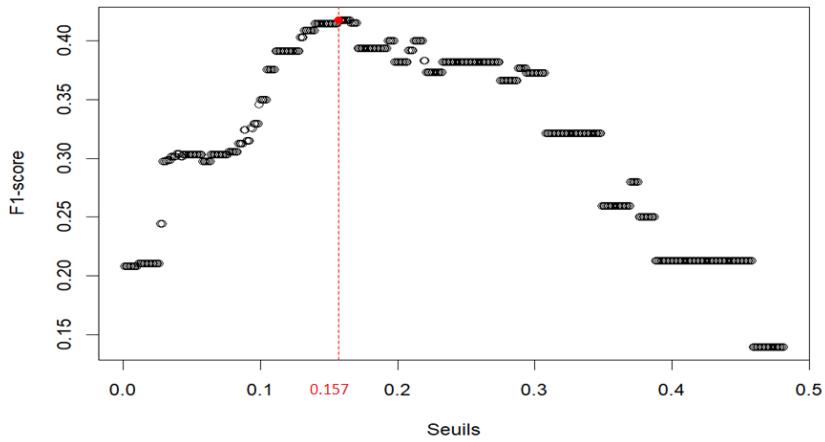

Figure 4 : F1-score en fonction des différentes valeurs du seuil de classification de la régression logistique

Modèle CART de classification

Ensuite, nous utilisons l'algorithme CART pour modéliser le risque d'avoir un sinistre extrême. Nous commençons par construire un arbre de classification maximal, suivi d'un arbre élagué.

Il est à noter que l'algorithme CART utilise les quatre variables suivantes dans la construction de l'arbre élagué : « Basel Business Line - Level 2 », « Event Risk Category », « Region of Incident » et « Taille Entreprise ». Il serait donc intéressant de construire un modèle de régression logistique uniquement avec ces variables, afin de le comparer avec les autres modèles déjà construits.
Le tableau suivant résume les performances des différents modèles GLM logistique et de l'arbre de classification CART, testés sur la base de test. Les métriques de performance des modèles GLM logistique sur la base de test sont très proches, avec une légère surperformance du modèle de la régression Stepwise Forward. Certes, l'arbre CART présente une accuracy plus grande (0,88) mais un F1-score plus petit (0,2). Compte tenu du déséquilibre des classes dans la base de données, le F1-score est à privilégier par rapport à l'accuracy.

Parmi tous les modèles construits, le modèle de régression logistique de Stepwise Forward présente la meilleure performance sur la base de test, avec le plus grand F1-score (0,42). De plus, il affiche le plus grand rappel (0,7) ce qui est plus souhaitable dans le contexte du risque cyber, où ne pas détecter un risque extrême peut avoir des conséquences graves. C'est donc le modèle que nous retiendrons.

	Modèle de régression logistique			**CART**
Modèle	Full model	Modèle Stepwise Forward	Modèle avec choix des variables par CART	
Accuracy	0,77	0,77	0,77	**0,88**
Precision	0,29	0,30	0,28	0,45
Recall	0,65	0,70	0,65	0,13
F1-score	0,40	**0,42**	0,39	0,20

Tableau 7: Métriques de performance des modèles GLM logistique et arbre de classification CART appliqués à la base de test

Limites et conclusions

Le modèle de la charge d'un sinistre cyber d'une entreprise proposé est ainsi la combinaison des meilleurs modèles testés auparavant :

- Le modèle GLM gamma, utilisant les deux variables explicatives « Basel Business Line - Level 2 » et « Event Risk Category », considérées comme les plus importantes par CART, prédictif de la charge attritionnelle ;
- La loi GPD calibrée aux excès de la charge au-delà du seuil des extrêmes de 50 M$, de paramètres de forme $\boldsymbol{\xi = 0,58}$ et d'échelle $\boldsymbol{\beta = 91,89}$;
- Le modèle de la régression logistique Stepwise Forward de la probabilité du risque extrême, utilisant les deux variables explicatives « Basel Business Line - Level 2 » et « Event Risk Category ».

La charge totale serait ainsi la somme de la modélisation attritionnelle via le modèle GLM Gamma et de la modélisation extrême, raccordées au seuil choisi, et pondérée par la probabilité de dépassement du seuil.

Ce modèle permettant de quantifier les pertes liées au risque cyber d'une organisation, peut être utilisé pour la détermination de la prime pure. Pour autant, il est à souligner que ce jeu de données ne permet pas de modéliser la fréquence de survenance des incidents cyber. Le modèle de la charge construit dépend de la ligne métier (Business line) de l'organisation et de la classification des événements opérationnels selon le régime bâlois. Ainsi, grâce au modèle construit, il est possible de proposer diverses couvertures en fonction de la garantie souscrite ainsi que de la ligne métier de l'organisation. A noter, que les autres variables explicatives testées auparavant peuvent être gardées dans la modélisation finale de la charge du sinistre cyber, notamment que les niveaux de performance des modèles avec l'ensemble des variables explicatives et ceux avec uniquement les deux variables retenues par la régression Stepwise Forward, sur la base de test, sont proches. De plus, l'utilisation des variables du secteur, de la taille de l'entreprise et des régions géographiques permettra une segmentation plus fine du tarif proposé selon le profil de l'entreprise et une diversification dans le portefeuille de l'assureur.

Pour conclure, nous avons souligné l'importance de modéliser séparément les charges cyber extrêmes et attritionnelles. Cette approche permet une prise en compte plus précise de la distribution des charges. Cependant, il est suggéré de déterminer deux seuils d'écrêtement distincts pour les charges des sinistres cyber dans les secteurs financiers et non financiers. Bien que le secteur financier soit parmi les plus touchés, le coût des incidents y est plus faible, potentiellement en raison d'investissements plus élevés en cybersécurité. Néanmoins, il est important de noter que la base de données utilisée présente des limitations, notamment un biais géographique à travers une concentration de sinistres aux Etats-Unis et l'absence de pertes liées à la réputation. Afin de mieux évaluer la méthodologie et les résultats, il est recommandé d'appliquer cette approche à une base de données plus étendue avec davantage de variables explicatives. En perspective, il est essentiel de souligner que les variables explicatives utilisées dans cette étude ne sont pas directement liées au risque cyber de l'entreprise. Bien que l'utilisation de données opérationnelles offre des possibilités de segmentation du processus de modélisation, l'incorporation de variables liées au risque cyber, telles que les mesures de prévention et de protection, pourrait améliorer la modélisation.

Synthetis Note

Cyber risk encompasses all potential dangers associated with the utilization of digital technologies. Presently, it stands as a significant threat to both businesses and institutions. Insurers play a crucial role in assisting companies in safeguarding themselves against cyber risk by offering specific coverages. However, the cyber insurance market is encountering challenges in its development. Due to the recent nature of cyber risk, insurers possess a limited claims history. Additionally, the multifaceted nature of cyber risk contributes to its insufficient understanding and presents difficulties in control.

To enhance their expertise and gain a better understanding of this risk, insurers may consider utilizing external data. With access to SAS data pertaining to operational losses experienced by companies globally, our primary goal is to evaluate the suitability of incorporating this external data into the modeling of cyber risk. This thesis is divided into two main sections. The initial section concentrates on modeling losses associated with cyber incidents in this database, while the second part delves into modeling the probability of a cyber incident being extreme or not.

Identification of cyber incidents in SAS OpRisk Global data

The identification of cyber incidents in the SAS OpRisk Global data is performed using a text mining algorithm applied to the descriptions of each incident. To optimize the efficiency of this algorithm, we preprocessed the incident descriptions. Before initiating the search, we established an initial comprehensive list of keywords related to cyber risks, followed by a second, more refined list consisting solely of the most specific keywords related to cyber risk. An incident was then classified as a cyber incident if it met one of the following three conditions :

1. At least one keyword from the refined list was present in the incident's description.
2. At least 10 words from the comprehensive list were present.
3. The incident belonged to operational events that could be associated with the associated risk according to the Basel classification.

To validate the identification procedure, random samples of incidents were taken to confirm their classification as genuinely related to cyber risks. **In total, we identified 1136 cyber incidents.**

Modeling the Severity of Cyber Losses

For the cyber incidents identified, the target variable for modeling is the "Current Value of Loss," representing the amount of losses in millions of dollars (M$), adjusted for the Consumer Price Index (CPI) to account for inflation. The analysis of the statistical summary of the costs of cyber incidents in Table 8, as well as Figures 5 and 6, reveals that the distribution of cyber loss severities exhibits a fat right tail, indicating the presence of extreme values.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation	Skewness	Kurtosis
0.11	0.53	2.16	31.65	10.41	3 505.35	149.07	14.14	279.76

Tableau 8 : Descriptive Statistics of Indexed Costs of Cyber Incidents in M$

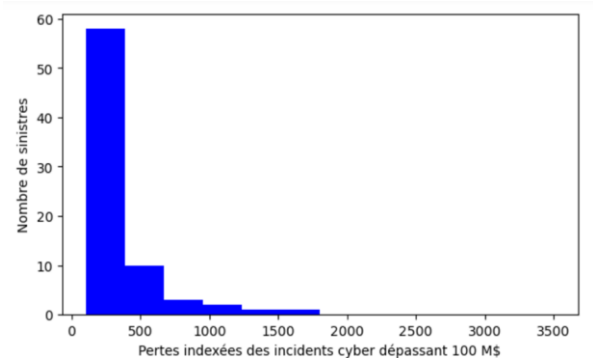

Figure 6 : Histogram of Indexed Losses of Cyber Incidents exceeding of 100 M$

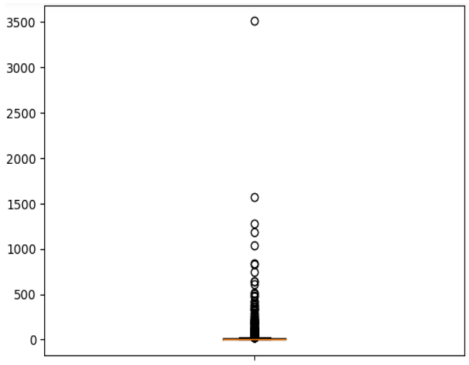

Figure 5 : Boxplot of Indexed Losses of Cyber Incidents in M$

Modeling losses can be complicated by the most extreme incidents. Therefore, modeling the amounts of losses using two models, connected at a certain threshold, is proposed.

Choice of the threshold

To determine the threshold, we rely on extreme value theory, specifically the Pickands' theorem, which demonstrates that there exists a threshold beyond which observations follow a Generalized Pareto Distribution (GPD).The Generalized Pareto Distribution, with scale parameter $\xi \in \mathbb{R}$ and shape parameter $\beta > 0$, is defined by its cumulative distribution function :

- $In\ the\ case\ of\ \xi \neq 0: G(x) = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$;
- $In\ the\ case\ of\ \xi = 0: G(x) = 1 - e^{-\frac{x}{\beta}}$;

Where : $y \geq 0\ for\ \xi \geq 0\ and\ 0 \leq y \leq -\frac{\beta}{\xi}\ for\ \xi < 0$

The choice of the threshold requires finding a suitable balance between variance and bias. One way to determine a threshold is through a graphical approach, using the Mean Residual Life Plot (MRL). This corresponds to the plot of points $(u, e(u))$, where $e(u)$ is the mean of excesses beyond the threshold u, defined by :

$$e(u) = \mathrm{E}(\mathrm{X} - \mathrm{u} < x / \mathrm{X} > \mathrm{u})$$

Under the assumption of a GPD distribution with a shape parameter $\xi < 1$, there exists a threshold $v > u$, where the expectation of the mean excess becomes linear in v. The mean excesses can thus be graphically represented, and when the Mean Residual Life Plot (MRL) begins to exhibit linear behavior, an appropriate threshold can be selected.

A second graphical method is the Parameter Stability Plot, which involves determining an optimal threshold by fitting the data to a GPD distribution and using different thresholds. Two parameter stability plots can then be created by plotting the estimated values of parameters ($\xi\ and\ \beta$) against various thresholds u, with 95% confidence intervals for these estimators. Beyond a threshold v and

under the assumption of a constant shape parameter, the modified scale parameter is estimated by $\hat{\beta}_u - \hat{\xi}u$.The stability of parameters (shape and scale) can then be observed and identified.

The use of these two methods did not yield a definitive conclusion regarding the threshold determination but provided a rough estimate of around \$50M.

A sensitivity analysis was conducted by selecting the threshold with the smallest standard error of the estimated shape parameter. We estimated the shape and scale parameters and their corresponding standard deviations for thresholds ranging from \$50M to the 90th percentile + \$0.5M, with a step of \$0.5M. Only a sample of the sensitivity analysis results is presented in Table 9.

As we are interested in the distribution tail of losses, the shape parameter draws our attention more than the scale parameter. Therefore, we choose the threshold with the smallest standard error of the estimated shape parameter. This corresponds to the \$50M threshold (highlighted in blue in the table).

Threshold	Number of Exceedances	Scale Parameter Estimator	Standard Deviation of Scale Parameter	Shape Parameter Estimator	Standard Deviation of Shape Parameter
50	123	91,8877	14,9586	0,5839	0,145
52,5	122	87,5148	14,7041	0,618	0,1517
57	119	81,4836	14,552	0,6757	0,1649
60	114	85,519	15,5653	0,6629	0,1675

Tableau 9 : Sample of Results from the Sensitivity Analysis of Thresholds ranging from \$50M to the 90th percentile + \$0.5M, with a step of \$0.5M.

Modeling Extreme Cyber Risk Losses

We calibrate a GPD distribution on the excesses of cyber risk losses with a threshold of \$50M. In this study, we employ the classical Maximum Likelihood (ML) method to estimate the parameters ξ and β, as shown in the following table :

Threshold	50
Number of Exceedances	123
Scale Parameter Estimator	91.89
Shape Parameter Estimator	0.58

Tableau 10 : Maximum Likelihood (ML) Estimators of Scale and Shape Parameters of the GPD with a Threshold of \$50M

The empirical density function of losses beyond the threshold and the one associated with the calibrated GPD appear relatively close in Figure 7, suggesting a good fit of the distribution.

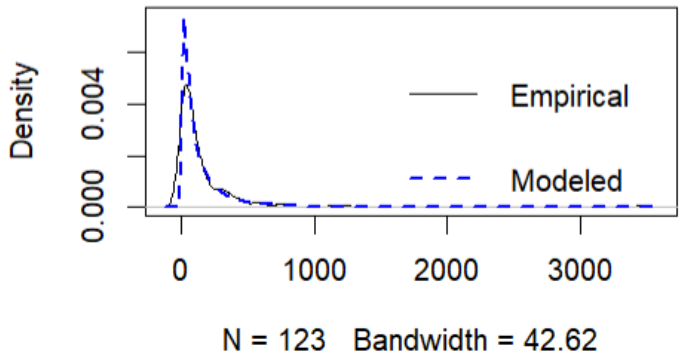

Figure 7: Empirical and Fitted GPD Cumulative Distribution Functions

Modeling Attritional Cyber Risk Losses

Losses below the threshold are first modeled using Generalized Linear Models (GLM) and then by Classification and Regression Trees (CART). A crucial step in this process is the division of the data through random sampling into a training set (70 %), on which the model definition will be based, and a test set (30 %), used to assess the predictive quality of the model.

Choice of Modeling Distributions

The initial analysis focuses on evaluating the suitability of different distributions (Exponential, Gamma, Log-normal, and Weibull) for modeling attritional cyber risk losses. Proximity among the Exponential, Gamma, and Weibull distributions is observed in terms of results compared to the log-normal distribution. The analysis of log-likelihood as well as AIC and BIC criteria indicates that the Weibull distribution provides the best results, with the highest log-likelihood and the lowest AIC and BIC criteria, followed by the Gamma distribution.

Model	Log-likelihood	AIC	BIC
Exponential	-1 863.93	3 729.86	3 734.42
Gamma	-1 755.47	3 514.94	3 524.07
Log-normal	-11 468.62	22 941.24	22 950.37
Weibull	**-1 725.79**	**3 455.57**	**3 464.70**

Tableau 11: Analysis of the Fit Adequacy of Classic Distributions to Attritional Cyber Risk Losses

Choice of Explanatory Variables

The SAS OpRisk Global data set comprises 49 variables. To select explanatory variables for attritional cyber risk losses, literature review utilizing this dataset, or focusing on cyber risk, along with statistical tests (such as the Kruskal-Wallis test and Cramer's V coefficient), were employed. The final selection includes the following variables:

Existing Variables in the dataset :

- Basel Business Line - Level 2: referring to the 2nd level of the business line (according to BIS[6] standards),
- Event Risk Category: corresponding to the classification of operational events (according to Basel II regulations),
- Region of Domicile: indicating the geographical region of the company's headquarters.

Constructed Variables :

- Region of Incident: indicating the region where the incident occurred, derived from the "Country of Incident" variable,
- Company Size: providing the classification of the company (Large Enterprise, Intermediate-sized Enterprise (ETI), Small and Medium-sized Enterprises (TPE, PME)) based on the number of employees,
- Sector: corresponding to the company's sector (financial or non-financial) based on the 1st level of the business line "Basel Business Line - Level 1."

To simplify the modeling and considering the very large number of different values with a non-significant number of observations, some modalities of the variables were grouped by frequency.

Weibull and Gamma GLM

We model the losses of attritional cyber risk incidents using the Weibull and Gamma GLM.
For both GLM, the approach begins with the construction of a full model, incorporating all pre-selected explanatory variables, referred to as the "Full model." Subsequently, the Stepwise regression, specifically the Forward method, is employed to select the most significant variables to be integrated into the model. Starting from a model without explanatory variables, known as the "Null model," variables are tested one by one, and the one that optimizes the selection criterion is added to the model. This procedure concludes when the improvement in model quality is no longer significant, regardless of the added variable.

To assess the predictive quality of the constructed models, we test them on the test set. The metric used is the Root Mean Squared Error (RMSE), which evaluates the dispersion of discrepancies between observed and model-predicted costs. $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$ where $\hat{y}_i$ are the predicted values by the model, y_i are the actual observed values, and n the data size. The RMSE is expressed in the same unit as the variable being predicted and is therefore easy to interpret. The lower the RMSE, the better the accuracy of the model.

CART Models

Another approach to model attritional losses of cyber incidents is to use Classification and Regression Trees (CART). This algorithm relies on an iterative binary partition of the observation space, followed by the determination of optimal sub-partitions to predict the data. The construction of a CART tree occurs in two distinct phases. Firstly, the construction of a maximal tree that establishes a range of models from which the best one will be selected. Then, a second phase, known as "pruning," focuses

[6] BIS is the Bank for International Settlements.

on creating optimal subtrees from the maximal tree. Similar to GLMs, CART models are built on the training set and validated on the test set.

It is noteworthy that the pruned tree gives more importance to the two variables "Basel Business Line - Level 2" and "Event Risk Category." Therefore, it would be interesting to build Weibull and gamma GLM models only with these two variables. These models will serve as a baseline for comparisons with the other already constructed models.

The following table summarizes the performance of different GLM and CART tree models, tested on the test set. The RMSE levels are generally close and relatively low. However, the gamma GLM model, using the two explanatory variables considered most important by CART, has the smallest RMSE of $9.2 million, representing 0.61% of the total charge of cyber incidents in the test set. Thus, this model performs the best and is chosen. Hence, the combination of both GLM and CART approaches in our study enhances the predictive power of the attritional cyber incident charge model.

	Weibull			**Gamma**			**CART**
Model	Full model	Stepwise Forward Model	Model with Variable Selection by CART	Full model	Stepwise Forward Model	Model with Variable Selection by CART	
RMSE test set	9.53	9.39	9.32	9.60	9.32	**9.20**	9.38

Tableau 12 : RMSE (in M$) of different GLM and CART models applied to the test set

Modeling the Probability of a Cyber Incident Being Severe or Not

Now that we have developed a combined model to represent the charge of cyber incidents, incorporating a GPD model for extreme incidents and a predictive model with explanatory variables for attritional incidents, our focus shifts to modeling the probability of a cyber incident being severe or not.

Logistic Regression Models

We start by implementing logistic regression to model the probability of a cyber incident being severe or not.

First, we implement a complete model, the "Full Model," which includes the six explanatory variables. Then, we proceed with a Stepwise Forward regression. To evaluate the model's performance, we use metrics based on the confusion matrix, as presented in the following table.

Metric	Accuracy	Recall	Precision	F1-score
Description	It measures the rate of correct predictions over the entire set of predictions.	It measures the model's ability to identify all relevant instances.	It measures the accuracy of the positive predictions.	It is a harmonic mean of precision and recall.
Formula	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

Tableau 13 : Performance metrics based on the confusion matrix[7]

Selection of the Optimal Classification Threshold

A good way to determine the optimal classification threshold is to plot Figure 8 of the F1-score metric against different threshold values. We choose to work with the F1-score because it is more robust in the presence of imbalanced data, as is the case here with significantly more attritional incidents than severe ones. The optimal classification threshold is the one with the highest F1-score. In this case, it is the threshold of 0.157.

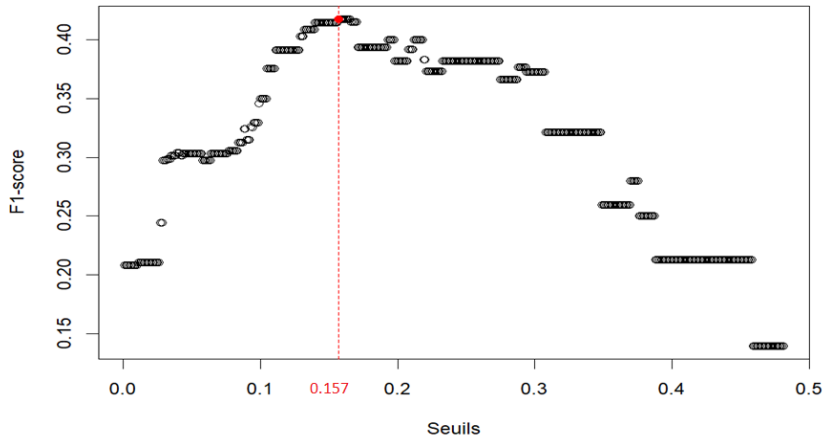

Figure 8 : F1-score as a function of different values of the logistic regression classification threshold

Model CART for Classification

Next, we use the CART algorithm to model the risk of having a severe incident. We start by building a maximal classification tree, followed by a pruned tree.

It is noteworthy that the CART algorithm uses the following four variables in the construction of the pruned tree: "Basel Business Line - Level 2," "Event Risk Category," "Region of Incident," and "Taille Entreprise." It would be interesting to build a logistic regression model only with these variables to compare it with the other models already constructed.

The following table summarizes the performance of different logistic GLM and CART classification tree models tested on the test set. The logistic GLM models' performance metrics on the test set are very

[7] TP : True Positives, TN :True Negatives, FP :False Positives, FN :False Negatives.

close, with a slight advantage for the Stepwise Forward regression model. While the CART tree exhibits a higher accuracy (0.88), it has a lower F1-score (0.2). Given the class imbalance in the database, F1-score should be prioritized over accuracy.

Among all the constructed models, the Stepwise Forward logistic regression model shows the best performance on the test set, with the highest F1-score (0.42). Moreover, it has the highest recall (0.7), which is more desirable in the context of cyber risk, where failing to detect a severe risk can have serious consequences. Therefore, this is the model we will retain.

	Logistic regression model			**CART**
Model	Full model	Stepwise Forward Model	Model with Variable Selection by CART	
Accuracy	0.77	0.77	0.77	**0.88**
Precision	0.29	0.30	0.28	0.45
Recall	0.65	0.70	0.65	0.13
F1-score	0.40	**0.42**	0.39	0.20

Tableau 14: Performance Metrics of Logistic GLM Models and CART Classification Tree Applied to the test Set

Limits and Conclusions

The model for the charge of a company's cyber incident presented is a combination of the best models tested previously:

- The GLM gamma model, using the two explanatory variables "Basel Business Line - Level 2" and "Event Risk Category," considered the most important by CART, predictive of attritional charges.
- The GPD law calibrated to excess charges beyond the extreme threshold of $50 million, with shape parameter $\xi = 0.58$ and scale parameter $\beta = 91.89$.
- The Stepwise Forward logistic regression model for the probability of extreme risk, using the two explanatory variables "Basel Business Line - Level 2" and "Event Risk Category."

The total loss would thus be the sum of the attritional modeling through the GLM Gamma model and the extreme modeling, linked to the chosen threshold and weighted by the probability of exceeding the threshold.

This model, quantifying losses related to cyber risk for an organization, can be used for determining the pure premium. However, it is important to note that this dataset does not allow modeling the frequency of cyber incidents.The constructed charge model depends on the organization's business line and the classification of operational events according to the Basel regime. Thus, with the built model, it is possible to offer various coverages based on the subscribed guarantee and the organization's business line. It should be noted that the other explanatory variables tested previously can be kept in the final modeling of the cyber incident charge, especially since the performance levels of models with all explanatory variables and those with only the two variables retained by Stepwise Forward regression, on the test set, are close. Furthermore, using variables related to the sector, company size, and

geographical regions will allow for a finer segmentation of the proposed rate based on the company's profile and diversification in the insurer's portfolio.

In conclusion, we have emphasized the importance of separately modeling extreme and attritional cyber charges. This approach allows for a more precise consideration of the charge distribution. However, it is suggested to determine two distinct clipping thresholds for cyber incident charges in financial and non-financial sectors. Although the financial sector is among the most affected, the cost of incidents is lower, potentially due to higher investments in cybersecurity.However, it is important to note that the database used has limitations, including geographical bias and the absence of reputation-related losses. To better assess the methodology and results, it is recommended to apply this approach to a more extensive database with more explanatory variables.In perspective, it is crucial to highlight that the explanatory variables used in this study are not directly related to the company's cyber risk. While using operational data offers opportunities for segmentation of the pricing process, incorporating variables related to cyber risk, such as prevention and protection measures, could enhance the modeling. Additionally, the current database does not allow modeling the frequency of cyber incident occurrences.

Table des matières

Introduction

De façon générale, le risque cyber correspond à tous les risques liés à l'usage des technologies numériques. Il représente aujourd'hui l'un des risques économiques majeurs. En raison de sa nouveauté et de sa nature en constante évolution, il n'existe pas de définition unique du risque cyber. La définition proposée par Cebula et Young (2010)[8] peut être retenue : le risque cyber est un risque opérationnel pouvant affecter la confidentialité, l'intégrité ou la disponibilité des données ou systèmes d'information.

Le risque cyber peut revêtir diverses formes, englobant à la fois les actes malveillants, comme le piratage ou l'hameçonnage ainsi que les incidents non intentionnels issus d'erreurs humaines ou d'accidents, comme la défaillance d'un système informatique.

Aujourd'hui, la cybercriminalité constitue l'une des menaces les plus importantes pour les entreprises et les institutions, capable de paralyser leur fonctionnement et pouvant mettre en péril leur survie. Les coûts associés aux risques cyber restent difficiles à évaluer, qu'ils soient directs ou indirects, mais sont estimés à plusieurs centaines de milliards d'euros annuellement pour l'économie mondiale et croissent chaque année.

Il est possible d'assurer un niveau de sécurité adéquat grâce à des bonnes pratiques, des outils de sensibilisation, des normes de sécurité et une réglementation appropriée. Toutefois, devant cette menace croissante, certaines entreprises choisissent de transférer une partie de leur risque cyber en souscrivant une assurance. Des couvertures d'assurances spécifiques sont ainsi proposées par certains acteurs du marché, dont le but de protéger les entreprises contre de telles attaques. Malgré cela, le risque cyber demeure encore relativement peu assuré. Ce constat est dû principalement à deux facteurs : d'une part, la difficulté des entreprises à appréhender ce risque, et d'autre part, la difficulté à le maîtriser par les assureurs.

Les acteurs du marché de l'assurance sont confrontés à un défi majeur : comment modéliser le risque cyber face au manque de données pertinentes ? Etant un risque relativement nouveau, l'historique des sinistres sur les portefeuilles des assureurs est limité et peu représentatif du risque cyber. En conséquence, la plupart des assureurs ne sont pas en mesure de s'appuyer seulement sur leurs propres bases historiques pour construire leur modélisation.

[8] Cebula et Young (2010), "A Taxonomy of Operational Cyber Security Risks", Software Engineering Institute Technical Note CMU/SEI-2010-TN-028

Pour pallier ce manque d'informations, les assureurs peuvent se tourner vers des bases de données externes. Cependant, ces sources externes peuvent poser question, notamment sur la qualité de leur alimentation ou sur leur pertinence vis-à-vis du portefeuille d'un assureur. Elles seront difficilement employées pour construire des modèles complets de tarification et de provisionnement. Néanmoins, elles peuvent au moins être utilisées pour vérifier certaines hypothèses et développer des méthodologies qui seront ensuite mises en œuvre par les compagnies d'assurances. Le jeu de données utilisé dans ce mémoire est SAS OpRisk Global data, qui répertorie des événements opérationnels, dont les risques cyber dans plusieurs pays. Alimentée en continu par SAS, elle fournit des informations sur le niveau de sinistralité de ces événements ainsi que des données sur les entreprises les ayant subis.

L'objectif principal de ce mémoire est de développer des modèles permettant de quantifier les pertes liées au risque cyber d'une organisation à partir de données externes, permettant à Deloitte de mieux s'approprier le risque cyber. Il est important de noter que les incidents cyber ne sont pas directement identifiables dans cette base, nécessitant ainsi le développement et l'application d'une méthodologie de fouille de texte pour sélectionner les sinistres cyber à modéliser. Les modèles de prédiction qui seront construits à partir de cette base devraient fournir un outil fiable pour quantifier le risque cyber.

Pour autant, il est à souligner que ce jeu de données ne permet pas de modéliser la fréquence de survenance des incidents cyber. En conséquence, ce mémoire se concentre uniquement sur la modélisation des coûts engendrés par les risques cyber.

Nous proposons une modélisation des coûts des sinistres cyber par deux distributions : d'une part, un modèle basé sur la famille de Distribution de Pareto généralisée (GPD) pour les sinistres extrêmes, et d'autre part, une distribution usuelle pour les sinistres attritionnels. Cette approche est basée sur deux distributions raccordées à un certain seuil. Elle vise à améliorer l'évaluation de la distribution des coûts, une information cruciale pour la détermination de la prime pure.

Nous essayerons également de modéliser la probabilité que le sinistre cyber soit extrême ou non. Les résultats de ces modèles peuvent servir d'indicateurs précieux pour les souscripteurs et de paramètres discriminants pour les actuaires travaillant sur ce sujet.

Ce rapport est organisé en quatre grandes parties. Dans la première partie, nous aborderons les définitions du risque cyber, mettant en lumière les spécificités de ce risque. De plus, nous fournirons un état des lieux du risque cyber dans le monde et par secteur d'industrie. En poursuivant, nous explorerons le marché de la cyber-assurance. Après avoir exposé les enjeux liés à la modélisation actuarielle du risque cyber, nous détaillerons les principales garanties et les perspectives de croissance de la cyber-assurance.

La deuxième partie se concentrera sur une revue de la littérature concernant la modélisation du risque cyber. Nous présenterons également le jeu de données SAS OpRisk Global data utilisé dans ce mémoire, et décrirons la procédure suivie pour identifier les sinistres cybers. Une fouille de termes dans la description des incidents de la base SAS OpRisk Global Data sera entreprise pour identifier les incidents cyber non reconnus jusqu'alors.

Dans la troisième partie, nous exposerons le cadre théorique de cette étude. Nous débuterons en présentant des éléments de la théorie des valeurs extrêmes, qui serviront tout d'abord à déterminer un seuil d'écrêtement approprié des sinistres en fonction de leurs coûts, et puis à la modélisation des dépassements du seuil choisi. Par la suite, nous introduirons les modèles linéaires généralisés (GLM) et les arbres de classification et de régression (CART), qui serviront à expliquer, le coût d'un sinistre cyber attritionnel d'une part, et la probabilité que le sinistre cyber soit extrême ou non d'autre part, en fonction des variables explicatives. Les bases théoriques sous-jacentes au choix des variables explicatives seront également présentées.

La quatrième et dernière partie sera consacrée à l'application des méthodes théoriques aux sinistres cyber de la base de données. Nous sélectionnerons un seuil d'écrêtement approprié. A partir du seuil retenu, un écrêtement sera effectué sur la base des coûts des sinistres : tous les montants supérieurs à ce seuil seront considérés comme graves et distingués des sinistres attritionnels. La distribution des sinistres extrêmes sera modélisée en appliquant la distribution de Pareto Généralisée (GPD). Quant aux coûts des sinistres attritionnels, ils seront modélisés à l'aide des modèles Linéaires généralisés (GLM) et des arbres de régression CART. Nous réaliserons une analyse descriptive des variables des deux jeux de données pour choisir celles qui expliquent les charges attritionnelles. Nous comparerons les pertes prédites par les différents modèles implémentés aux pertes réelles. Les résultats de toutes ces méthodes seront confrontés et analysés pour conclure sur le meilleur modèle à retenir. Nous développerons aussi les modèles prédictifs de la probabilité que le sinistre cyber soit extrême ou non.

Enfin, nous établirons les avantages et limites de notre modélisation et de la base de données sur laquelle elle repose afin de bien saisir tous les conditions et contexte qui sous-entendent l'utilisation potentielle de ce mémoire. Des pistes de travaux complémentaires seront également exposées.

Première partie

Contexte et motivation de l'étude du risque cyber

A mesure que le monde digital évolue en taille et en complexité, les menaces cyber prolifèrent. Cette exposition accrue au risque est particulièrement facilitée par le passage au tout numérique, l'introduction de nouvelles technologies et l'adoption du travail hybride. Il est indubitablement devenu une menace majeure pour la stabilité financière des institutions. Dans cette première partie, nous introduisons le risque cyber, ses caractéristiques et évolution dans le monde avant de présenter la cyber-assurance.

Chapitre 1 : Le risque cyber, définition et classification

Si elle peut sembler une question simpliste de prime abord, la définition du risque est plus complexe qu'il n'y paraît. Arriver à un consensus clair et net sur le sujet est complexe.

Dans son rapport « Cyber resilience - The cyber risk challenge and the role of insurance » en 2014, le CRO Forum[9] qui réunit les directeurs des risques (Chief Risk Officers (CRO)) des grandes sociétés d'assurance et de réassurance européennes, définit le risque cyber comme « tous les risques découlant de l'utilisation de données électroniques et de leur transmission, y compris les outils technologiques tels qu'internet et les réseaux de télécommunication. Il englobe également les dommages matériels pouvant être causés par des cyberattaques, la fraude commise par la mauvaise utilisation des données, toute responsabilité découlant du stockage des données, ainsi que la disponibilité, l'intégrité et la confidentialité des informations électroniques, qu'il s'agisse d'informations liées à des individus, des entreprises ou des gouvernements ». Le risque cyber englobe donc un large spectre de risques associés à l'usage des technologies numériques.

La dernière version du Cyber Lexicon[10] de 2023, publiée par le Conseil de stabilité financière (Financial Stability Board (FSB)), définit le risque cyber comme la combinaison de la probabilité que des incidents cyber se produisent et de leur impact. L'incident cyber, à son tour, est défini comme tout événement observable dans un système informatique qui soit : (i) compromet la cybersécurité d'un système d'information ou des informations que le système traite, stocke ou transmet ; soit (ii) viole les politiques et procédures de sécurité, ou les politiques d'utilisation acceptables, que cela résulte d'une activité malveillante ou non.

[9] Le CRO Forum a été créé en 2004 pour faire progresser les pratiques de gestion des risques dans le secteur de l'assurance. Les sociétés membres du CRO Forum sont de grandes compagnies d'assurance multinationales, basées dans le monde entier, avec une concentration en Europe.

[10] Le Cyber Lexicon est un ensemble de termes fondamentaux liés à la cybersécurité et à la cyber-résilience dans le secteur financier. Il est destiné à soutenir le travail du FSB, des organismes de normalisation, des autorités et des acteurs du secteur privé, pour aborder la cyber-résilience du secteur financier.

Les causes et les méthodes du risque cyber varient, et comprennent à la fois des incidents involontaires, humains ou des systèmes, et des actes malveillants. Compte tenu de l'omniprésence croissante et de la donnée et de son utilisation, ces risques peuvent toucher les particuliers mais aussi les entreprises et les administrations. Les auteurs peuvent être internes ou externes à l'organisation victime.

Dans ce chapitre, nous commençons par présenter les risques cyber, leur potentielle évolution et état des lieux. Ensuite, nous abordons les mécanismes d'assurance possibles pour ces risques. Enfin, nous examinons les difficultés entravant le développement du marché de la cyber-assurance.

1.1 Le risque cyber est un risque opérationnel

Le risque cyber est aussi une forme du risque opérationnel d'une entreprise. Le Comité de Bâle défini ce dernier comme « le risque de pertes résultant de l'inadaptation ou de la défaillance de procédures internes, de personnes et de systèmes ou résultant d'événements extérieurs ».

Le risque cyber est donc le risque opérationnel lié à l'information ou des systèmes informatiques et qui entraîne des conséquences affectant la confidentialité, la disponibilité ou l'intégrité des informations. La définition proposée par Cebula et Young (2010) peut être retenue : le risque cyber est un risque opérationnel pouvant affecter la confidentialité, l'intégrité ou la disponibilité des données ou systèmes d'information.

1.2 Classification du risque cyber

Afin de mieux circonvenir le risque cyber, quatre éléments peuvent entrer en ligne de compte, comme présenté dans le graphe ci-après :

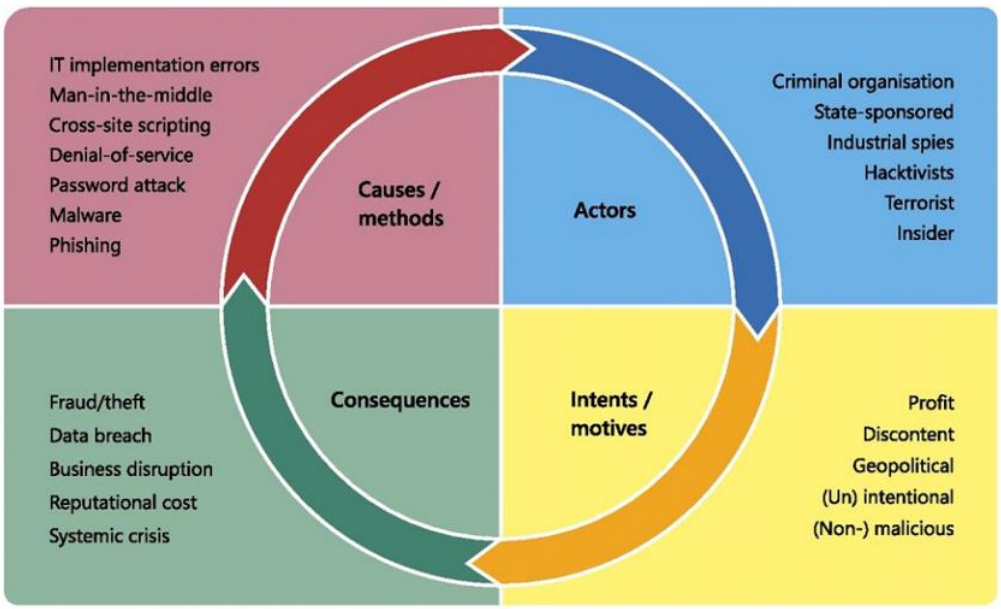

Figure 9 : Classification des risques cyber. (Source: BIS[11])

[11] BIS Working Paper, n°1039 "Cyber risk in central banking", Bank for International Settlements BIS), Septembre 2022

Nous retenons de la classification des risques cyber précédente, les quatre éléments suivants :

Les causes : elles peuvent être multiples, qu'elles soient intentionnelles ou accidentelles. Pour en citer quelques-unes, cela peut passer par une erreur de configuration dans le système informatique, une divulgation accidentelle de données ou un logiciel malveillant.
Les acteurs : le risque cyber peut provenir de réseaux criminels ou terroristes à dimension internationale, mais également d'un individu isolé : concurrent, client mécontent, hacker, un collaborateur ayant branché une clé USB sur le réseau,… Un sinistre peut aussi résulter d'un incendie ou d'une catastrophe naturelle.
Les motivations : un incident cyber peut être accidentel ou intentionnel. Le but dans ce dernier cas peut être purement lucratif (cas de rançon) ou géopolitique (attaques sponsorisées par les Etats).
Les conséquences : au-delà des dégâts matériels, les incidents cyber peuvent aussi engendrer des dommages immatériels. Les défaillances du système informatique peuvent nuire à l'intégrité et perturber les services. Les violations de données peuvent compromettre la confidentialité et la réputation. La fraude et le vol peuvent impliquer la perte de fonds ou de toute information à caractère personnel. Nous entendons par une information à caractère personnel, la définition retenue par la Commission nationale de l'informatique et des libertés (CNIL)[12] : « toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement ». Par exemple : un nom, une photo, une empreinte, une adresse postale, une adresse mail, un numéro de téléphone, un numéro de sécurité sociale, un matricule interne, une adresse IP, un identifiant de connexion informatique, un enregistrement vocal, ... Peu importe que ces informations sont confidentielles ou publiques.

1.3 Cybercriminalité : les menaces les plus répandues

Les cyberattaques peuvent revêtir diverses formes et méthodes. Même parmi les méthodes les plus courantes, certaines sont bien connues et d'autres encore trop méconnues. On retrouve ainsi notamment :

- **Hameçonnage (phishing) :** est une technique frauduleuse destinée à leurrer la victime pour l'inciter à communiquer des données personnelles (comptes d'accès, mots de passe…) en se faisant passer pour un tiers de confiance.
- **Malware ou logiciels malveillants :** il s'agit de logiciels hostiles ou intrusifs conçus pour endommager, perturber ou accéder de manière non autorisée aux systèmes informatiques. Cela inclut les virus, les vers, les chevaux de Troie, et les rançongiciels ou ransomwares.
- **Attaques par déni de service ou DDoS (Distributed Denial of Service) :** est une attaque informatique ayant pour but de rendre inaccessible un serveur, d'empêcher les utilisateurs légitimes d'un service de l'utiliser, grâce à l'envoi de multiples requêtes ou par l'exploitation de failles de sécurité afin de provoquer une panne d'un service.
- **Ingénierie sociale (social engineering) :** s'appuie sur la nature humaine. Elle fait appel aux vulnérabilités humaines, comme la volonté de faire confiance aux autres, pour obtenir des informations sensibles ou un accès à un terminal personnel. Un email semblant provenir d'un fournisseur de confiance qui demande de mettre à jour les informations de carte de crédit, un message vocal menaçant d'un interlocuteur se faisant passer pour une administration fiscale : ce sont quelques exemples d'ingénierie sociale.
- **Attaques par force brute :** consistent à tester, l'une après l'autre, chaque combinaison possible d'un mot de passe ou d'une clé pour un identifiant donné afin de se connecter au service ciblé.

[12] Source : https://www.cnil.fr/fr/cnil-direct/question/une-donnee-caractere-personnel-cest-quoi

- **Attaques par injection SQL :** ces attaques exploitent les failles de sécurité des applications web en injectant du code SQL malveillant pour accéder et manipuler les bases de données.
- **Attaques "Man-in-the-Middle" (MitM) :** ces attaques interceptent les communications entre deux parties légitimes, soit en écoutant, soit en se faisant passer pour un participant légitime, permettant à un attaquant de lire, modifier ou altérer les échanges de données.
- **Attaques "Zero-Day" :** sont des tentatives réussies par les cybercriminels de trouver et d'exploiter des vulnérabilités ou des failles inconnues dans un logiciel ou un système d'exploitation, avant que les développeurs n'aient la possibilité de les rectifier.
- **Attaques de la chaîne d'approvisionnement (supply chain attaque) :** elles se produisent lorsqu'un pirate accède au réseau d'une entreprise par l'intermédiaire de vendeurs tiers ou de la chaîne d'approvisionnement. C'est un moyen très efficace d'enfreindre la sécurité en injectant des virus ou des composants malveillants dans un produit sans que le développeur, le fabricant ou le client final ne s'en aperçoive.

1.4 Retour sur quelques cyberattaques marquantes

Le schéma ci-après, montre les principales cyberattaques au niveau mondial au cours des huit dernières années. Le rançongiciel constitue le risque le plus avéré, avec notamment un sinistre de 10 milliards de dollars chez NotPetya[13].

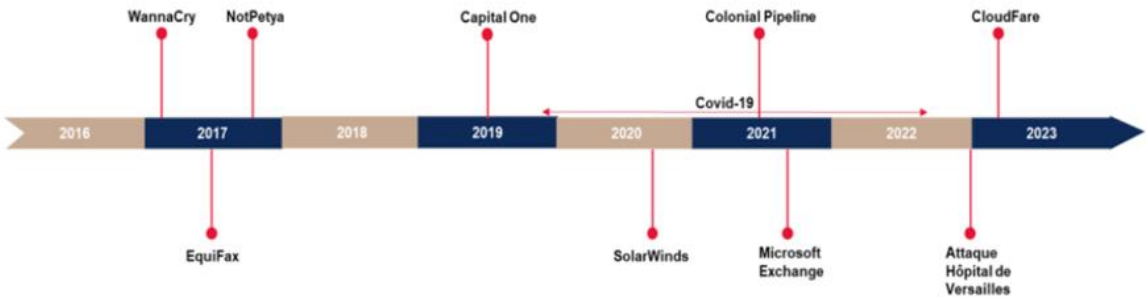

Figure 10 : Principales cyberattaques au cours des 8 dernières années (Source[14] : Seabird conseil)

Dans le détail, ces attaques ont eu un impact significatif sur les entreprises et systèmes informatiques entraînant des perturbations majeures. Les coûts associés sont considérables : pertes financières directes, coûts indirects tels que la perte de clients, atteinte à la réputation, litiges potentiels, interruption des activités ou encore baisse de productivité.

[13] Incident détaillé dans la partie 2.3

[14] Source : https://www.seabirdconseil.com/nos-decryptages/le-cyber-risque-de-quoi-parle-t-on-lexique-seabird/

Chapitre 2 : Evolution continue du risque cyber

La nature du risque cyber évolue rapidement et constamment avec des attaques de plus en plus sophistiquées et mieux ciblées. L'enquête « Global Risk Management Survey » réalisée par Aon en 2017[15] révèle une augmentation rapide tant en termes de fréquences que de niveau du risque cyber. Les pirates informatiques utilisent des méthodes de plus en plus élaborées et ciblent un éventail élargi d'organisations qui peinent à assurer un rythme d'amélioration suffisant pour mieux prévenir.

2.1 Une menace majeure

Ces dernières années, les risques cyber ont pris une place majeure. En 2023, et pour la deuxième année consécutive, le risque cyber se place en tête du baromètre des risques d'Allianz[16], juste avant l'interruption d'activité et les évolutions macroéconomiques. Ce baromètre est un classement annuel des risques d'entreprise établi par Allianz Global Corporate & Specialty (AGCS), l'assureur du groupe Allianz.

La cybercriminalité et les défaillances des systèmes informatiques figurent parmi les risques les plus sévères pesant sur l'économie mondiale. « Il y a une tempête cybernétique qui se prépare », a déclaré Sadie Creese, professeur en cybersécurité à l'Université d'Oxford, lors d'une interview à la Réunion annuelle du Forum économique mondial 2023 à Davos, en Suisse[17]. « Cette tempête se forme, et il est vraiment difficile d'anticiper à quel point elle sera dévastatrice ».

Les cyberattaques contre les entreprises constitueraient, selon le président de la Réserve fédérale des États-Unis, le risque actuel le plus important pour l'économie américaine, surpassant même une crise financière similaire à celle de 2008[18].

2.2 Un nombre considérable de victimes

Les organisations touchées par les cyberattaques sont extrêmement nombreuses. Le 6ème baromètre annuel du Club des Experts de la Sécurité de l'Information et du Numérique (CESIN)[19], révèle que plus d'une entreprise française sur deux déclare avoir subi entre une et trois attaques cyber au cours de l'année 2020. Ce chiffre tient compte uniquement des attaques réussies. Plus précisément, 80 % des incidents ont été répertoriés comme étant de l'hameçonnage et plus de la moitié ont exploité une faille de sécurité.

---

[15] Le 2017 Global Risk Management Survey d'Aon est conçu pour offrir aux entreprises des perspectives et des connaissances leur permettant d'être compétitives dans un environnement commercial complexe. Réalisée tous les deux ans, l'étude a recueilli des données auprès de 1.843 répondants d'entreprises publiques et privées du monde entier.

[16] Source : https://newsroom.allianz.fr/actualites/barometre-des-risques-allianz-2023-a554-98cdb.html

[17] Du 16 au 20 janvier 2023, plus de 2500 participants issus du monde économique, politique, scientifique et culturel ont participé à la rencontre annuelle du forum économique mondial à Davos.

[18] Source : Rapport d'information N° 678 fait au nom de la délégation aux entreprises relatif à la cybersécurité des entreprises, par MM. Sébastien MEURANT et Rémi CARDON, Sénateurs, Enregistré à la Présidence du Sénat le 10 juin 2021, page 24. Lien : https://www.senat.fr/rap/r20-678/r20-6781.pdf

[19] Baromètre de la cyber-sécurité des entreprises, vague 6 – Janvier 2021,page 39, enquête réalisée par Opinionway pour le CESIN. Lien : https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/marketing/internet-et-ntic.html

Au Royaume-Uni, 39 % des entreprises ont identifié des attaques cyber en 2022, d'après le rapport « UK Official Statistics Cyber Security Breaches report »[20] pour l'année 2022.

2.3 Des coûts pesants et croissants

Un incident cyber peut coûter très cher. En octobre 2020, l'entreprise de service numérique française Sopra Steria, a été victime d'une attaque de rançongiciel, spécifiquement du logiciel malveillant Ryuk. Les assaillants ont chiffré les fichiers de Sopra Steria et ont exigé le paiement d'une rançon en échange de la clé de déchiffrement. L'entreprise a estimé ses pertes à 50 millions d'euros[21].

Une des cyberattaques les plus couteuses de l'histoire, NotPetya de 2017, dirigée contre l'Ukraine, mais qui s'est propagée vers des systèmes dans le monde entier, a infligé des dégâts estimés à plus de 10 milliards de dollars[22], soit un peu plus de 10 % du PIB de l'Ukraine à l'époque. Initialement, il a été dissimulé dans une mise à jour logicielle légitime appelée MeDoc, largement utilisée en Ukraine. Par conséquent, de nombreux clients, qui utilisent le logiciel, ont reçu le malware sous la forme de mise à jour, et il s'est ensuite répandu sur tout le réseau. Une fois infiltré dans les systèmes, NotPetya a rapidement dépassé le cadre d'une simple opération de rançon. Contrairement à d'autres rançongiciels, NotPetya semblait avoir un objectif destructeur plutôt que lucratif. Après avoir chiffré les fichiers des utilisateurs, il a rendu les données inaccessibles, mais le processus de paiement de la rançon n'a pas conduit à la restauration des fichiers. De plus, le malware a également affecté la fonction de démarrage des ordinateurs, provoquant des dommages considérables et des perturbations au-delà de la simple perte de données. Certaines entreprises mondiales ont subi d'importantes pertes financières, notamment Maersk, le géant du transport maritime, qui a enregistré des dommages allant jusqu'à 300 millions de dollars. Même des infrastructures critiques, dont la centrale nucléaire de Tchernobyl, ont été perturbées, ce qui donne une idée de la portée et de la criticité potentielle de ces attaques.

En matière de pertes agrégées, l'article « Top 10 Cybersecurity Predictions And Statistics For 2023 » de Cybersecurity Ventures[23], chercheur mondial et principale source d'information dans le domaine de la cybersécurité, annonce que la cybercriminalité est prévue pour causer des pertes d'un total de 8 milliards de dollars à l'échelle mondiale en 2023. Le magazine prévoit aussi que les coûts mondiaux de la cybercriminalité augmenteront de 15 % par an au cours des trois prochaines années, atteignant 10,5 milliards de dollars par an d'ici 2025, contre 3 milliards de dollars en 2015.

Si pour un particulier, le sinistre est souvent de quelques milliers, le dommage peut rapidement se compter en millions, comme le cas de la violation de données, dont le coût moyen est en progression par ailleurs chaque année. Selon le rapport « Cost of a data Breach 2023 » d'IBM[24] sur les violations

---

[20] Source : https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2022/cyber-security-breaches-survey-2022

[21] Source : https://www.fintechfutures.com/2020/11/sopra-steria-ransomware-attack-costs-group-e50m/

[22] Source : https://www.cyber-cover.fr/cyber-documentation/cyber-criminalite/cybercriminalite-notpetya-le-malware-a-10-milliards-de-dollars#:~:text=L'ancien%20conseiller%20de%20la,%C3%A0%2010%20milliards%20de%20dollars.

[23] Source : https://cybersecurityventures.com/stats/

[24] International Business Machines Corporation (IBM), est une entreprise multinationale américaine présente dans les domaines du matériel informatique, du logiciel et des services informatiques.Source :https://www.ibm.com/reports/databreach#:~:text=Take%20a%20deep%20dive%20into,h

des données, le coût moyen est passé de 4,35 millions de dollars en 2022 à 4,45 millions de dollars en 2023.

Pour conclure, les dommages causés par un sinistre cyber peuvent parfois être difficile à évaluer car liés à des éléments immatériels tels que la perte de confiance, la dépréciation de la valeur de la propriété intellectuelle, la perte de données, la mise en place de nouveaux systèmes de sécurité, la contagion à d'autres entreprises, ou encore l'impact sur les futures primes de cyber-assurance. Le risque cyber peut même remettre en question la continuité et/ou la pérennité de l'activité de l'entreprise.

2.4 Des attaquants de plus en plus performants

La nature du risque cyber évolue rapidement et constamment marquée par des attaques de plus en plus élaborées et mieux ciblées. Un véritable écosystème cybercriminel aux ressources considérables, a émergé progressivement et s'est perfectionné, ce qui lui permet de conduire des attaques hautement sophistiquées. La spécialisation et la professionnalisation des attaquants s'expliquent notamment par les gains financiers déjà accumulés, et le potentiel de gains à venir très important.

Suivant le rapport « Emerging Risks Initiative Major Trends and Emerging Risk Radar » de 2023 du CRO forum[25], relatif aux risques émergents dans le secteur assurantiel, quoique certains aspects du risque cyber sont déjà apparus et doivent être considérés comme une menace persistante, de nouveaux éléments peuvent être continuellement observés dans les attaques de plus en plus sophistiquées et dont les impacts s'élargissent, moyennant l'évolution des nouvelles technologies.
L'intelligence artificielle par exemple, qui entre dans une ère actuellement très prometteuse et pleine d'élan est aussi une source de craintes comme facilitatrice des activités cybercriminelles. Son utilisation pourrait notamment permettre plusieurs activités liées à la cybercriminalité : de la distribution massive d'informations truquées, entraînant encore plus de difficultés à discerner la vérité, au développement de logiciels malveillants sophistiqués. Les méthodes assistées par l'Intelligence Artificielle (IA), telles que les deepfakes et le clonage vocal, augmentent les chances de succès de l'ingénierie sociale, en trompant un individu souvent peu conscient de ce que l'IA peut permettre.

En outre, le temps de détection d'un risque cyber peut être parfois long. L'activité cyber est passée d'une activité très perturbatrice, comme dans les premiers temps des logiciels malveillants, à une activité très secrète. Les violations les plus graves restent indétectées pendant des périodes considérables. En se référant au rapport IBM de 2022 sur le coût d'une violation de données, il s'écoule en moyenne 277 jours, presque 9 mois, avant d'identifier la violation de données. Les attaques se font à un niveau très professionnel, et cherchent à prendre le contrôle total sur l'environnement. Elles durent longtemps.

2.5 L'impératif de la sensibilisation pour une cybersécurité renforcée

Dans la majorité des cas, les menaces liées aux cyberattaques proviennent avant tout d'une erreur humaine. Cette situation souligne l'importance cruciale de sensibiliser les individus aux enjeux de la

ow%20to%20mitigate%20the%20risks.&text=The%20global%20average%20cost%20of,15%25%20increase%20over%203%20years.

[25] Source : https://www.thecroforum.org/emerging-risk-initiative-major-trends-and-emerging-risk-radar-2023/

cybersécurité, car l'erreur d'un seul individu peut avoir des conséquences dévastatrices, pouvant suffire à endommager l'ensemble d'une entreprise. La nécessité d'une sensibilisation individuelle et collective aux conséquences potentielles des attaques cyber devient ainsi un impératif pour renforcer la résilience des organisations face à ces menaces croissantes.

Chapitre 3 : Un état des lieux du risque cyber dans le monde

Les risques cyber menacent tant les grandes que les petites institutions, les pays riches comme les pays pauvres, et ne connaissent pas de frontières. Aucun secteur d'activité n'est épargné. Les tensions géopolitiques ne font qu'aggraver les risques. Si le risque cyber était un pays, il serait la 3ème économie mondiale, en termes de PIB, selon le rapport d'information n° 678 (2020-2021) du Sénat[26].

3.1 Risque cyber aux Etats-Unis

En tant que puissance mondiale, les États-Unis sont une cible attractive pour les cybercriminels qui cherchent à voler des données sensibles, à compromettre des systèmes essentiels ou à perturber l'activité économique. Les États-Unis abritent de nombreuses grandes entreprises qui dominent le marché lié à Internet. Ces entreprises sont des cibles de choix en raison de la valeur de leurs données et de leur influence mondiale.

Le rapport « Internet Crime Report » publié par le Federal Bureau of Investigation (FBI) en 2021[27], signale un accroissement sans précédent, des cyberattaques enregistrées aux Etats-Unis. Le nombre des plaintes déposées à l'Internet Crime Complaint Center (IC3) du FBI, a augmenté de 7% par rapport à 2020, impliquant une perte totale de plus de 6,9 billions de dollars. Le phishing, le rançongiciel, les schémas de compromission des courriels professionnels (BEC) et l'utilisation des cryptomonnaies à des fins criminelles figurent parmi les principaux incidents réclamés en 2021. En 2022, le même rapport montre que bien que le nombre total des plaintes ait diminué de 5%, les pertes financières ont considérablement augmenté de 49%. Le phishing était toujours le crime cyber le plus signalé.

3.2 Risque cyber dans l'Union européenne

L'agence de l'Union Européenne pour la cybersécurité (ENISA), confirme dans la 11ème édition de son rapport annuel « ENISA Threat Landscape »[28], qui couvre la période de Juillet 2022 à Juin 2023, une nette augmentation des attaques cyber dans l'Union européenne, tout au long de la période examinée.

La guerre en cours en Ukraine demeure un facteur significatif qui façonne le domaine de la cybersécurité. En parallèle de la guerre sur le terrain, un véritable champ de bataille numérique prend forme. Cette guerre a pu cristalliser l'arrivée de ce nouvel axe de bataille rendant les guerres d'autant plus hybrides. Le piratage informatique a connu une expansion, marquée par l'émergence de nombreux groupes organisés. Ceux-ci ont particulièrement développés des logiciels de rançongiciel dont le nombre d'incidents a été en très forte croissance durant la première partie de 2023.

Cependant, il est important de noter que le nombre d'événements observés ne signifie pas nécessairement une augmentation réelle du nombre d'attaques ou de la gravité de leurs impacts. Cette

[26] Source : https://www.senat.fr/rap/r20-678/r20-6781.pdf
[27] Source : https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf
[28] Source : https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

hausse pourrait être attribuée à une attention médiatique ou publique portée sur des événements spécifiques pendant une certaine période.

Conformément au rapport de 2022 de l'ENISA[29], la liste des cyberattaques les plus éminentes au sein de l'Union européenne, sur la période entre avril 2021 et juillet 2022 est la suivante :

- Rançongiciel
- Logiciel malveillant (malware)
- Ingénierie sociale
- Menaces contre les données (threats against data) qui englobe différentes formes de dangers ou risques susceptibles de compromettre la sécurité et l'intégrité des données
- Désinformation qui consiste à fournir des informations trompeuses ou des fausses alertes par exemple, aux victimes. Cela peut distraire et détourner l'attention des équipes de sécurité, les empêchant de se concentrer sur de réelles menaces.
- Ciblage de la chaîne d'approvisionnement (supply chain targeting).

En raison de la taille de l'économie et la présence des entreprises technologiques majeures, les États-Unis sont massivement plus touchés que l'Union européenne par les cyberattaques. Au niveau des types des attaques enregistrées, une analogie entre les deux est constatée. Le rançongiciel et l'hameçonnage (phishing) sont les principales attaques.

3.3 Risque cyber en France

Le Club des Experts de la Sécurité de l'Information et du Numérique (CESIN), dévoile les résultats de sa septième grande enquête sur la cybersécurité des entreprises françaises[30]. En 2021, plus d'une entreprise française sur deux déclare avoir subi entre une et trois attaques cyber au cours de l'année. Ce chiffre tient compte uniquement des attaques réussies, ayant eu des répercussions flagrantes pour les victimes. L'hameçonnage (phishing) reste le vecteur d'attaque le plus fréquent. Malgré la prise de conscience croissante sur ce risque, 73 % des entreprises déclarent l'hameçonnage comme vecteur d'entrée principal pour les attaques subies.

Dans son panorama de la cybermenace en 2022, l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI)[31], estime que le contexte géopolitique et les évènements majeurs que sont la coupe du monde de rugby 2023 et les jeux olympiques et paralympiques de Paris 2024 sont susceptibles de fournir aux attaquants de nombreuses opportunités de nuire.

3.4 Risque cyber par secteur d'industrie

Si le numérique est aujourd'hui notre quotidien, certains secteurs d'activité sont d'autant plus à risque vis-à-vis du risque cyber que ce soit du fait du potentiel de nuisance pour certains secteurs comme la santé, l'administration et l'éducation ou la nature du secteur d'activité qui repose sur le numérique comme le secteur financier et les services informatiques.

[29] Source : https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022
[30] Source : https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/marketing/internet-et-ntic.html
[31] Source : https://www.cert.ssi.gouv.fr/cti/CERTFR-2023-CTI-001/

- **En termes de fréquence des sinistres**

Le secteur financier figure parmi les secteurs les plus victimes d'incidents cyber en raison des bénéfices potentiels significatifs qu'il représente. S'ajoute à cela, la forte numérisation de ses infrastructures, en l'occurrence les services de compensation et de paiement, ce qui augmente sa surface d'exposition au risque cyber. La forte interconnexion entre ses acteurs, laisse le risque de contagion très élevé. De plus, la prévalence du travail à domicile en lien avec la pandémie de Covid-19, a fortement exposé le secteur financier aux menaces cyber, selon le bulletin 2020 de la Banque des règlements internationaux[32]. En effet, le recours rapide au télétravail, a nécessité la numérisation rapide des activités des entreprises et a rendu difficile le maintien des normes de cybersécurité. Le graphique ci-après, publié dans le même bulletin, donne les proportions des incidents cyber par secteur d'activité. Le secteur financier est touché par le risque cyber plus que tous les autres secteurs [33] depuis la pandémie. En deuxième position, se trouve le secteur des services, qui englobe la télécommunication, le tourisme et la technologie de l'information parmi d'autres.

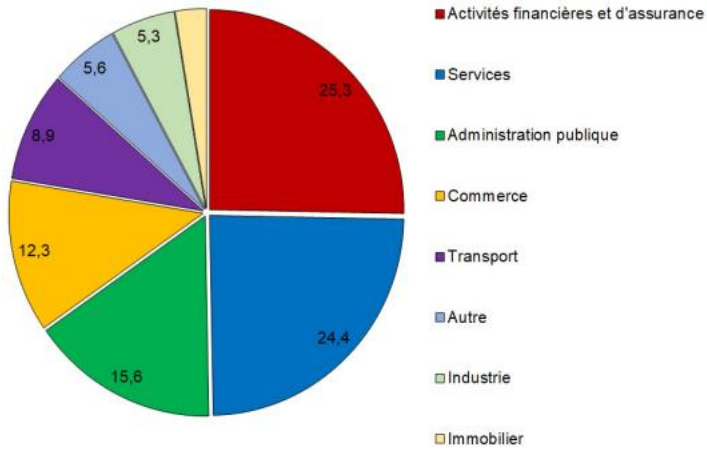

Figure 11: Répartition des incidents cyber par secteur lors des six premiers mois de la pandémie de Covid-19 (mars-septembre 2020)[34]

- **En termes de coûts des sinistres**

Si les études publiées ne détaillent pas nécessairement pour chaque type de sinistres la répartition des coûts moyens par secteur d'activité, l'exemple des violations de données peut être retenu comme exemple des différences fondamentales à ce niveau. En effet, selon les rapports de 2022[35] et 2023[36] d'IBM sur les violations des données, c'est le secteur de la santé qui enregistre le plus grand coût moyen des violations des données (10,93 millions de dollars en 2023), soit plus du double du coût moyen de toutes les violations, suivi par le secteur financier (5,90 millions de dollars).

---

[32] Source : https://www.bis.org/publ/bisbull37.pdf
[33] Les données portent principalement sur les États-Unis et exclut le secteur de la santé.
[34] BIS Bulletin N°37, « Covid-19 and cyber risk in the financial sector », par Iñaki Aldasoro, Jon Frost, Leonardo Gambacorta et David Whyte, page 24.
[35] Source : https://www.ibm.com/reports/data-breach actionguide#:~:text=Data%20breach%20costs%20averaged%20USD,Breach%20Report%20has%20been%20published.
[36] Source : https://www.ibm.com/reports/data-breach

Le coût élevé des violations des données à caractère personnel dans le secteur de la santé, est principalement dû à la nature sensible de ces informations. Les violations de données dans le secteur de la santé peuvent impliquer des informations médicales, des dossiers de patients, des antécédents médicaux, et d'autres données confidentielles. S'ajoute à cela, le fait que le secteur de la santé est soumis à des réglementations strictes en matière de protection des données, telles que la loi HIPAA[37] (Health Insurance Portability and Accountability Act) aux États-Unis. Les entreprises du secteur de la santé sont tenues de respecter des normes élevées en matière de sécurité des données, et les violations peuvent entraîner des amendes importantes.

Tout comme dans le secteur de la santé, le secteur financier traite des données sensibles, notamment des données financières, des informations sur les cartes de crédit, des détails sur les comptes bancaires, ... Le système financier est aussi caractérisé par la complexité de ces systèmes ce qui accroît la surface d'attaque.

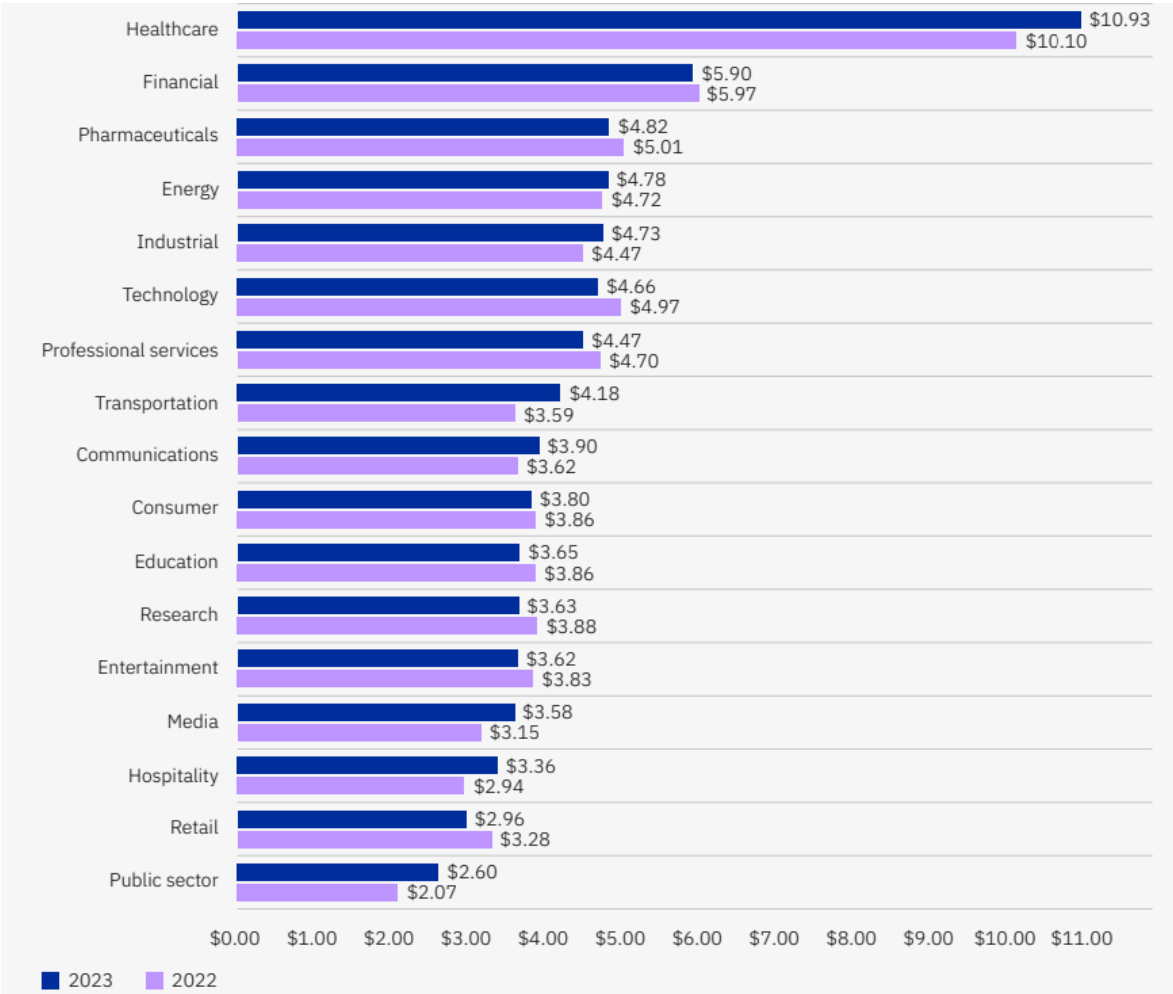

Figure 12 : Coût moyen d'une violation des données en millions de dollars par secteur pour 2022 et 2023 (Source : Cost of a Data Breach Report 2023, IMB Security[38])

[37] La Health Insurance Portability and Accountability Act de 1996 (HIPAA) et les réglementations émises en vertu de hipaa sont un ensemble de lois américaines sur les soins de santé qui établissent des exigences en matière d'utilisation, de divulgation et de protection des informations médicales identifiables individuellement.

[38] Source : https://www.ibm.com/reports/data-breach

3.5 Risque cyber par région

- **En termes de fréquences des sinistres**

En termes de nombre d'attaques, les entreprises américaines demeurent la cible privilégiée des attaquants. Le fournisseur en solutions et services de sécurité Specops, en se basant sur des données issues du CSIS (Center for Strategic and International Studies), a publié le classement ci-dessous, des pays ayant subi le plus de cyberattaques visant des gouvernements ainsi que des entreprises technologiques et de défense « significatives », c'est à dire ayant chacune causée des pertes de plus d'un million de dollars, entre mai 2006 et juin 2020. Sur les 14 dernières années, ce sont les Etats-Unis qui ont connu le plus grand nombre de cyberattaques de grande ampleur (156), loin devant la Grande-Bretagne (47) et l'Inde (23).

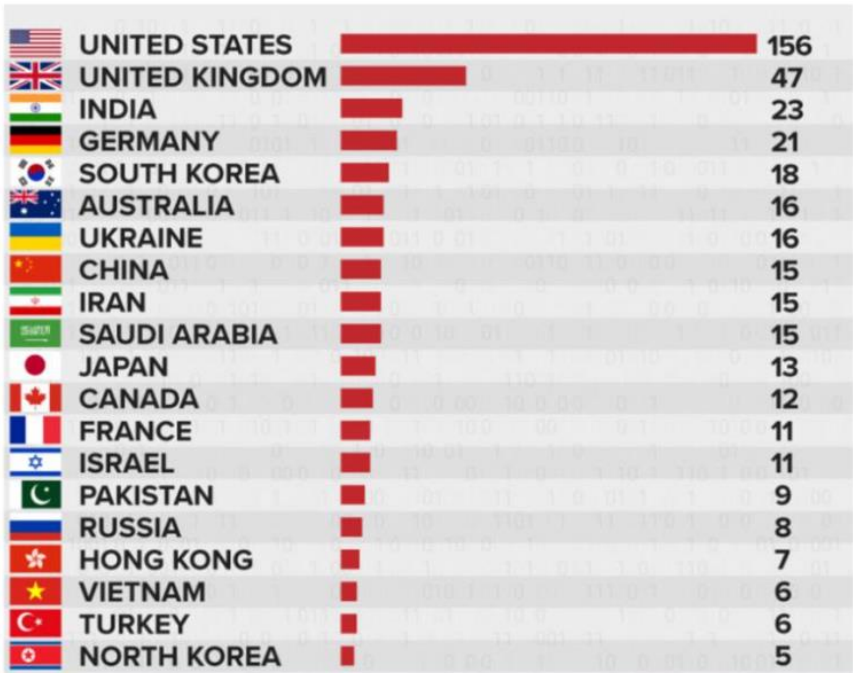

Figure 13 : Classement des pays touchés par une cyberattaque significative de mai 2006 et juin 2020 (Source : Specops/CSIS[39])

- **En termes de coûts des sinistres**

Les rapports d'IBM sur les violations des données de 2022 et 2023, soulignent une différence régionale des niveaux des coûts des incidents. Dans le graphe ci-dessous, les violations des données en 2023 par exemple, sont plus coûteuses aux Etats-Unis par rapport aux autres pays, avec un coût moyen de 9,48 millions de dollars. Le Moyen-Orient vient en seconde position, avec une moyenne de 8,07 millions de dollars, probablement à cause des tensions géopolitiques omniprésentes, ce qui peut être exploité par les acteurs malveillants pour mener des attaques. En troisième position, vient le Canada avec un coût moyen de 5,13 millions de dollars en 2023.

[39] Source : https://specopssoft.com/blog/countries-experiencing-significant-cyber-attacks/

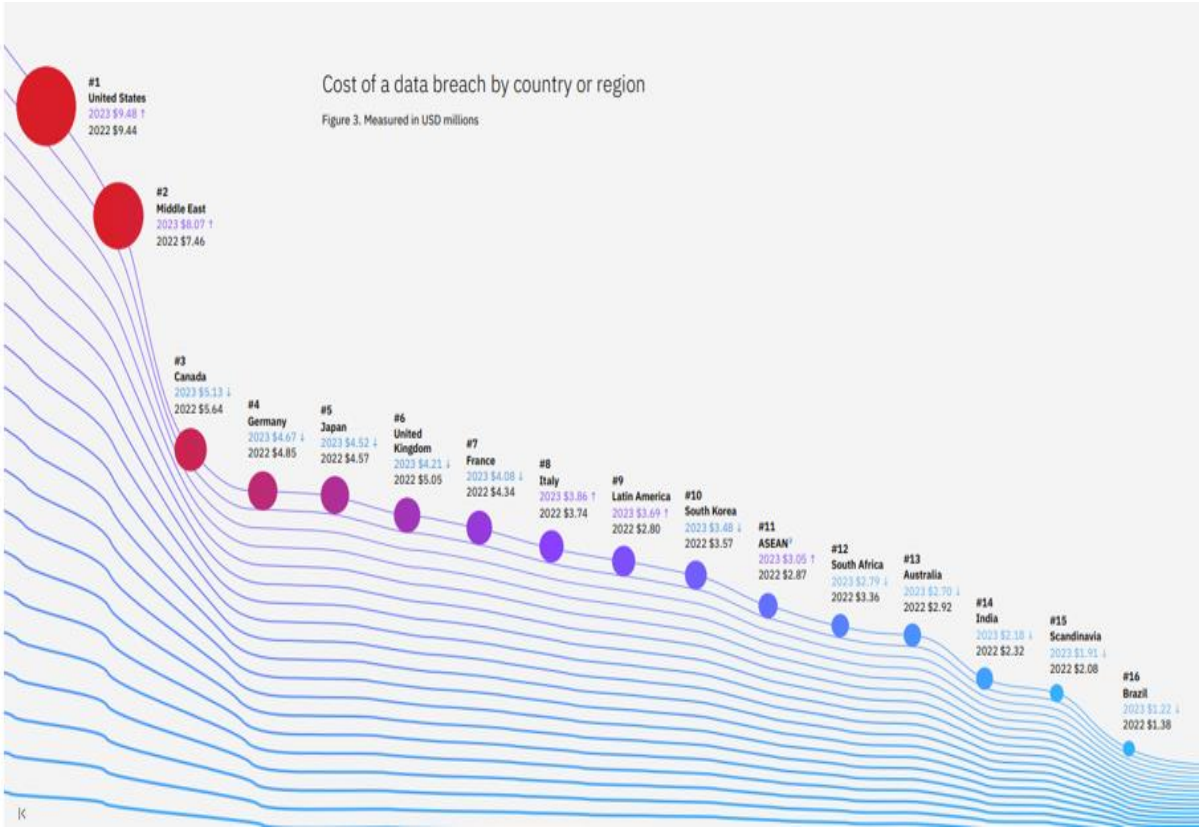

Figure 14 : Coût moyen d'une violation des données en millions de dollars par pays ou région pour 2022 et 2023 (Source : Cost of a Data Breach Report 2023, IMB Security[40])

Ainsi, il est observé que les États-Unis occupent la première position en tant que pays le plus impacté par les cyberattaques, que ce soit en matière de fréquence ou de coûts.

[40] Source : https://www.ibm.com/reports/data-breach

Chapitre 4 : Marché de la cyber-assurance

Compte tenu des niveaux de sinistres liés au risque cyber et le risque croissant généré par ce dernier, un marché assurantiel a commencé à émerger et à se développer. Bien que la mise en place de mesures de sécurité et de bonnes pratiques soit indéniablement essentielle, les gestionnaires des risques devraient également explorer d'autres options pour faire face aux risques cyber résiduels. Une alternative est le transfert du risque dans le cadre de l'assurance. Les premiers contrats de la cyber-assurance apparaissent aux Etats-Unis en 1998. Depuis, le marché ne cesse de croître à travers le monde. Assurer le risque cyber, s'accompagne d'une panoplie de défis : des formes en constante évolution, peu des données sur la sinistralité survenue, des charges des sinistres incertaines, des niveaux complexes d'interconnexion des systèmes informatiques, et la liste est longue.
Le développement du marché de l'assurance cyber est partiellement freiné par le coût jugé élevé des primes des polices, la confusion quant à l'étendue de la couverture et l'incertitude quant à la probabilité d'une attaque.

4.1 Le cyber en assurance, un risque presque comme les autres ?

Le risque cyber en assurance présente des caractéristiques d'un risque traditionnel en assurances dommages : il est incertain et imprévisible. Pourtant, il présente des caractéristiques qui lui sont propres. Il est systémique au sens où il peut propager et toucher simultanément plusieurs assurés. Il est également mal connu. Or, l'assureur appuie son estimation du risque sur la base des pertes antérieures.

Les formes des risques cyber sont nombreuses et en constante évolution. C'est donc un facteur à tenir en compte lors de la conception des produits de l'assurance-cyber. La classification des événements et risques cyber doit se faire dans un bon niveau de granularité. La séparation des événements est aussi importante, par exemple séparer le terrorisme cyber du terrorisme dans son sens le plus large.
Une définition de ce risque peut devenir obsolète au bout d'un certain temps. Une revue continue de la définition du risque, et une mise à jour des garanties des polices proposées s'avèrent nécessaires.

Contrairement aux produits d'assurances classiques, l'exposition des risques cyber dépassent les pertes physiques directes pour couvrir des pertes de type arrêt de l'activité, coût de réputation, mais aussi une exposition liée à la responsabilité civile suivant la violation des données à caractère personnel par exemple, des pénalités et amendes. Un cernement de l'exposition et des pertes potentielles au moment de la souscription, permet de proposer un tarif adéquat.

Pour un assureur cyber, le risque d'accumulation des pertes est critique. La diversification industrielle, géographique, clientèle reste la méthode la plus simple pour gérer ce risque, mais ce n'est pas suffisant, vu la complexité du risque et l'interconnexion des systèmes informatiques. Les assureurs sont donc encouragés à analyser les zones d'accumulation probables et développer une approche permettant d'analyser comment et à quelle échelle leurs portefeuilles pourraient être impactés.

4.2 De la couverture implicite à la couverture explicite

Pour l'assureur comme pour l'assuré, deux cas de figure se présentent :

- **Les garanties silencieuses ou implicites :** correspondent à la prise en charge des pertes et dommages consécutifs à la réalisation d'un risque cyber par des polices d'assurance traditionnelles, c'est-à-dire, biens, risques divers et autres sinistres qui ne sont pas des sinistres cyber. Elles sont qualifiées de silencieuses parce que leur couverture n'est ni explicitement mentionnée ni explicitement exclue du contrat, puisqu'elles n'ont initialement pas été rédigées pour couvrir les risques cyber. Ainsi, à titre d'exemple, le contrat qui couvre les pertes d'exploitation, sans condition de mise œuvre de la garantie à l'existence d'un dommage matériel empêchant la poursuite de l'exploitation, peut être sollicité lorsque l'entreprise assurée est dans l'incapacité de commercialiser ses produits à cause d'une attaque en déni de service. De même, cela s'applique lorsque l'assureur a fourni une couverture « tous risques sauf… » et n'a pas expressément exclu le risque cyber.

- **Les garanties cyber pures ou affirmatives :** sont délivrées explicitement et spécifiquement pour couvrir le risque cyber. Par exemple, une telle garantie pourrait indemniser une entreprise assurée pour les pertes financières résultant d'une violation de données, de la perte d'informations confidentielles, ou des coûts liés à la restauration des systèmes après une cyberattaque.

Au sein des contrats cyber affirmatifs ou silencieux, il faut aussi différencier les couvertures dites « First Party » des couvertures « Third party » :

- **les couvertures dommages propres (ou First Party) :** cette catégorie de contrats couvre les pertes subies sur les biens de l'assuré, par exemple le vol ou la destruction de données, l'extorsion de fonds ou encore l'interruption d'activité.

- **les couvertures responsabilité civile (ou Third Party) :** cette catégorie de contrats indemnise les pertes subies par un tiers au titre de la responsabilité civile de l'assuré engagée à la suite d'un évènement cyber, par exemple une fuite de données personnelles ou une erreur dans un programme informatique causant des pertes à des tiers.

4.3 La cyber-assurance dans le monde

A fin 2022, plus de 220 groupes assureurs souscrivent au niveau mondial des risques cyber, d'après une étude réalisée par la société de notation et d'analyse des assurances Insuramore[41]. La même étude indique que le volume global des primes de la cyber-assurance a approché 14 milliards de dollars en 2022, contre 8,6 milliards de dollars une année auparavant, avec les Etats-Unis qui comptabilisent plus de la moitié des primes totales.

4.4 La cyber-assurance en France

En France, le marché de la cyber-assurance fait état d'un retard dans la maîtrise du risque cyber par rapport au marché américain. Selon la 2ème édition 2022 de l'étude « LUmière sur la CYberassurance » (LUCY) de l'AMRAE (l'Association pour le Management des Risques et des

[41] Source : https://www.insuramore.com/rankings/insurers/premiums-cyber/

Assurances de l'Entreprise)[42], la sous-couverture des entreprises françaises constitue une menace pour le tissu économique tout entier.

Cependant, selon la 3ème édition de LUCY publiée par l'AMRAE en mai 2023[43], le marché de la cyber-assurance en France gagne en maturité. A la fin de 2022, le nombre d'entreprises assurées a augmenté de 25%. Le volume global des primes cyber en France en 2022 s'élève à 316 millions d'euros, enregistrant une hausse de 72% par rapport à 2021, tandis que la sinistralité historiquement basse atteint un niveau de 70,80 millions d'euros, en baisse de 57% par rapport à 2021. Les assureurs bénéficient ainsi d'une marge de manœuvre grâce à un ratio sinistres/primes de 22,3% pour l'exercice, notamment après l'année fortement déficitaire de 2020. A partir de 2021, les assureurs, ont opté pour des conditions de souscription plus strictes, impliquant une augmentant des taux de prime, une réduction des capacités et une hausse des franchises. Egalement, la sinistralité a connu une très forte baisse en 2022, après le pic de 2020, lié surtout à l'intensité des sinistres survenus.

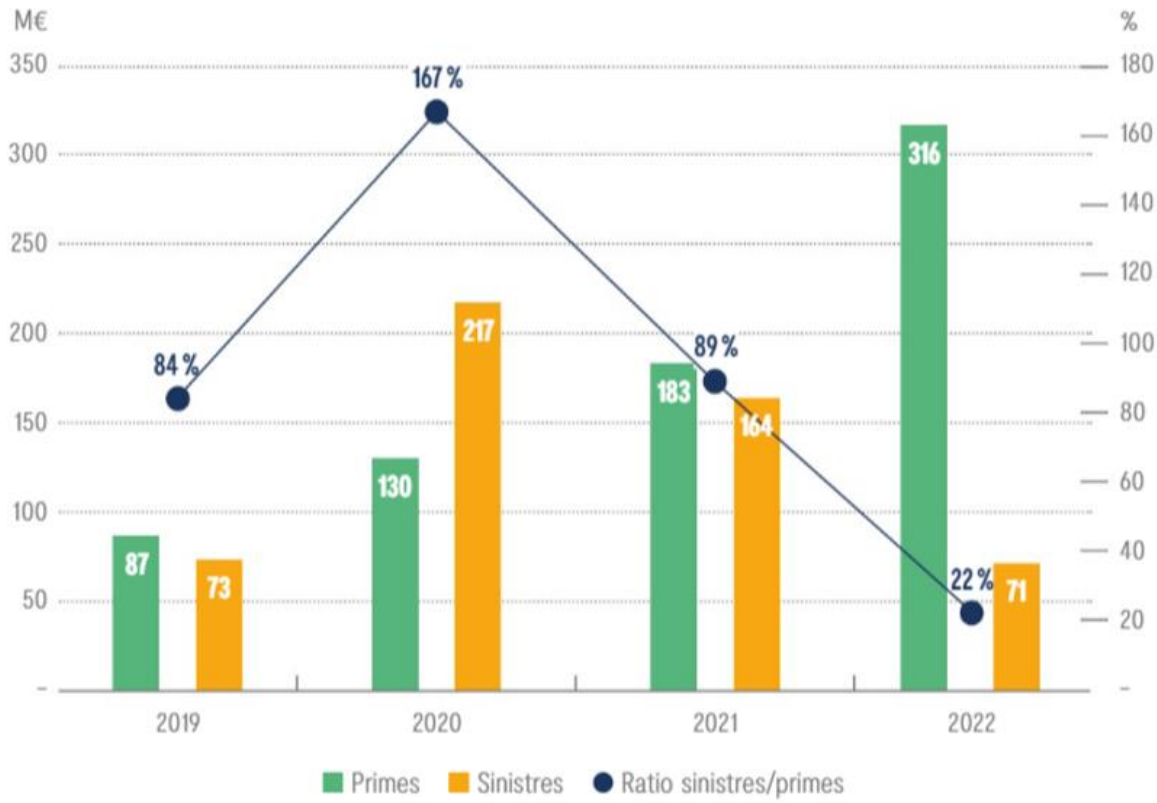

Figure 15 : Le marché français de l'assurance cyber (Source : 2ème édition LUCY par AMRAE[44])

[42] Source : https://www.amrae.fr/recherche?type=All&search_api_fulltext=2%C3%A8me+%C3%A9dition+lucy
[43] Source : https://www.amrae.fr/bibliotheque-de-amrae/lucy-lumiere-sur-la-cyberassurance-amrae-mai-2023
[44] Source : https://www.amrae.fr/bibliotheque-de-amrae/lucy-lumiere-sur-la-cyberassurance-amrae-juin-2022

D'après la même étude, le marché de la cyber-assurance en France présente des disparités marquées en fonction de la taille des entreprises. Il est principalemengt porté par le segment des grandes entreprises, qui représentent à lui seul 83 % du volume total de primes versées au titre de la garantie cyber en 2022, ainsi que par le segment des entreprises de taille intermédiaire. Ces deux segments semblent parvenir à une maturité avec une augmentation des souscriptions et une sinistralité relativement faible par rapport au niveau de souscription. En revanche, le segment des entreprises de taille moyenne a connu une tendance similaire à celle observée en 2020 pour les grandes entreprises et en 2021 pour les entreprises de taille intermédiaires. Il a enregistré une augmentation significative de la sinistralité par rapport aux niveaux de souscription, entraînant le ratio prime/sinistralité à un niveau élevé, atteignant 100 %.

4.5 Facteurs favorisant la croissance de la cyber-assurance

Cependant, plusieurs facteurs sont susceptibles d'entraîner une croissance du marché de la cyber-assurance :

- Premièrement, **la médiatisation des incidents cyber**, qui montre l'ampleur des pertes potentielles d'une cyberattaque et accroît la prise de conscience du risque associé.

- Deuxièmement, **la réglementation** tant aux États-Unis, par exemple, le décret présidentiel « Executive Order on Improving the Nation's Cybersecurity » promulgué en mai 2021[45], avec des mesures coercitives pour les entreprises et visant à améliorer la cybersécurité nationale, qu'en Europe, par exemple la mise en pratique du règlement général européen sur la protection des données à partir de 2018[46], qui augmentent le coût des violations de données. La réglementation renforce la responsabilisation des entreprises dans la protection des données et incite à mettre en place des mesures de sécurité et impose la notification d'incidents.

- Troisièmement, **les modifications des polices d'assurance traditionnelles**. L'exclusion du risque cyber dans le cadre des polices, va accroître la nécessité d'une assurance cyber dédiée. L'autorité européenne, l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) et la Direction générale du Trésor s'accordent sur la nécessité de poursuivre l'effort de clarification des contrats d'assurance traditionnels. A titre illustratif, à partir du 1er janvier 2023, AXA exclut les garanties cyber de ses contrats responsabilité civile. Le Lloyd's de Londres durcit, pour sa part, les règles d'exclusion dans le but de se prémunir contre un éventuel risque systémique. Les attaques soutenues par des Etats seront éliminées de son champ de couverture à compter du 1er mars 2023.[47]

[45] Le décret présidentiel "Executive Order on Improving the Nation's Cybersecurity" est une initiative gouvernementale des États-Unis visant à renforcer la cybersécurité nationale. Il a été signé par le président Joe Biden le 12 mai 2021.

[46] Le règlement général de protection des données (RGPD) est un texte réglementaire européen qui encadre le traitement des données de manière égalitaire sur tout le territoire de l'Union européenne (UE). Il est entré en application le 25 mai 2018.

[47] Source : https://www.atlas-mag.net/category/tags/focus/le-marche-mondial-de-la-cyberassurance

Conclusion

Au cours des dernières années, les incidents cyber se sont accrus et les rapports se font de plus en plus alarmants sur l'ampleur des attaques dirigées contre les entreprises. Malgré la multiplication des solutions de protection développées par l'industrie depuis les premières cyberattaques, leur mise en œuvre seule ne suffira pas à éliminer le risque cyber. Les compagnies d'assurance, qui assument une partie des risques de leurs clients, jouent un rôle crucial dans ce contexte. Aujourd'hui, l'offre de la cyber-assurance est encore peu développée, avec des obstacles structurels au développement de ces assurances, notamment, le manque d'historicité sur les impacts des sinistres cyber. Une approche potentielle pour surmonter ces défis consiste à recourir à des données externes sur le risque cyber, permettant ainsi de construire des statistiques fiables et indépendantes, offrant une meilleure analyse du risque cyber.

Deuxième partie

Revue de la littérature et présentation des données

Cette partie a pour but de faire un tour d'horizon de la littérature existante au sujet de la modélisation actuarielle du risque cyber. Nous procédons ensuite à la présentation de la base de données utilisée dans le cadre de ce mémoire et de la méthodologie suivie pour l'identification des sinistres cyber au sein de celle-ci.

Chapitre 5 : Etudes académiques sur la modélisation du risque cyber

5.1 Données publiques disponibles

Le manque de données et le recul sur le risque cyber expliquent le peu d'études quantitatives sur ce sujet. Les données cyber sont dans la majorité des cas non publiques, car les entreprises victimes préfèrent ne pas révéler ces incidents de peur de nuire à leur réputation. De plus, l'absence d'une définition standard du risque cyber complique la collecte des sinistres y afférents.

Néanmoins, les violations des données sont un des risques cyber les mieux documentés. Une des principales bases de données largement utilisée dans ce contexte est la « Chronology of Data Breaches » par Privacy Rights Clearinghouse (PRC[48]). Ce jeu de données recense les incidents de fuites de données à caractère personnel aux Etats-Unis, accompagnés du nombre d'enregistrements compromis pour chaque incident. Les études académiques existantes cherchent principalement à prévoir le nombre d'enregistrements compromis à la suite d'une violation de données personnelles en fonction des caractéristiques d'une organisation.

Un exemple d'utilisation de la base PRC, est l'étude « Cyber Risk Management Strategies » de Marco PIRRA (2023)[49], qui modélise le nombre d'enregistrements compromis par un modèle de comptage négatif binomial zéro inflation (ZINB), adapté à la présence excessive des zéros dans la base PRC. Cette étude arrive notamment à la conclusion qu'il existe une corrélation entre le nombre de transactions en bitcoin et le nombre des futures violations des données. En tant que première devise numérique, le bitcoin est utilisé par les cyber criminels pour monétiser leurs attaques.

Le risque cyber est aussi le risque opérationnel lié à l'information ou aux systèmes informatiques (section 1.1). Pour comprendre le risque cyber, il est avisé d'exploiter les bases de données consacrées au risque opérationnel. Ces ressources permettent de couvrir un spectre plus large d'incidents cyber par

[48] Privacy Rights Clearinghouse est une organisation à but non lucratif basée aux États-Unis qui se consacre à la protection de la vie privée des consommateurs.
[49]Source :https://www.actuaries.org/IAA/Documents/SECTIONS/JointColloquium2022/Day3_AFIRERM_Pirra.pdf

rapport, par exemple, à la base Chronology of Data Breaches, qui se concentre uniquement sur les violations de données.

Les données SAS OpRisk Global data constituent le plus grand recueil au monde des pertes opérationnelles publiquement déclarées, excédant 100 000 dollars. Ces données présentent plusieurs avantages notables, notamment leur fiabilité et leur exhaustivité. En effet, chaque incident répertorié a été confirmé par au moins une source médiatique majeure, garantissant ainsi la traçabilité et la vérifiabilité de ces informations. Ces données sont collectées à partir des sources publiques et sont maintenues à jour par les équipes de SAS selon des normes strictes de qualité des données.

Le risque opérationnel a été examiné de près par les autorités réglementaires dans le cadre de Bâle II à la fin des années 90. Cela est pertinent dans la mesure où les données ont été systématiquement collectées assez récemment, entraînant donc un biais de déclaration dans toutes les données sur le risque opérationnel.

5.2 Quelques études menées avec la base de données SAS OpRisk Global data

Il est à noter que la base de données SAS OpRisk Global Data a fait l'objet de quelques études académiques, aussi bien sur le risque opérationnel global que sur le risque cyber spécifiquement. Nous citons, à titre d'exemples, les études suivantes :

Le risque opérationnel global :

- « An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates » par Chavez-Demoulin, Embrechts et Hofert (2015)[50],
- « Using SAS® OpRisk Global Data to Improve Decision-Making at a Bank » par Gericke et Raubenheimer (2020)[51].

Le risque cyber spécifiquement :

- « Insurability of Cyber Risk: An Empirical Analysis » par Bienner, Eling et Wirfs (2014)[52],
- « Modelling and Management of Cyber Risk » par Eling et Wirfs (2015)[53],
- « What are the actual costs of cyber risk events ? » par Eling et Wirfs (2019)[54],
- Le mémoire d'actuariat de MARTINEZ (2019) intitulé « Modélisation assurantielle du risque cyber »[55].
- « Heterogeneity in cyber loss severity and its impact on cyber risk measurement » par Eling et Jung (2022)[56]

[50] Source : https://www.jstor.org/stable/43998282
[51] Source : https://support.sas.com/resources/papers/proceedings20/5069-2020.pdf
[52]Source : https://www.internationalinsurance.org/sites/default/files/2018-03/Insurability%20of%20Cyber%20Risk.pdf
[53] Source : https://www.actuaries.org/oslo2015/papers/iaals-wirfs&eling.pdf
[54] Source : https://www.sciencedirect.com/science/article/abs/pii/S037722171830626X
[55] Source : https://www.institutdesactuaires.com/docs/mem/d39dd709e4ba0bb10e621651a7c8fbcb.pdf
[56]Source :https://www.researchgate.net/publication/361117286_Heterogeneity_in_cyber_loss_severity_and_its_impact_on_cyber_risk_measurement

5.3 Focus sur les études cyber avec SAS OpRisk Global data

S'ils l'on se concentre sur les études ayant trait au risque cyber, celle de Bienner, Eling et Wirfs (2014) cherche à répondre à la question de l'assurabilité du risque cyber. En analysant les caractéristiques statistiques des sinistres cyber, extraits de la base SAS OpRisk Global data, il est constaté que les risques cyber présentent des caractéristiques distinctes par rapport aux autres risques opérationnels de la base de données, et donc doivent être traités séparément. Ils concluent que le risque cyber peut être assurable selon l'approche introduite par Berliner[57] (1982), qui permet de différencier entre les risques assurables et non assurables, basée sur les conditions actuarielles, les conditions de marché, et les conditions sociétales. Cependant, ils rajoutent que le manque de données, le caractère évolutif du risque cyber et l'asymétrie d'information (par exemple, l'antisélection et l'aléa moral), entravent le développement du marché de la cyber-assurance.

A travers une analyse empirique approfondie du risque cyber, Eling et Wirfs (2015), confirment que les risques cyber sont très différents par rapport aux autres risques opérationnels, et donc les deux types de risques nécessitent des modélisations distinctes. En effet, les pertes des risques cyber identifiés dans la base sont beaucoup plus faibles que celles des risques non cyber. La distribution des risques non cyber montre des valeurs de VaR (Value at Risk) et de TVaR (Tail Value at Risk) bien plus élevées que celle de des risques cyber.

Dans une étude ultérieure, Eling et Wirfs (2019) explorent la possibilité d'appliquer les lois usuelles (exponentielle, gamma, log-normale, ...), utilisées pour la modélisation des charges en assurance, aux risques cyber. Il est constaté qu'aucune des distributions paramétriques simples ne modélise de manière adéquate les pertes, nécessitant l'utilisation de l'approche de modélisation de la théorie des valeurs extrêmes (TVE). La distribution de Pareto généralisée (GPD) offre le meilleur ajustement aux pertes cyber mesurées par les valeurs de log-vraisemblance et le critère d'information d'Akaike (AIC[58]). De plus, leur constatation principale met en évidence la variabilité significative du risque cyber par rapport aux autres risques, mettant en lumière le rôle crucial du comportement humain dans la genèse des risques cyber et complexifiant la capacité à prédire ce risque. De plus, les pertes sont influencées par le type d'événement et des spécificités propres à l'entreprise telles que le pays, l'industrie ou la taille.

Dans certains cas, l'incident cyber peut se solder par de lourdes pertes financières. En utilisant la base de données SAS OpRisk Global data, Strupczewski (2019)[59] conduit une analyse de la queue de distribution des pertes des risques cyber, et emploie particulièrement la théorie des valeurs extrêmes et conclut que la distribution de Pareto généralisée (GPD) est la meilleure pour modéliser les risques cyber extrêmes.

Dans son mémoire d'actuariat intitulé « Modélisation assurantielle du risque cyber », Martinez (2019) construit un modèle de prédiction du coût d'un incident cyber selon les caractéristiques propres d'une

[57] Cette approche, basée sur neuf critères d'assurabilité, est fréquemment utilisée pour analyser les marchés et les produits d'assurance. Les critères sont catégorisés en trois grandes catégories classifiant les risques en termes de conditions actuarielles, de conditions de marché et de conditions sociétales.
[58] Le critère AIC est présenté en détail dans la section 8.5.
[59]Source :https://www.researchgate.net/publication/332461145_What_Is_the_Worst_Scenario_Modeling_Extreme_Cyber_Losses

entreprise et le type d'incident. Il utilise des modèles de régression, des arbres de régression et des forêts aléatoires. Cependant, la séparation des sinistres cyber extrêmes des autres sinistres attritionnel n'est pas conduite. Par conséquent, les modèles construits ne peuvent pas bien capter ces cas extrêmes.

Un modèle de régression Tweedie est utilisé par Eling et Jung (2022)[60], pour la modélisation de la sévérité des pertes liées aux risques cyber des entreprises du secteur financier. Leur analyse met en évidence plusieurs facteurs associés à des pertes liées au cyber plus importantes et plus sévères au sein des entreprises du secteur financier.

Après avoir présenté les principales études du risque cyber utilisant la base SAS OpRisk Global data, ce jeu de données sera présenté en détail dans la section suivante, ainsi que la méthodologie d'identification des incidents cyber au sein de la base de données.

[60]Source :https://www.researchgate.net/publication/361117286_Heterogeneity_in_cyber_loss_severity_and_its_impact_on_cyber_risk_measurement

Chapitre 6 : Présentation des données

6.1 Base de données

Nous utilisons le jeu de données SAS OpRisk Global data extraite en juillet 2023. Cette base de données étant constamment alimentée par SAS, il est en effet important de noter la date limite d'extraction dans le cadre de cette étude. Elle recense 39 573 pertes opérationnelles survenues entre mars 1971 et juin 2023 et concernant des entreprises de différents pays et industries.

La base SAS OpRisk Global data comprend 49 variables. Pour chaque sinistre, elle apporte des informations en relation avec la ligne métier (business Line), la catégorie d'événement, la situation économique et géographique de l'entreprise victime, les informations sur la société mère en cas de groupe, la localisation géographique et enfin, les coûts associés à ces incidents. Les différentes variables seront détaillées par la suite de cette partie.

Etant donné que le risque opérationnel est régi réglementairement par Bâle II[61], nous retrouvons naturellement dans la base les catégories d'évènements et les lignes métier selon les classifications bâloises.
Des informations relatives à l'identification des entreprises, telles que les références, les noms des entreprises et éventuellement le nom de l'entreprise mère en cas de groupe, sont disponibles dans la base de données, mais sans importance pour notre modélisation. La liste complète des variables en annexe (Annexe A).

La base de données fournit des informations détaillées sur la ligne et la sous-ligne métier (business line) de chaque établissement conformément aux normes de la Banque des règlements internationaux ou « Bank for International Settlements » (BIS), le code et le nom du secteur d'industrie. Elle comprend également des informations sur la région du domicile, qu'il s'agisse de la région de l'entreprise ou de celle de l'entreprise mère en cas de groupe. Des données sur la taille de l'entreprise sont aussi disponibles, telles que le nombre de salariés, le chiffre d'affaires et les fonds propres.

Les sinistres sont classés selon les sept catégories de classification des événements opérationnels définies par le comité de Bâle[62] :

- **Fraude interne :** par exemple, des informations inexactes sur les positions, des falsifications, des vols commis par un employé et des délits d'initié perpétrés par un employé agissant pour son propre compte ;
- **Fraude externe :** cela inclut des événements comme des braquages, des faux en écriture et des dommages causés par le piratage informatique ;
- **Pratiques en matière d'emploi et sécurité sur le lieu de travail :** cela concerne des problématiques telles que le non-respect des règles de santé et sécurité au travail, les violations

[61] Les normes Bâle II (le second accord de Bâle) constituent un dispositif prudentiel destiné à mieux appréhender les risques bancaires et principalement le risque de crédit ou de contrepartie et les exigences, pour garantir un niveau minimum de capitaux propres, afin d'assurer la solidarité financière. Dans le cadre du dispositif Bâle II, la définition du risque opérationnel, les procédures à mettre en place pour le limiter et les méthodes de quantification ont été normalisées.
[62] Source : https://www.bis.org/publ/bcbs128.pdf

des règles en matière d'enmploi, les plaintes pour discrimination et la responsabilité civile en général ;

- **Clients, produits et pratiques commerciales :** cela englobe des violations de l'obligation fiduciaire, l'utilisation frauduleuse d'informations confidentielles sur la clientèle, des opérations boursières malhonnêtes pour le compte de la banque, le blanchiment d'argent et la vente de produits non autorisés ;
- **Dommages aux actifs corporels :** cela concerne des actes tels que le terrorisme, le vandalisme, les incendies et les inondations ;
- **Dysfonctionnement de l'activité et des systèmes :** par exemple, les pannes de matériel et de logiciels informatiques, les problèmes de télécommunications et les pannes d'électricité ;
- **Exécution, livraison et gestion des processus :** cela inclut des erreurs d'enregistrement des données, des défaillances dans la gestion des sûretés, des lacunes dans la documentation juridique, des erreurs d'accès aux comptes des clients et des défaillances des fournisseurs ou des conflits avec eux.

Les pertes dans cette base représentent une estimation des coûts, à la fois directs et indirects des événements. Cependant, la perte de réputation due à l'événement n'est pas incluse, car ce type de perte est exclu par définition du risque opérationnel selon Bâle II[63]. Toutes les pertes sont en dollar américain. Une limite de cette base est qu'elle ne couvre que les pertes supérieures à 100 000 dollars. Néanmoins, étant donné que les coûts des sinistres cyber sont souvent élevés (à titre illustratif, le coût moyen d'une violation des données en 2023 est estimé à 4,45 millions de dollars, comme mentionné dans la section 1.3.6), la censure des pertes devient moins préoccupante. Par ailleurs, les montants des pertes sont ajustés par l'indice des prix à la consommation américain pour prendre en compte l'inflation, ce qui permet une meilleure comparabilité. La base fournit l'indice des prix à la consommation américain ou le « Consumer Price Index » (CPI)[64], qui mesure la variation des prix à la consommation, autrement dit de l'inflation, et est utilisé dans ce contexte.

La base de données renseigne également le pays où est situé le siège social de l'entreprise ainsi que le pays où s'est produit l'incident. Elle fournit des dates clés des événements, telles que l'année de début et de fin du sinistre.

Chaque incident est accompagné d'un texte décrivant les circonstances de survenance et les problèmes qui en découlent. Compte tenu que les variables précédentes ne sont pas suffisantes, nous allons utiliser ces textes pour identifier les risques cyber.

6.2 Identification des incidents cyber dans SAS OpRisk Global Data

6.2.1 Revue de la littérature

La base de données SAS OpRisk Global Data est la plus grande collection des pertes opérationnelles déclarées publiquement. Néanmoins, les sinistres cyber au sein de la base ne sont pas explicitement identifiés. Afin de les déterminer, Bienner, Eling et Wirfs (2014) élaborent une méthodologie décrite dans leur rapport intitulé « Modelling and Management of Cyber Risk ». Ils établissent trois critères pour définir un événement cyber :

[63] La définition du risque opérationnel donnée par Bâle II inclut le risque juridique (notamment le risque d'amendes, de pénalités, de dommages et intérêts), mais exclut les risques stratégiques et les risques de réputation.
[64] Aux États-Unis, l'indice des prix à la consommation (CPI) mesure les variations dans les prix payés par les consommateurs pour un panier de biens et de services.

1. **« Critical asset » ou Actif important :** il s'agit de l'actif impacté par l'incident cyber, pouvant être un système informatique, un serveur, un site internet, un disque dur, des données confidentielles, un secret commercial,…
2. **« Actor » ou Acteur :** désignant l'acteur à l'origine de l'incident. Quatre types d'acteurs peuvent être impliqués :
    - « Actions by people » ou actions humaines : actes criminels tels que l'hameçonnage ou la fraude, ou des erreurs accidentelles telles que des erreurs humaines ;
    - « System and technical failure » ou défaillances système et techniques : pannes techniques des systèmes ;
    - « Failed internal processes » ou échecs de processus internes : procédures internes inadéquates ;
    - « External events » ou événements externes : comme les catastrophes naturelles, par exemple, les inondations et les tempêtes ;
3. **« Outcome » ou Résultat** : indiquant la conséquence de l'incident, par exemple, une fuite de données ou une interruption d'un système ou d'une activité.

Ensuite, les deux chercheurs ont défini des familles de mots-clés pour chacun des trois critères. Enfin, ils ont cherché les familles de ces mots-clés dans les descriptifs des événements disponibles dans la base de données. Un sinistre est donc considéré comme cyber, si son descriptif comporte au moins un terme appartenant à chaque famille de mots des trois critères.

Eling et Wirfs ont continué à utiliser cette méthodologie dans leurs études antérieures. Cependant, dans le cadre de son mémoire d'actuariat (2019), Martinez estime que cette méthodologie n'est pas suffisante pour identifier tous les sinistres cyber dans la base SAS OpRisk Global data. Par conséquent, il a complété cette démarche par une fouille de texte plus pointue, en recherchant des expressions spécifiques liées au risque cyber et des termes associés à chaque catégorie d'incident cyber. Il suppose qu'il existe trois catégories distinctes d'incidents cyber :

- Les incidents cyber d'origine criminelle externe, représentant les actes intentionnels perpétrés par des criminels extérieurs à l'entreprise victime ;
- Les incidents cyber d'origine criminelle interne, englobant les actions intentionnelles menées par des criminels qui travaillent ou ont déjà travaillé dans l'entreprise victime ;
- Les incidents cyber d'origine accidentelle, incluant à la fois les erreurs humaines et les événements naturels.

Pour chacune des deux catégories d'incidents énumérées ci-dessus, une liste des types d'incidents potentiels est élaborée. Pour chacun de ces types, une famille de mots-clés est établie. Pour identifier les incidents cyber de nature accidentelle, des filtres sont appliqués aux variables « Event risk category », « Sub risk category » et « Activity » de la base SAS OpRisk Global data.

Après avoir examiné la littérature existante sur la modélisation actuarielle du risque cyber, nous présenterons la méthodologie construite pour identifier les risques cyber à partir de la base de données.

6.2.2 Procédure d'identification des sinistres cyber

Pour identifier les risques cyber dans la base SAS OpRisk Global Data, une méthode basée sur la fouille du texte dans la description de chaque sinistre a été conduite. La fouille de texte est particulièrement

utile pour traiter de grandes quantités de données textuelles, ce qui peut être fastidieux, voire impossible à faire manuellement. De plus, la méthodologie élaborée par Eling et Wirfs, comme présentée dans la section précédente, s'avère insuffisante et sous-estime les incidents cyber de la base de données. Ainsi, nous procédons selon une méthodologie spécifique.

Initialement, un dictionnaire de mots-clés liés au risque cyber est constitué. Ces mots-clés sont rassemblés à partir de divers sites Internet traitant des risques cyber. Cette liste est ensuite complétée par celle construite par Eling et Wirfs (2014). Au total, le dictionnaire comprend 209 mots-clés. Le dictionnaire est en annexe (Annexe B).

Préalablement à la recherche de ces mots-clés, une étape de traitement des descriptifs des incidents a été effectuée pour réduire leurs tailles et optimiser le temps de l'algorithme de la fouille du texte.

La première étape du traitement des descriptifs est la **tokenisation** qui consiste à découper un texte en unités linguistiques plus petites appelées tokens. Un token peut être un mot, un caractère, ou un sous-mot. Compte tenu des différentes structures linguistiques existantes, la tokenisation varie d'une langue à une autre. Pour cette base de données en anglais, la ponctuation ainsi que les caractères alphanumériques sont conservées comme des "tokens", et doivent donc être éliminés.
Certains mots contiennent des majuscules. Toutes les majuscules sont alors réduites puis la ponctuation et les caractères non numériques comme les dates et les chiffres sont supprimés. Les stopwords ou mots vides, c'est-à-dire tous les mots n'ayant pas de réelle signification tels que les mots de liaison, sont aussi supprimés. Ce processus qui essaie d'harmoniser les tokens s'appelle la **normalisation**.

La dernière étape du traitement des descriptifs est la **lemmatisation**. Il s'agit de faire correspondre à chaque mot, son lemme ou encore la forme canonique du mot. A la différence de la racinisation (en anglais « stemming ») qui est le fait de réduire un mot à sa racine, la lemmatisation a l'avantage de sauvegarder la signification du mot en fonction de son contexte.
Enfin, tous les doublons sont supprimés pour éviter de compter doublement les mots-clés lors de la recherche.

Le traitement des descriptifs est réalisé à l'aide d'un code Python. La fouille des termes du dictionnaire est effectuée et génère un score nommé « Score 1 » pour chaque incident, indiquant le nombre de mots-clés retrouvés dans le descriptif retraité. Les scores varient sur l'ensemble de la base de données, de 0 à 22. Intuitivement, on pourrait penser que les sinistres ayant un faible score ne sont pas des risques cyber, tandis que ceux ayant un score élevé en sont. Cependant, après vérification aléatoire de quelques incidents à faible score, certains se sont avérés être des risques cyber.

Pour affiner la détection, un second dictionnaire plus restreint est construit, à partir du premier dictionnaire, en conservant uniquement les mots-clés les plus spécifiques au risque cyber. Ce deuxième dictionnaire est en annexe (Annexe C). Une deuxième fouille de texte plus pointue est effectuée en utilisant ce deuxième dictionnaire, générant un nouveau score appelé « Score 2 », variant de 0 à 8. Les sinistres ayant un Score 2 différent de zéro sont alors considérés comme risque cyber. Les autres

sinistres avec un Score 2 nul sont également classés comme risque cyber s'ils ont un Score 1 supérieur à 10.
Pour le reste des sinistres, des filtres sont appliqués sur les variables de la base suivantes :

- « Event risk category »
- « Sub risk category »
- « Activity »

Par exemple, pour détecter les incidents liés à la défaillance de système, on filtre en sélectionnant la catégorie d'événement « Business disruption and system failure » dans la variable « Event risk category » ainsi que la sous-catégorie « Systems » dans la variable « Sub risk category ». De manière similaire, les cas de piratage sont identifiés en choisissant les catégories « Internal Fraud » dans la variable « Event Risk Category » et « Hacking damage (if not physical damage) » dans la variable « Activity ».

Les deux fouilles de textes sont réalisées en utilisant du code Python. La dernière étape de validation consiste à effectuer des échantillonnages aléatoires des sinistres pour confirmer qu'ils correspondent effectivement à des sinistres cyber. Nous avons tiré aléatoirement plusieurs cas des sinistres identifiés comme risque cyber et lu les descriptifs correspondants. Tous les cas testés étaient bien des risques cyber.

Au terme de ce processus, 1 136 incidents cyber sont identifiés dans la base SAS OpRisk Global data.

6.2.3 Exemple

Pour illustrer la démarche employée pour identifier les risques cyber à partir de la base de données, examinons l'événement suivant en guise d'exemple.

- **Descriptif de l'incident tel qu'il apparaît dans la base de données :**

« In June 2023, Indigo Books & Music Inc, a Canadian bookseller, reported that it lost an estimated $23.95M (31.7M CAD) due to a cyberattack. On February 8, 2023, the company was hit by a ransomware attack that encrypted its data while the culprits demanded payment for decryption. The attack affected the company's online and in-store sales, causing a revenue loss of $20.02M (26.5M CAD) in the fourth quarter of its fiscal year. The attack also compromised some personal information of current and former employees, such as emails, home addresses, social insurance numbers and bank account details. The hacker group claimed to be affiliated with LockBit, a ransomware site, and threatened to release the stolen data publicly if Indigo did not pay by March 2, 2023. Indigo refused to pay the ransom as it could not guarantee the money would not end up in the hands of terrorists. The deadline passed without any evidence of the data being released, but Indigo said it was working with its cyber insurance provider and law enforcement to investigate the incident and protect its employees. »

- **Descriptif traité :**

« june indigo book music inc canadian bookseller report lose estimate cad due cyberattack february company hit ransomware attack encrypt data culprits demand payment decryption affect online instore sales cause revenue loss fourth quarter fiscal year also compromise personal information current former employees email home address social insurance number bank account detail hacker group claim affiliate lockbit site threaten release steal publicly pay march refuse ransom could guarantee money would end hand terrorists deadline pass without evidence say work cyber provider law enforcement investigate incident protect ».

- **Fouille de texte en utilisant le Dictionnaire 2 :**

Nous identifions les mots-clés suivants : cyberattack – ransomware – attack - personal information - hacker - ransom - cyber.

Ainsi, nous obtenons un Score 2 égal à 7 et donc ce n'est pas la peine d'effectuer la deuxième fouille de texte en utilisant le Dictionnaire 1. L'incident est identifié comme incident cyber.

6.3 Synthèse de la procédure d'identification des incidents cyber

Le schéma suivant synthétise les différentes étapes de la procédure d'identification des incidents cyber dans la base de données SAS OpRisk Global data.

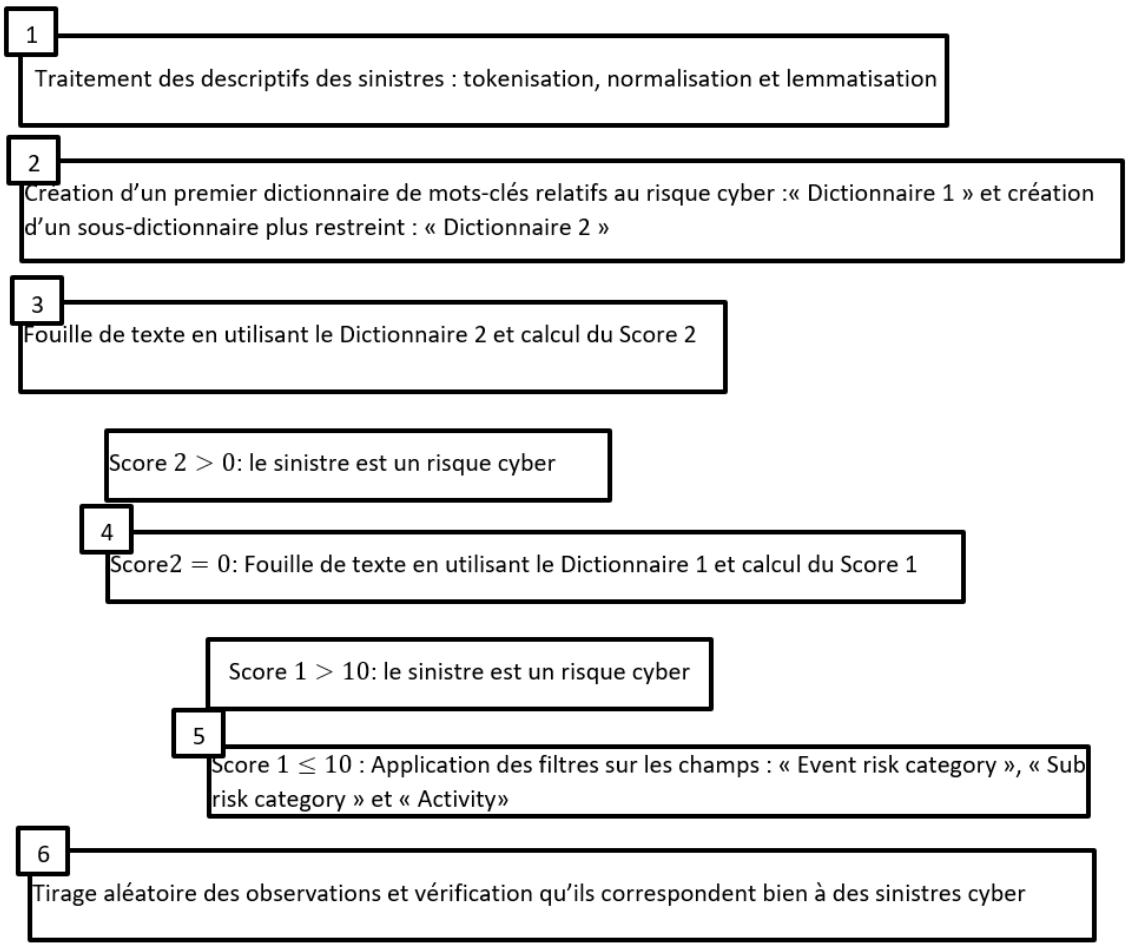

Figure 16 : Synthèse de la procédure d'identification des incidents cyber

6.4 Etude descriptive de la variable cible

6.4.1 Distribution de la variable cible

Maintenant que les incidents cyber sont identifiés, le coût indexé de ces incidents : « Current Value of Loss », peut être désormais analysée. Pour rappel, ces pertes sont exprimées en millions de dollars (M$) et ajustées en fonction du CPI[65].

Les sinistres cyber de la base engendrent une charge totale de 35,96 M$. Le résumé statistique des coûts des incidents cyber est présenté dans le tableau suivant. Les coûts varient de 0,11 à 3 505,35 M$. Le coût moyen est de 31,65 M$. Cette moyenne est plus élevée que la médiane (2,16 M$), illustrant ainsi l'asymétrie de la distribution des pertes.
La skewness positive indique une asymétrie positive dans la distribution des pertes. Cela signifie que la queue droite de la distribution est plus longue, les valeurs étant davantage regroupées du côté gauche.
Le kurtosis excédant 3 indique que les valeurs sont plus concentrées autour de la moyenne, avec une queue épaisse. De plus, le troisième quartile, qui est supérieur à 10 M$, est loin derrière le maximum observé (3505,352 M$).

Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum	Ecart-type	Skewness	Kurtosis
0,11	0,53	2,16	31,65	10,41	3 505,35	149,07	14,14	279,76

Tableau 15: Statistiques descriptives des coûts indexés des incidents cybers exprimés en M$

Le graphique 17 et 18 de la distribution des pertes indexées des sinistres cyber met en évidence une concentration des pertes sur des valeurs basses, tout en révélant la présence de quelques valeurs extrêmes. Cela confirme que la distribution est à queue épaisse.

[65] Aux États-Unis, l'indice des prix à la consommation (CPI) mesure les variations dans les prix payés par les consommateurs pour un panier de biens et de services.

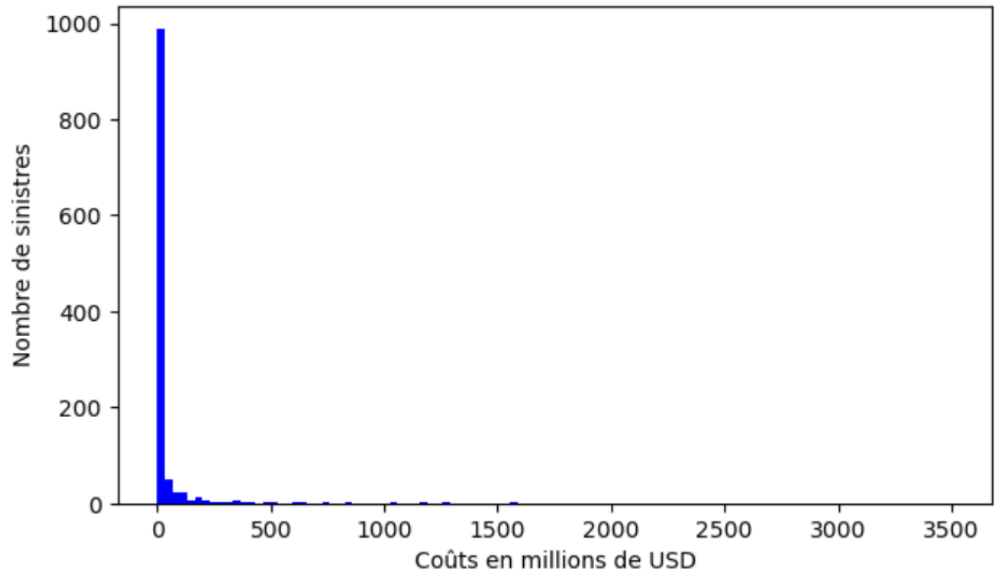

Figure 17 : Histogramme des pertes indexées d'incidents cyber en millions de dollars

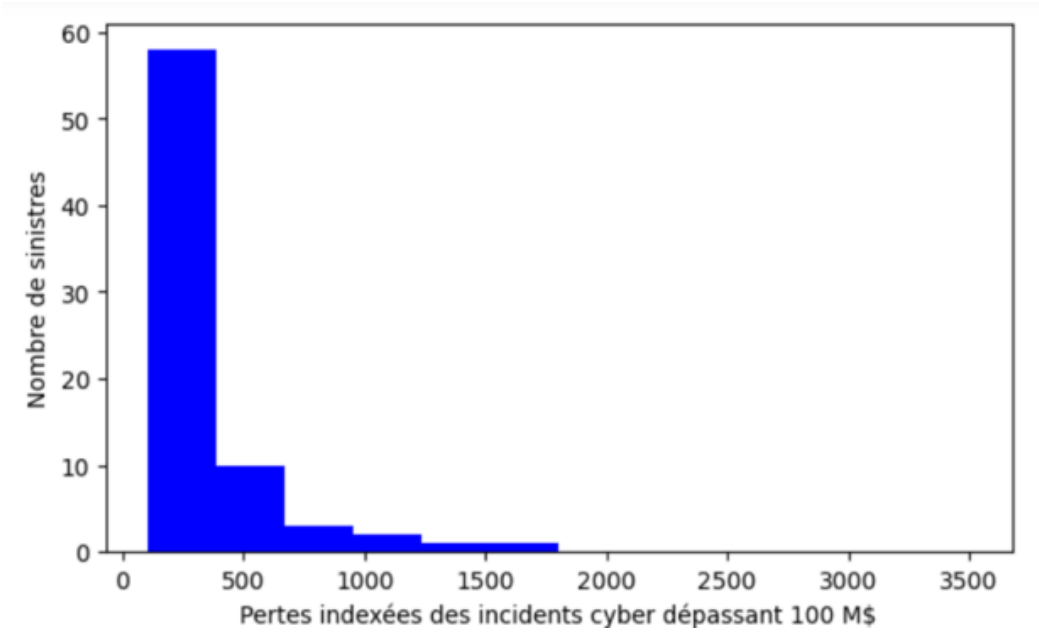

Figure 18: Histogramme des pertes indexées d'incidents cyber supérieurs à 100 millions de dollars

La boîte à moustaches, dans la figure ci-après, rend visible l'asymétrie de la distribution des pertes des incidents cyber et la présence des valeurs extrêmes.

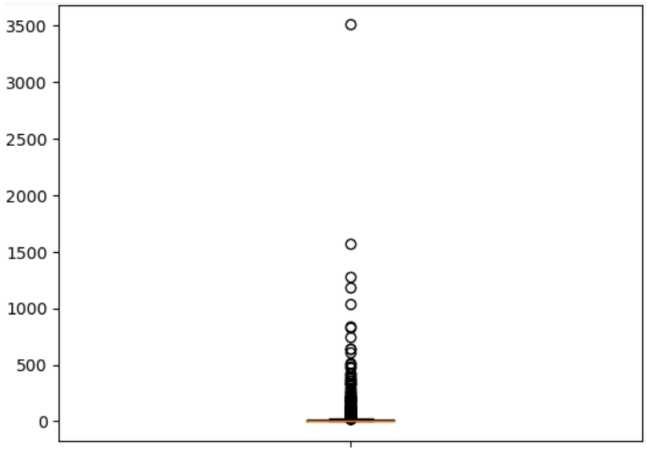

Figure 19 : Boîte à moustaches des pertes indexées d'incidents cyber en millions de dollars

Il apparaît plus nettement grâce à la fonction de répartition empirique que la distribution des pertes des sinistres cyber de la base de données présente des valeurs extrêmes.

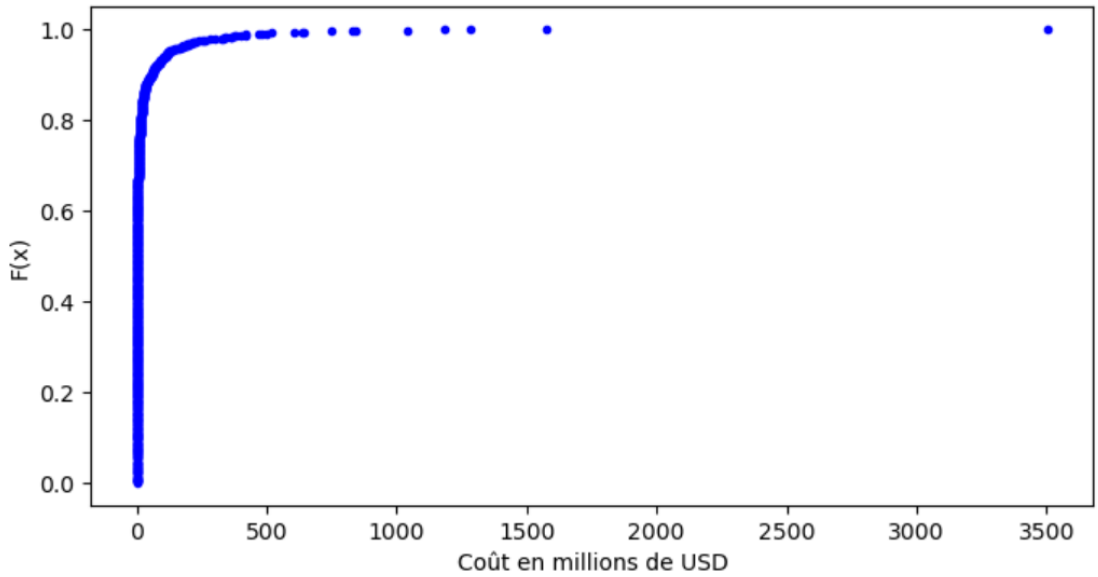

Figure 20 : Fonction de répartition empirique des pertes indexées des incidents cyber en millions de dollars

Si la queue de distribution est regardée de plus près, en examinant les incidents dont le coût dépasse 100 millions de dollars, soit 6,7 % des observations de la base, on compte 76 incidents pour un montant total de pertes atteignant 27 M$, soit 75,41 % de la perte totale. En considérant les incidents dont le coût dépasse 200 M$, soit 3,3 % des observations de la base, on recense 38 incidents pour une perte totale de 21 M$, représentant ainsi 60,67 % de la perte totale. Ainsi, une infirme proportion de 3,3% des observations englobe 60,67 % de la charge totale de la base.

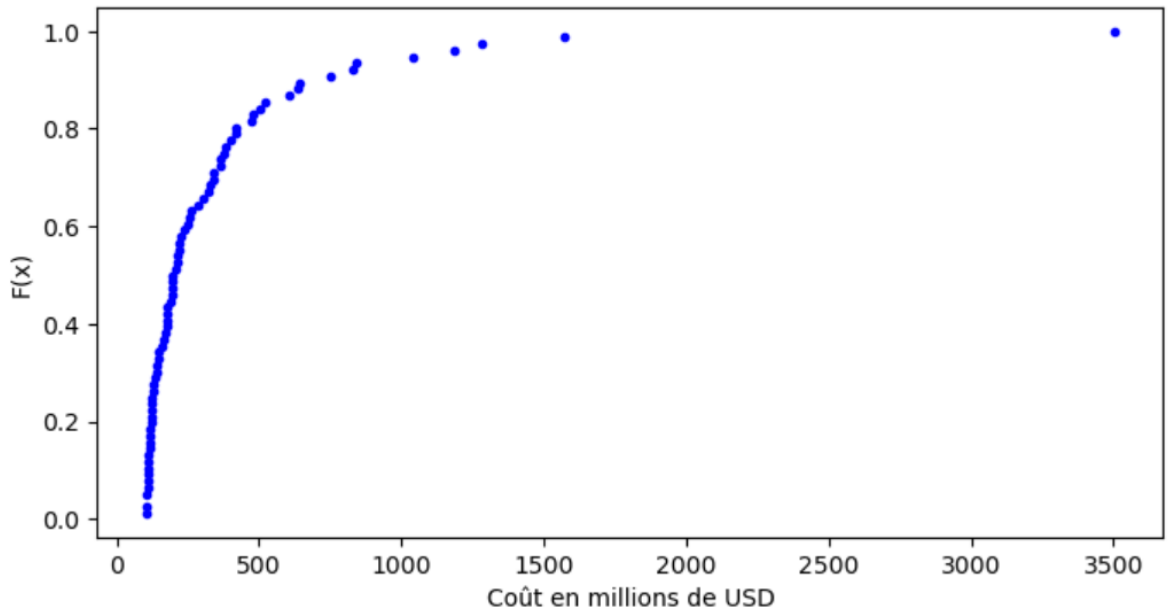

Figure 21 : Fonction de répartition empirique des pertes indexées des incidents cyber dépassant 100 millions de dollars

Pour modéliser les coûts des sinistres cyber, il est impératif de segmenter la base de données en distinguant les sinistres graves des sinistres attritionnels, procédant ainsi à un écrêtement. Les méthodes classiques de tarification ne sont pas adaptées aux sinistres extrêmes, d'où la nécessité de sélectionner un seuil pour définir les sinistres graves.

6.4.2 Analyse descriptive des sinistres cyber identifiés

Nous analysons la variable cible du coût des sinistres cyber en fonction des différentes variables disponibles dans la base de données. Le tableau suivant, donne le nombre des incidents cyber, le pourcentage du nombre total des incidents cyber ainsi que le coût moyen en millions de dollars en fonction des différentes classes de certaines variables.

Variable	Nombre	Pourcentage du nombre total des incidents cyber	Coût moyen (M$)
Répartition A : Secteur			
Financier	756	66,55%	21,59
Non Financier	380	33,45%	51,68
Répartition B: Region of Domicile			
North America	606	53,35%	32,84
Europe	277	24,38%	22,79
Asia	160	14,08%	48,93
Other	59	5,19%	46,17
Africa	34	2,99%	17,06
Répartition C: Event Risk Category			

External Fraud	578	50,88%	23,34
Internal Fraud	266	23,42%	29,85
Execution, Delivery & Process Management	149	13,12%	9,96
Business Disruption and System Failures	92	8,10%	69,9
Clients, Products & Business Practices	50	4,40%	131,89
Employment Practices and Workplace Safety	1	0,09%	20,36
Répartition D: Taille de l'entreprise[66]			
Grande Entreprise	751	66,11%	32,02
Entreprise à taille Intermédiaire	249	21,92%	27,79
Petite et moyenne Entreprise	114	10,04%	36,05
Très petite entreprise	22	1,94%	40,22

Tableau 16 : Analyse descriptive des sinistres cyber identifiés en fonction de quelques variables de la base de données

La Répartition A dans le tableau 16, donne la répartition des sinistres cyber de la base de données par secteur Financier et non Financier. Sans surprise, plus que de deux tiers des sinistres cyber surviennent dans le secteur financier. Cependant, le coût moyen des incidents cyber des entreprises du secteur non financier (51,68 M$) est à un peu plus que le double du coût moyen du secteur financier (21,59 M$). Ce constat est retrouvé en réalité, comme déjà discuté dans la section 3.5. Le secteur financier reste le plus touché par le risque cyber, en même temps, il est le mieux préparé pour en faire face par rapport aux autres secteurs.

La répartition géographique des incidents cyber (Répartition B) montre que plus que la moitié des incidents cyber de la base de données concernent des entreprises de l'Amérique du Nord, suivie par les entreprises européennes (24,38 %). Ce constat est conforme à ce qui a été vu dans la section 3.1. En effet, les Etats-Unis, faisant partie de l'Amérique du Nord, connaissent le plus grand taux de cybercriminalité au niveau mondial. De plus, SAS OpRisk global data étant américaine, la proportion des sinistres américains au sein de cette base est importante. Sans oublier la réglementation américaine, qui impose aux entreprises la déclaration de leurs sinistres cyber. A priori, il existe un biais géographique important dans la base de données utilisée. En termes de coûts, ce sont les entreprises asiatiques qui enregistrent le plus grand coût moyen des incidents cyber (48,93 M$) suivies par les entreprises de l'Amérique du Nord (32,48 M$), puis les entreprises européennes (22,79 M$). Les entreprises africaines quant à elles connaissent les plus petites proportions des sinistres cyber aussi bien en nombres qu'en montants.

La Répartition C répertorie les incidents cyber selon les catégories de classification des événements opérationnels définies par le comité de Bâle[67]. Plus de la moitié des incidents cyber de la base de données

[66] La création de la variable Taille de l'entreprise en fonction du nombre des salariés est détaillée dans la section 11.3.

[67] Détail dans la section 6.1.

font partie de la fraude externe (external fraud) (50,88%) suivie par la fraude interne (internal fraud) (23,42%), puis par l'exécution, livraison et gestion des processus **(**Execution, Delivery & Process Management) (13,12%). Ainsi, dans plus que la moitié des cas, le comportement humain est responsable de l'incident cyber. En termes de coût moyen, nous trouvons en tête du classement, la catégorie Clients, produits et pratiques commerciales (Clients, Products & Business Practices) avec un coût moyen de 131,89 M$ relatifs à 50 incidents. Le deuxième plus grand coût moyen est celui de la catégorie dysfonctionnement de l'activité et des systèmes (Business Disruption and System Failures) (69,90 M$), suivi par les catégories de fraude interne et fraude externe (respectivement 29,85 M$ et 23,34 M$).

Enfin, la Répartition D, donne la répartition des incidents cyber selon la taille des entreprises. Nous constatons que plus la taille de l'entreprise augmente, plus le nombre des incidents cyber augmente. En effet, 751 sinistres cyber concernent des grandes entreprises, contre 22 des très petites entreprises. Encore une fois, ce constat est conforme à ce qui est observé en réalité, les entreprises à grandes tailles représentent des visées à grand potentiel pour les cybercriminels. Cependant, en termes de coûts moyens, les très petites ainsi que les petites et moyennes entreprises enregistrent les plus grands coûts moyens (respectivement 40,22 M$ et 36,05 M$), ce qui peut être expliqué par le fait que ces entreprises ne sont pas bien préparées pour faire face aux risques cyber et donc ne peuvent pas limiter leurs dégâts, contrairement aux autres entreprises grandes et intermédiaires, qui généralement investissent beaucoup plus dans leurs systèmes de prévention et de protection du risque cyber. A noter, que le coût moyen des grandes entreprises (32,02 M$) est supérieur à celui des entreprises intermédiaires (27,79 M$). Ce constat peut être dû à l'amplitude des actifs et données des premières entreprises par rapport aux deuxièmes.

6.5 Objectifs de l'étude

En termes généraux, Deloitte vise à mieux s'approprier le risque cyber, un risque peu connu, en évolution continue, et dont le marché d'assurance est non mature. L'objectif de ce mémoire est d'évaluer la pertinence de l'utilisation de données publiques pour modéliser le risque cyber. Cette étude s'attache à construire un modèle composite du coût d'un incident cyber en se basant sur la base SAS OpRisk Global Data : un modèle prédictif des coûts d'incidents attritionnels utilisant des modèles GLM et des arbres de régression, ainsi qu'un modèle GPD pour les sinistres extrêmes. Cette approche vise à mieux prendre en compte la diversité des niveaux de gravité des coûts liés au risque cyber.

Une deuxième caractéristique distinctive de ce mémoire réside dans la modélisation de la probabilité qu'un incident cyber soit qualifié de grave. Les modèles développés au cours de cette étude ne contribueront pas seulement à approfondir la compréhension du risque cyber, mais seront également utiles aux assureurs cherchant à élaborer des produits d'assurance cyber.

Troisième partie

Contexte théorique de l'étude

Cette partie est axée sur la présentation du cadre théorique de cette étude. Nous débutons par exposer des éléments de la théorie des valeurs extrêmes puis nous abordons les modèles linéaires généralisés et enfin les arbres de régression et de classification CART. Pour rappel, et comme il le sera présenté dans la partie suivante, les valeurs extrêmes seront principalement utilisées par la suite, pour la modélisation de la sinistralité grave et ces derniers pour la modélisation attritionnelle.

Chapitre 7 : Eléments de la théorie des valeurs extrêmes

7.1 Présentation générale

Alors que la statistique « classique » a une appétence particulière à la partie centrale de la loi modélisant au mieux un phénomène donné, la Théorie des Valeurs Extrêmes (TVE) s'intéresse à l'étude du comportement asymptotique des queues de distribution de la loi. La difficulté principale réside dans l'application de cette théorie, puisque par nature, un évènement extrême est très peu observé. Nous parlons d'évènements rares. La TVE s'avère précieuse dans plusieurs domaines d'application, comme en finance, en météorologie et dans les sciences environnementales. En particulier, elle trouve également son utilité dans le domaine de l'assurance pour modéliser les risques extrêmes. Pour une présentation assez complète de la TVE, nous renvoyons à l'ouvrage de référence « An Introduction to Statistical Modeling of Extreme Values » par Stuart Coles (2001). Cet ouvrage présente les principaux résultats théoriques concernant la TVE et propose divers exemples pratiques pour une meilleure compréhension.

En TVE, deux approches distinctes sont utilisées : la première, connue sous le nom de Block maxima, se concentre sur l'étude de la distribution asymptotique du maximum, tandis que la seconde se penche à l'étude de la loi des excès. Dans cette deuxième approche, seules les données dépassant un seuil spécifique sont prises en considération. Cette approche, connue sous le nom de Peaks-Over-Threshold (POT), est celle qui nous aide à définir ce seuil. Le graphe ci-après, illustre les deux approches.

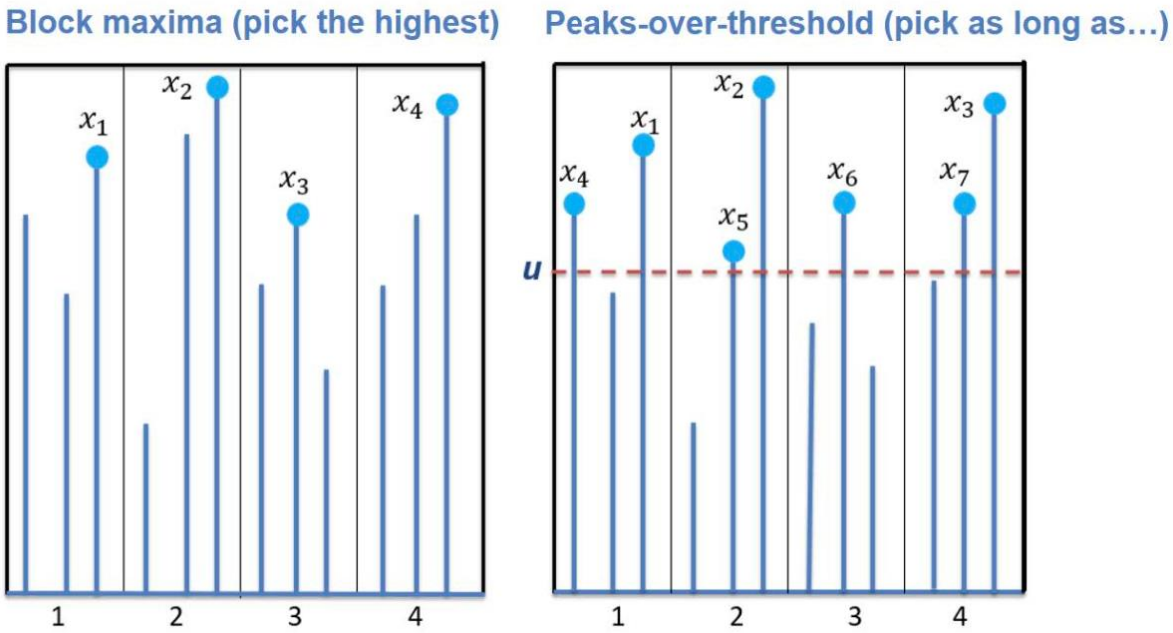

Figure 22: Illustration des deux approches Block maxima et POT (Source : « Extreme Value Theory ('EVT') & Extreme Value At Risk ('EVaR') » de LibertyRoad)

Nous aborderons les fondements de la théorie des valeurs extrêmes et explorerons plusieurs méthodes graphiques. Ces méthodes visent à estimer un seuil optimal permettant de définir une observation comme étant extrême, tout en offrant un aperçu du comportement de la queue de distribution de la variable aléatoire étudiée.

7.2 Méthode des dépassements des seuils (ou Peaks Over Threshold (POT))

7.2.1 Distributions des valeurs extrêmes

Les distributions de valeurs extrêmes sont une famille de distributions liées à la TVE. Considérons $X_1, \dots, X_n$, un échantillon de n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition commune F. Notre intérêt se porte sur la queue de distribution, ce qui nous conduit à étudier le comportement du maximum de cet échantillon, noté par $M_n = \max(X_1, \dots, X_n)$. La fonction de répartition de M_n est donnée par :

$$P(M_n \leq x) = \prod_{t=1}^{n} P(M_t \leq x) = F(x)^n$$

De façon analogue au théorème central limite, nous pouvons trouver deux suites de normalisation $a_n > 0$ et b_n et une loi non-dégénérée de loi G, telles que quand $n \to \infty$:

$$\frac{M_n - b_n}{a_n} \xrightarrow{loi} G$$

Le théorème de Fisher-Tippet fournit un résultat très intéressant : il n'existe que trois distributions qui sont considérées asymptotiques à la loi limite de M_n. Il s'agit des distributions de Fréchet, Gumbel et Weibull, dont les fonctions de répartitions s'expriment comme suit :

- **Distribution de Fréchet :** $\phi_\alpha(x) = exp(-x^{-\alpha}), x > 0\ pour\ \alpha > 0$
- **Distribution de Gumbel :** $\Lambda(x) = exp(-e^{-x}),\ -\infty < x < +\infty$
- **Distribution de Weibull :** $\Psi_\alpha(x) = exp(-|x|^{-\alpha}), x \leq 0\ pour\ \alpha < 0$

7.2.2 Loi des valeurs extrêmes généralisée

Bien que le comportement des trois lois de Fréchet, Gumbel et Weibull soit différent, elles peuvent être combinées en une seule loi, caractérisée par un paramètre ξ mesurant la lourdeur de la queue de distribution. Cette loi est appelée la loi des valeurs extrêmes généralisée (Generalized Extreme Value) notée GEV et est définie par la fonction de répartition suivante :

$$G_\xi(x) = \begin{cases} exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right), & \xi \neq 0, \quad 1+\xi x > 0 \\ exp(-exp(-x)), & \xi = 0, \quad -\infty \leq x \leq +\infty \end{cases}$$

En introduisant les paramètres de localisation μ et de dispersion σ, nous obtenons la forme la plus générale de la GEV :

$$G_{\xi,\mu,\sigma}(x) = \exp\left(-\left(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\right)\right),\ \xi \neq 0, \quad 1+\xi\frac{x-\mu}{\sigma} > 0$$

- Le cas de $\xi > 0$ correspond à la loi de Fréchet de paramètre $\alpha = \frac{1}{\xi}$
- Le cas de $\xi = 0$ correspond à la loi de Gumbel.
- Le cas de $\xi < 0$ correspond à la loi de Weibull de paramètre $\alpha = -\frac{1}{\xi}$.

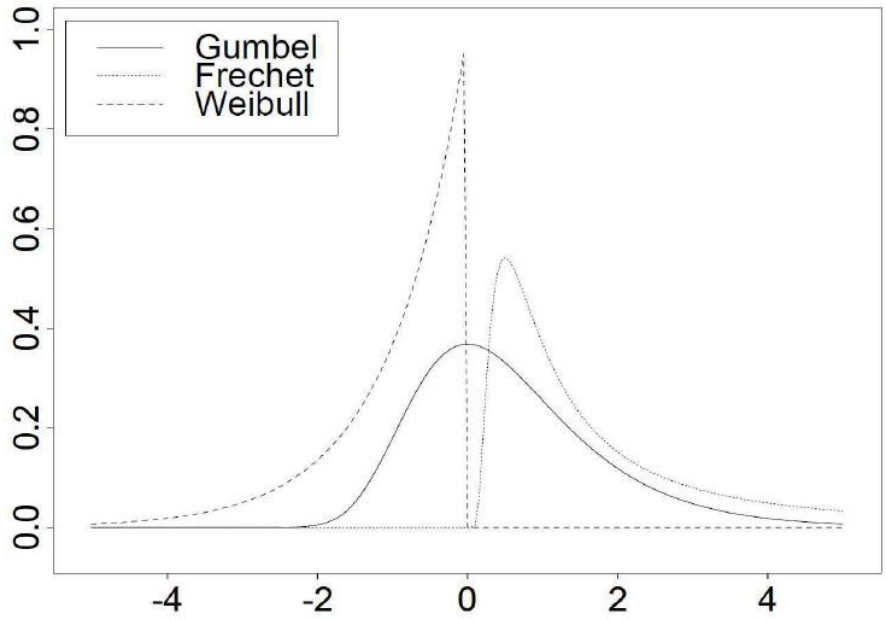

Figure 23 : Densités des distributions de Gumbel, Fréchet et Weibull (Source : cours ENSAE-IPP « Extreme Value Theory » par Christian Y. Robert)

Quand $\xi = 0$, la loi présente une décroissance de type exponentiel dans la queue de la loi, on dit que l'on est alors dans le domaine d'attraction de Gumbel (c'est le cas par exemple de la loi normale, exponentielle, gamma ou log-normale). Le domaine de Fréchet, correspondant à une loi non bornée et est le domaine où $\xi > 0$ (c'est le cas par exemple de la loi de Cauchy ou Pareto). Enfin quand $\xi < 0$, la loi est bornée et on dit que l'on est dans le domaine d'attraction de Weibull (c'est le cas par exemple de la loi uniforme ou béta).

7.2.3 Définition de la distribution conditionnelle des excès

Soit X une variable aléatoire de fonction de répartition F et u un seuil donné. La distribution conditionnelle des excès au-delà de u est définie par la fonction de répartition de $X - u / X > u$, notée F_u :

$$F_u(x) = \mathrm{P}(\mathrm{X} - \mathrm{u} < x / \mathrm{X} > \mathrm{u}) = \frac{F(u+x) - F(u)}{1 - F(u)}, \qquad 0 \leq x < \infty$$

L'espérance de la loi des excès est :

$$e(u) = \mathrm{E}(\mathrm{X} - \mathrm{u} < x / \mathrm{X} > \mathrm{u})$$

Cette distribution conditionnelle convient à la modélisation d'une sinistralité au-delà d'un certain seuil qui correspond à l'objectif de modélisation de la sinistralité grave des sinistres cyber. L'objectif de la méthode POT est de déterminer par quelle loi de probabilité cette distribution conditionnelle peut être approchée.

7.2.4 Théorème de Pickands (1975) - Loi de Pareto Généralisée

La fonction de répartition F de point terminal x_F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull) $G_{\xi,\beta}$ de paramètre de forme $\xi > 0$, si et seulement s'il existe une fonction positive $\beta(u)$ telle que :

$$\sup | F_u(x) - G_{\xi,\beta(u)}(x)| \to 0 \; pour \; 0 \leq x \leq x_F - u \; et \; u \to \infty$$

Où $G_{\xi,\beta(u)}$ est la fonction de répartition de la loi de Pareto Généralisée et F_u est la fonction de répartition des excès au-delà du seuil u.

Ainsi, pour u suffisamment grand, la loi des excès est approchée par une loi Pareto généralisée :

$$F_u(x) \approx G_{\xi,\beta(u)}$$

La loi de Pareto Généralisée, notée GPD (Generalized Pareto Distribution) de paramètres de forme $\xi \in \mathbb{R}$ et d'échelle $\beta > 0$, se définit par sa fonction de répartition :

- $Cas \; \xi \neq 0 : G(x) = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$;
- $Cas \; \xi = 0 : G(x) = 1 - e^{-\frac{x}{\beta}}$;

Où : $y \geq 0 \; pour \; \xi \geq 0 \; et \; 0 \leq y \leq -\frac{\beta}{\xi} \; pour \; \xi < 0$

7.3 Choix du seuil d'écrêtement

La principale difficulté de la méthode POT (Peaks Over Threshold) est de définir un seuil approprié. S'il est trop élevé, il y aura peu de valeurs pour modéliser correctement la queue de la distribution car la variance est susceptible d'être grande en raison de la présence uniquement d'observations très extrêmes. En revanche, un seuil bas inclura trop de valeurs, entraînant un biais élevé. Il est donc important de trouver un compromis adapté entre la variance et le biais.

Il existe plusieurs façons de choisir un seuil, et chacune présente à la fois des aspects positifs et négatifs. Ci-après, seront discutées quelques méthodes courantes.

7.3.1 Approche empirique (Rule of Thumb)

Une méthode pour déterminer le seuil consiste à utiliser une règle empirique pour choisir les k plus grandes observations et de procéder à la modélisation. Le $90^{ème}$ percentile est couramment utilisé à cette fin, mais d'autres seuils sont également proposés, tels que :

$k = \sqrt{n}$ et $k = n^{(2/3)} \Big/ \log(\log(n))$, n étant la taille des observations.

Ces approches sont souvent utilisées en pratique, bien qu'elles puissent manquer de justification théorique complète. Pour une compagnie d'assurance par exemple, les informations pertinentes pourraient se situer dans la distribution des sinistres dépassant une certaine valeur ou dans le niveau de certains quantiles de sinistres. Cette méthode est la plus rapide pour établir un seuil, mais en raison des inévitables disparités entre la plupart des données, elle ne peut pas être considérée comme une méthode fiable pour définir un seuil optimal.

7.3.2 Approche graphique

Une autre façon de déterminer un seuil consiste à utiliser des outils graphiques. Dans cette approche, le niveau de seuil défini dépend davantage des données elles-mêmes. Cependant, ces méthodes graphiques prennent beaucoup plus de temps que l'utilisation d'une règle prédéterminée et lorsque nous travaillons avec plusieurs ensembles de données, il peut être plus efficace en termes de précision, d'utiliser une règle empirique. Ci-dessous, nous présentons la théorie sous-jacente à certains de ces outils graphiques.

- **La fonction des excès moyens (Mean Residual Life Plot)**

Un outil graphique qui peut s'avérer utile pour définir le seuil u est le Mean Residual Life Plot (MRL) introduit par Davison et Smith (1990). Le MRL est le graphe des points $\big(u, e(u)\big)$, où $e(u)$ est la moyenne des excès au-delà du seuil u, définie par :

$$e(u) = E(X - u < x \mid X > u)$$

Si X suit une loi $GPD_{\xi,\beta}$, avec ξ le paramètre de forme et β le paramètre d'échelle, alors pour $\xi < 1$:

$$E(X - u \mid X > u) = \frac{\xi}{1-\xi}u + \frac{\beta}{1-\xi}$$

Nous pouvons démontrer que pour un seuil $v > u$, l'espérance de la moyenne des excès est linéaire en v :

$$E(X - v \mid X > v) = \frac{\beta + \xi v}{1 - \xi} = \frac{\beta_u + \xi(v - u)}{1 - \xi}$$

Avec $\beta_u = \xi u + \beta$.

Les excès moyens peuvent ainsi être représentés graphiquement et lorsque le MRL commence à montrer un comportement linéaire, un seuil approprié peut être sélectionné. Cependant, le MRL risque de perdre sa linéarité lorsque le seuil devient très élevé, en raison de la variance des quelques valeurs extrêmes restantes. Il existe des cas où le MRL est complètement ou jamais linéaire, et par conséquent, aucune conclusion ne peut être tirée de son observation.

- **Le graphe de stabilité des paramètres d'échelle et forme (Parameter stability plot)**

Cette méthode permet de déterminer un seuil optimal en ajustant les données à une distribution GPD et en utilisant différents seuils. Deux graphes de stabilité des paramètres peuvent ensuite être réalisés en traçant les valeurs estimées des paramètres ($\xi \; et \; \beta$) en fonction des différents seuils u, avec des intervalles de confiance à 95 % pour ces estimateurs. Au-delà d'un seuil v et sous l'hypothèse d'un paramètre de forme constant, le paramètre d'échelle modifié est estimé par $\hat{\beta}_u - \hat{\xi} u$. La stabilité des paramètres (forme et échelle) peut alors être observée et localisée.
Il est important de noter que selon les données, il peut arriver que les graphiques de stabilité des paramètres d'échelle et forme ne fournissent aucune information pertinente, tout comme c'est parfois le cas avec le MRL.

7.3.3 Une approche de compromis

Dans cette approche de compromis, nous combinons les méthodes graphiques et une estimation des paramètres de forme et d'échelle par diverses techniques. De plus, nous évaluons les erreurs standards associées aux paramètres estimés pour les seuils potentiels identifiés. Ces erreurs standards sur les paramètres estimés pour différents seuils, servent à orienter le choix du seuil. Le seuil retenu est celui affichant les plus petites erreurs standards.

7.4 Estimation des paramètres du modèle GPD

Une fois qu'un seuil est déterminé, les paramètres de forme et d'échelle associés à la distribution asymptotique des valeurs extrêmes peuvent être estimés. La méthode la plus couramment utilisée est la méthode du maximum de vraisemblance (Maximum Likelihood - ML), qui sera détaillée ci-dessous. Nous aborderons la méthode des moments pondérés par la probabilité (Probability Weighted Moments - PWM), une alternative qui peut être plus appropriée dans certains cas.

7.4.1 Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance repose sur la recherche et la maximisation d'une fonction de vraisemblance afin de trouver un estimateur approprié.

Considérons un échantillon $Y_1, \dots, Y_{k_n}$ indépendant et identiquement distribué selon une loi GPD de paramètres $\xi \; et \; \beta$, ce qui nous donne :

$\log\big(L(\xi,\beta_u)\big) = -k_n \log(\beta_u) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k_n}\log\left(1+\frac{\xi}{\beta_u}\,Y_i\right)$, lorsque $1+\frac{\xi}{\beta_u}Y_i > 0\ pour\ i = 1,\dots,k_n$ sinon $(L_{MV}(\xi,\beta_u)) = -\infty$.

La log-vraisemblance exprimée par la relation mentionnée ci-dessus, est maximisée à l'aide de méthodes numériques telles que l'algorithme de Newton-Raphson. Les estimateurs du maximum de vraisemblance sont asymptotiquement gaussiens et efficaces lorsque $\xi > -\frac{1}{2}$ (Smith (1987)).

7.4.2 Méthode des moments pondérés par la probabilité

L'idée fondamentale derrière la méthode des moments pondérés (Probability Weighted Moments - PWM) est d'égaliser les moments théoriques et empiriques de l'échantillon dans le but de trouver une méthode alternative à la méthode du maximum de vraisemblance (ML), de manière similaire à la méthode des moments.

Cette approche consiste à utiliser les deux moments pondérés, notés M_0 et M_1, où $M_s = E\left[X\big(1-G_{\xi,\beta}(Y)\big)^s\right]$. Nous avons :

$$M_s = \frac{\beta_u}{(s+1)(s-\xi+1)}, \text{si } s > \xi - 1$$

Un estimateur de ce moment est donné par :

$$\widehat{M_s} = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1-\frac{i}{k_n+1}\right)^s Y_{i,k_n},$$

où $Y_{1,k_n},\dots,Y_{k_n,k_n}$ sont les statistiques d'ordre associées à $Y_1,\dots,Y_{k_n}$.

Cette équation conduit aux estimateurs suivants pour $s=0$ et $s=1$:

$$\begin{cases}\hat{\xi} = 2 - \dfrac{\widehat{M_0}}{\widehat{M_0}-2\widehat{M_1}} \\ \widehat{\beta_u} = 2\dfrac{\widehat{M_0}\widehat{M_1}}{\widehat{M_0}-2\widehat{M_1}}\end{cases}$$

Les estimateurs des moments pondérés sont asymptotiquement gaussiens pour $-1 < \xi < \frac{1}{2}$ (Hosking et al. (2003)).

A partir du seuil retenu, un écrêtement est réalisé sur la base des coûts des sinistres. Tous les montants supérieurs à ce seuil sont considérés graves et sont distingués des sinistres ordinaires appelés attritionnels. La distribution des coûts des sinistres extrêmes est modélisée en utilisant la distribution de Pareto Généralisée - GPD. Quant aux coûts des sinistres attritionnels, ils sont modélisés en utilisant les Modèles Linéaires Généralisés – GLM, couramment utilisées en assurance non-vie, ainsi que les arbres de régression CART. Ces méthodes seront explorées plus en détail dans la section suivante.

Chapitre 8 : Modèles linéaires généralisés

8.1 Introduction

Les modèles linéaires généralisés (Generalized Linear Model - GLM) constituent des méthodes paramétriques qui permettent d'étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs $X_1, \dots, X_k$. Ils ont été introduits en 1972 par les statisticiens britanniques John Nelder et Robert Wedderburn. L'un des avantages des GLM par rapport aux modèles de régression linéaire classiques réside dans le fait que l'hypothèse selon laquelle la variable cible ne peut suivre que la loi Normale est désormais levée. Cette généralisation permet d'envisager des variables aléatoires provenant d'un panel de lois appelé la famille exponentielle. Un ouvrage qui offre une compréhension approfondie des modèles linéaires généralisés et de leurs applications est « Generalized Linear Models » écrit par McCullagh et Nelder. Nous nous basons aussi dans cette partie sur le cours de l'ENSAE « Assurance dommage » de Nicolas Baradel.

Un modèle GLM comporte trois composantes principales :

- **La composante aléatoire :** elle suppose que les observations Y_i sont indépendantes et proviennent d'une famille exponentielle.
- **La composante systématique :** elle définit $\eta = X\,\beta$ comme le prédicteur linaire où $\eta_i = \sum_{i=1}^{p} X_{ij}\ \beta_j$ pour $j = 1, \dots, n$. Les colonnes de X sont les variables explicatives ou leurs modalités si nous travaillons avec des variables qualitatives.
- **La fonction de lien :** celle-ci établit la relation fonctionnelle entre la combinaison linéaire des variables $X_1, \dots, X_p$ et l'espérance mathématique de la variable réponse Y, $E(Y) = \mu = g^{-1}(\eta)$, où g est une fonction différentiable et monotone, appelée une « fonction de lien ».

8.2 Famille exponentielle

Le modèle GLM repose sur une distribution de probabilité issue d'une famille exponentielle définie comme suit :

$$f(y_i, \theta_i, \varphi, \omega_i) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right)$$

Où :

- θ est un paramètre réel, appelé paramètre canonique.
- φ est un paramètre de dispersion.
- Les fonctions a, b et c sont spécifiées en fonction du type de la loi exponentielle.
- a est une fonction non nulle définie sur $\mathbb{R}$.
- $b(\theta_i)$ est la fonction de cumulant, supposée être deux fois différentiable, avec une dérivée seconde inversible.
- c est une fonction définie sur $\mathbb{R}^2$, deux fois dérivable.

Ainsi, pour une variable aléatoire Y appartenant à la famille exponentielle, on obtient les résultats suivants :

$$E(Y) = b'(\theta) \text{ et } Var(Y) = a(\varphi)b"(\theta)$$

Où $b'\ et\ b"$ représentent respectivement les première et seconde dérivées de la fonction b.

8.3 Fonction de lien

La fonction de lien g permet de traduire la relation fonctionnelle entre la composante aléatoire et la composante déterministe. Elle satisfait l'équation $E(Y) = \mu = g^{-1}(\eta)$ avec $\eta_i = \sum_{i=1}^{p} x_{ij}\ \beta_j$.

Dans le modèle GLM, nous avons une relation arbitraire $g(\mu_i) = \eta_i$. Ainsi, la fonction g doit être monotone et différentiable pour respecter cette relation.

Pour la modélisation des charges des sinistres dans l'assurance non-vie en utilisant les GLM, la loi Gamma est souvent utilisée. Dans le cadre de cette étude, nous allons également examiner la loi de Weibull. Ces deux lois que nous avons choisies pour la modélisation de la charge des sinistres ont un support positif.

8.4 Estimation des paramètres par maximum de vraisemblance

En considérant $y = (y_1, \ldots, y_n)$ comme étant une réalisation de l'échantillon de n variables aléatoires indépendantes $Y_1, \ldots, Y_n$, dont les fonctions de densité f_{Y_i} sont issues de la famille exponentielle et pour chaque i, y_i la réponse pour $x_i = (x_1, \ldots, x_n)$, la vraisemblance de y se formule comme suit :

$$\mathcal{L}(y; \beta, \varphi) = \prod_{i=1}^{n} f(y_i;\ \mu_i, \varphi) = \prod_{i=1}^{n} f(y_i;\ x_i\beta, \varphi)$$

Avec $\mu_i = g\big(E(Y_i)\big) = x_i\beta$ et g est la fonction canonique. Par conséquent, la log-vraisemblance s'exprime ainsi :

$$l(y; \beta, \varphi) = \sum_{i=1}^{n} \left[\frac{y_i x_i \beta - b(x_i\beta)}{a_i(\varphi)} + c(y_i, \varphi)\right]$$

La valeur maximale de $l(y; \beta, \varphi)$ est obtenue en résolvant l'équation aux dérivées partielles suivante :

$$\begin{cases} \dfrac{\partial l(y; \beta, \varphi)}{\partial \beta_j} = 0\ pour\ j = 1, \ldots, p \\ \dfrac{\partial l(y; \beta, \varphi)}{\partial \varphi} = 0 \end{cases}$$

La résolution de ce système ne possède pas de solution explicite. Toutefois, il existe plusieurs algorithmes d'optimisation itératifs pour obtenir les estimations du maximum de vraisemblance. Les deux algorithmes les plus couramment utilisés sont l'algorithme de Newton-Raphson[68] et l'algorithme du Fisher-scoring[69].

8.5 Qualité de l'ajustement et comparaison des modèles

- **Déviance :**

La déviance est une statistique utilisée pour évaluer la qualité d'ajustement d'un modèle linéaire généralisé. Elle compare le modèle estimé au modèle saturé, c'est-à-dire le modèle qui

[68] Lien pour plus d'informations sur l'algorithme : https://fr.wikipedia.org/wiki/M%C3%A9thode_de_Newton
[69] Lien pour plus d'informations sur l'algorithme : https://en.wikipedia.org/wiki/Scoring_algorithm

comporte autant de paramètres à estimer que d'observations, représentant ainsi exactement les données. Formellement, la déviance est définie comme la différence entre la log-vraisemblance du modèle ajusté et celle du modèle saturé. En notant M le modèle ajusté et S le modèle saturé, la déviance du modèle M se calcule comme suit :

$$D(M) = -2\varphi\left(\mathcal{L}(y;\hat{\beta}_M) - \mathcal{L}(y;\hat{\beta}_S)\right) = -2\,\varphi\left(ln\left(\frac{\mathcal{L}_M}{\mathcal{L}_S}\right)\right)$$

Où $\hat{\beta}_M$ est l'estimateur du maximum de vraisemblance de β dans le modèle M et $\hat{\beta}_S$ désigne l'estimateur du maximum de vraisemblance de β dans le modèle saturé.

Une faible valeur de la déviance indique un ajustement de bonne qualité. En effet, plus la déviance est faible, plus les log-vraisemblances des deux modèles sont proches. L'objectif est donc de minimiser la valeur de la déviance. Lorsque le modèle étudié est exact, la statistique D suit approximativement une loi de χ^2 à $n-k$ degrés de liberté, où n est le nombre d'observations et k est le nombre de paramètres estimés dans le modèle.

- **Critère d'information (AIC)**

Le Critère d'Information d'Akaike (Akaike Information Criterion – AIC) est une mesure de la qualité d'un modèle statistique introduite par Hirotugu Akaike en 1973. L'AIC utilise le maximum de vraisemblance, mais en pénalisant les modèles comportant beaucoup de variables, ce qui peut conduire à un « sur-apprentissage » et à une généralisation moins bonne.
En notant k le nombre de paramètres du modèle et $\mathcal{L}$ sa log-vraisemblance maximale, le critère AIC est défini de la manière suivante :

$$AIC = -2\ln(\mathcal{L}) + 2k$$

Le meilleur modèle est celui qui a le plus petit critère AIC. En effet, un AIC plus bas indique un équilibre entre la précision du modèle mesurée par la log-vraisemblance maximale, et sa complexité mesurée par le nombre de paramètres. Cela favorise les modèles qui fournissent un ajustement adéquat tout en évitant un excès de complexité.

- **Critère d'information (BIC)**

Le Critère d'Information Bayésien, Bayesian Information Criterion - BIC, est un critère d'information dérivé du critère d'information d'Akaike proposé par Gideon Schwartz en 1978. Contrairement au critère d'information d'Akaike, la pénalité dans le BIC dépend non seulement du nombre de paramètres, mais aussi de la taille des données. Le critère est défini comme suit :

$$BIC = -2\ln(\mathcal{L}) + \ln(n)k$$

Où :
- $\mathcal{L}$ est la log-vraisemblance maximale du modèle.
- n est la taille des données.
- k est le nombre de paramètres du modèle.

Dans le BIC, la pénalisation pour la complexité du modèle augmente avec la taille des données. Le meilleur modèle est celui qui possède le plus petit BIC, car cela indique un équilibre entre l'ajustement du modèle et sa complexité, tout en tenant compte de la taille de l'échantillon.

- **L'erreur quadratique moyenne :**

L'erreur quadratique moyenne, Root Mean Squared Error- RMSE, est la racine carrée de la moyenne des carrées des erreurs ou résidus d'un modèle. Les résidus représentent la différence entre les valeurs observées et les valeurs prédites par le modèle. La RMSE est une mesure qui évalue la dispersion de ces résidus. En d'autres termes, il permet de quantifier la proximité des données par rapport à la ligne de régression ou d'ajustement du modèle. Plus la RMSE est faible, plus les prédictions du modèle sont proches des valeurs réelles, indiquant ainsi une meilleure précision du modèle.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Où :
- $\hat{y}_i$ sont les valeurs prédites par le modèle
- y_i sont les valeurs réellement observées
- n est la taille des données.

La RMSE s'exprime dans la même unité que la variable à prédire et est par conséquent facile à interpréter.

8.6 Sélection des variables

Dans un modèle GLM, un des problèmes essentiels est le choix des variables explicatives, surtout lorsqu'il y a un grand nombre de variables dans la base de données. Certaines d'entre elles ne sont pas forcément significatives pour expliquer la variable cible.
Afin de remédier à cela, différentes techniques peuvent être utilisées, notamment la régression pas à pas et l'analyse exploratoire des données. Il est tout aussi possible de faire appel à l'expertise du marché, pour effectuer une première sélection des variables qui pourra être validée par des méthodes statistiques.

8.6.1 Régression pas à pas

L'objectif de la régression pas à pas (Stepwise regression) est de choisir un sous-ensemble utile de prédicteurs. A chaque étape, le programme ajoute la variable la plus significative ou supprime la variable la moins significative, selon un critère prédéfini tel que le critère AIC, ou le critère BIC. Une régression Stepwise propose en général trois méthodes :

- **Régression en avant (Forward regression) :** cette méthode implique de commencer sans variables explicatives dans le modèle, de tester l'ajout de chaque variable en utilisant un critère de qualité de modèle choisi, d'ajouter la variable dont l'inclusion apporte l'amélioration la plus statistiquement significative du modèle, et de répéter ce processus jusqu'à ce qu'aucune variable n'améliore le modèle de manière statistiquement significative.

- **Régression en arrière (Backward regression) :** cette méthode démarre avec un modèle contenant toutes les variables initiales. À chaque étape, elle évalue la suppression de chaque

variable, en se basant sur un critère de qualité du modèle prédéfini. La variable supprimée est celle qui, si retirée, induit la détérioration statistiquement la moins significative de l'ajustement du modèle, selon le critère défini. Ce processus est répété de manière itérative, jusqu'à ce qu'aucune autre variable ne puisse être supprimée sans entraîner de détérioration statistiquement significative de l'ajustement du modèle selon le critère spécifié.

- **Régression bidirectionnelle (Bidirectional regression) :** combine à la fois la méthode forward et backward. Elle débute avec un modèle initial comprenant toutes les variables ou un sous-ensemble choisi et procède à des ajouts ou suppressions de variables en fonction de leur impact statistique significatif sur l'ajustement du modèle.

8.6.2 Tests statistiques

- **V de Cramer : mesurer l'intensité des associations entre les variables explicatives**

Pour le cas des variables explicatives qualitatives, nous nous intéressons à l'association entre ces variables. Une mesure souvent utilisée pour évaluer cette association est le coefficient V de Cramer. C'est une mesure de la force d'association entre deux variables nominales et prend des valeurs comprises entre 0 et 1. Il est calculé en se basant sur le test du chi-deux (χ^2) entre les variables catégorielles. Le test du chi-deux (χ^2) est un test statistique permettant de déterminer s'il existe une relation entre deux variables catégorielles. Pour deux variables catégorielles X et Y, avec un nombre de modalités respectif de k et l, et pour n observations, nous considérons la statistique de Pearson définie par :

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{i,j} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i\, n_j}{n}}$$

Où :

- $n_{i,j}$ est le nombre d'observations pour lesquelles X prend la modalité i et Y la modalité j
- n_i est le nombre d'observations pour lesquelles X prend la modalité i
- n_j est le nombre d'observations pour lesquelles Y prend la modalité j

Le coefficient du V de Cramer est calculé à l'aide de la formule suivante :

$$V = \sqrt{\frac{\chi^2/_n}{min(k-1, l-1)}}$$

Plus le V de Cramer s'approche de 1, plus l'intensité de la relation entre les deux variables étudiées est forte. Il s'interprète comme suit :

V de Cramer	Interprétation
$[0 - 0{,}2]$	Les champs ont un lien d'association faible
$]0{,}2 - 0{,}6]$	Les champs ont un lien d'association modéré
$]0{,}6 - 1]$	Les champs ont un lien d'association fort

Tableau 17: Règle empirique d'interprétation du coefficient du V de Cramer

- **Test de Kruskal-Wallis : Détection des corrélations entre la variable dépendante et les variables explicatives**

Le test de Kruskal-Wallis (1952) est un test statistique non paramétrique visant à évaluer l'influence d'une variable explicative discrète sur une variable continue. Il sert à déterminer si la distribution de la variable continue varie significativement selon les différentes modalités de la variable explicative catégorielle. Ce test est souvent utilisé comme alternative à l'ANOVA (Analyse de la variance) lorsque l'hypothèse de normalité n'est pas respectée. Il cherche à évaluer si les rangs médians des sous-échantillons $(Y|_{x=1}, \dots, Y|_{x=k})$, d'une variable Y continue sur $k \geq 2$ groupes d'une variable X sont différents. Ce test propose une hypothèse nulle (H_0) selon laquelle les k groupes sont confondus et ont la même distribution aléatoire :

(H_0): Les groupes indépendants $i \in [1, k]$ ont tous la même tendance centrale et proviennent donc de la même population.

La statistique du test de Kruskal-Wallis mesure l'écart entre la moyenne des rangs dans chaque groupe et la moyenne de tous les rangs, qui est égale à $\frac{n+1}{2}$. Elle est calculée comme suit :

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(n+1)$$

Où :

- n_i représente l'effectif de la population $i \in [1, k]$ et $n = \sum_{i=1}^{k} n_i$ est la somme totale des effectifs
- X_{ij} désigne la $j^{ème}$ observation de l'échantillon i, avec $j \in [1, n_i]$
- R_{ij} indique le rang de l'observation X_{ij} parmi les n observations
- $R_i = \sum_{j=1}^{n_i} R_{ij}$ représente la somme des rangs associée à chaque groupe i.

Sous l'hypothèse nulle (H_0), la statistique du test suit asymptotiquement une loi du chi-deux à $(k-1)$ degrés de liberté. La p-value associée au test de Kruskal-Wallis est égale à $P(X \leq {\chi^2}_{k-1})$. Si la p-value est telle que l'on doit rejeter l'hypothèse H_0, alors au moins un groupe est différent d'un autre et donc mène à la conclusion que la variable catégorielle a un effet sur les rangs moyens des différents groupes.

8.6.3 Expérience et expertise métier

Tirer des conclusions à partir des seuls résultats statistiques n'est pas toujours suffisant. Sur les résultats de ses tests statistiques, un avis d'expert supplémentaire peut s'appliquer pour déterminer si la variable a une vraie pertinence vis-à-vis de la variable explicative.

8.7 Modélisation du coût des sinistres

La modélisation de la sévérité des sinistres peut être effectuée en utilisant plusieurs modèles qui diffèrent dans leurs hypothèses et sur la répartition des sinistres. Il convient de noter qu'une distribution de sévérité de sinistre est généralement composée de valeurs continues et positives. L'enjeu de cette section est d'étudier la modélisation du coût d'un sinistre cyber, en fonction des caractéristiques propres à une entreprise et du type du sinistre. Nous nous intéressons aux charges des sinistres attritionnels et

notons Y_i pour $i \in \{1, \dots, n\}$ le montant du $i^{ième}$ sinistre de la base de données. Les deux modèles étudiés sont le modèle de régression de Gamma puis le modèle de régression de Weibull.

- **Modèle de régression Gamma**

La loi Gamma est une des lois les plus couramment utilisées pour la modélisation des charges des sinistres en assurance-non-vie. Si Y est une variable aléatoire suivant une loi Gamma de paramètre r et α tous deux strictement positifs. La densité s'écrit :

$$f_{r,\alpha} = \frac{\alpha^r}{\Gamma(r)} y^{r-1} \exp(-\alpha y) \quad y \geq 0$$

Avec $\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} du$

Et

$$E(y_i) = \mu_i = \frac{r}{\alpha_i}$$
$$Var(y_i) = \frac{r}{{\alpha_i}^2} = \frac{{\mu_i}^2}{r}$$

En posant $\mu = r/\alpha$, la fonction de densité peut également être mise sous la forme :

$$f_{r,\alpha} = \exp(-rln(\alpha) - \alpha y + (r-1)\ln(y) - \ln \Gamma(r))$$

Ainsi, la loi Gamma appartient à la famille exponentielle avec :

- $\theta = -\alpha$
- $\omega = -1$
- $a(\varphi) = 1$
- $b(\theta) = rln(-\theta)$
- $c(y;\ \varphi) = (r-1)\ln(y) - \ln \Gamma(r)$

Pour des variables d'intérêt positives, un lien canonique qui est très souvent utilisé est le lien logarithmique de sorte que $\mu_i = \exp(\eta_i)$. En d'autres termes, $ln(\mu_i) = {x_i}^t \beta$. La vraisemblance associée à ce modèle peut être écrite comme suit :

$$\mathcal{L}(y;\beta) = \prod_{i=1}^{n} \frac{1}{\Gamma(r)} \left(\frac{r y_i}{\mu_i}\right)^r exp(-\frac{r y_i}{\mu_i}) \frac{1}{y_i}$$

En passant au logarithme, la maximisation de la vraisemblance revient à maximiser la quantité suivante en fonction de β :

$$l(y;\beta) = \sum_{i=1}^{n} r\left(-\frac{y_i}{\mu_i} - \ln(\mu_i)\right)$$

Pour estimer $\hat{\mu}_i = \exp(\hat{\eta}_i)$, il suffit de résoudre le système d'équations suivant :

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} r\left(-\frac{y_i}{\exp(\eta_i)} - \eta_i\right) = 0 \; pour \; j = 1, \dots, p$$

- **Modèle de régression de Weibull**

Une autre loi servant à modéliser le coût des sinistres est la loi de Weibull. Soit Y une variable aléatoire suivant une loi Gamma de paramètres de forme $s > 0$ et d'échelle r. La densité s'écrit :

$$f(y,s,r) = \frac{s}{r}\left(\frac{y}{r}\right)^{s-1} exp\left(-\left(\frac{y}{r}\right)^{s}\right) \quad y > 0$$

Pour un s fixé, la loi de Weibull fait partie de la famille exponentielle :

$$f(y,r) = \exp(y^s(-r^{-s}) + \ln(s) - sln(r) + (s-1)\ln(y))$$

tel que :

$$a(y) = y^s, b(r) = -r^{-s}, c(r) = -sln(r), v = sy^{r-1}$$

et

$$E(Y^s) = \mu(r) = r^s et\ Var(Y^s) = r^{2s}$$

Le lien canonique utilisé est le lien logarithmique de sorte que $\mu = \exp(\eta)$, avec $\eta = X\,\beta$ le prédicteur linéaire. La log-vraisemblance associée à ce modèle peut être écrite comme suit :

$$l(y,\beta) = \sum_{i=1}^{n}\left[{y_i}^s(-\left(e^{{X_i}'\beta}\right)^{-1} + \ln(s) - {X_i}'\beta + (s-1)\ln(y_i)\right]$$

Les paramètres β peuvent être estimés par la méthode du maximum de vraisemblance.

8.8 Modélisation de la probabilité que le sinistre cyber soit grave

Dans le chapitre précédent, nous avons défini le seuil pour classifier les sinistres en fonction de leur coût. Un sinistre dont le coût est supérieur à ce seuil est considéré comme extrême. Ainsi, nous disposons maintenant de réalisations binaires qui classifient les sinistres cyber en tant qu'extrêmes ou non. Notre objectif est de modéliser la probabilité p_i que le sinistre cyber i soit grave en fonction des variables explicatives disponibles. Pour ce faire, nous commencerons par utiliser la régression logistique. Nous optons pour la fonction de lien logit, comme elle est la plus employée en pratique pour la modélisation des probabilités :

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right)$$

Soit Y la variable aléatoire binomiale dont on cherche à modéliser la probabilité de succès étant $Y = 1$. La probabilité de succès peut s'écrire :

$$p_i = P[Y_i = 1|X_i = x_i] = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$$

Supposons $y_1, \dots, y_n$ les réalisations de variables aléatoires $Y_1, \dots, Y_n$ de loi binomiale $Bin(m_i, p_i)$. La vraisemblance associée s'écrit donc :

$$\mathcal{L}(y;\beta) = \prod_{i=1}^{n} \binom{m_i}{y_i} {p_i}^{y_i} (1-p_i)^{m_i - y_i}$$

En passant au logarithme, maximiser la vraisemblance revient à maximiser la quantité suivante selon β :

$$l(y;\beta) = \sum_{i=1}^{n} \left[y_i ln\left(\frac{exp(\eta_i)}{1+exp(\eta_i)}\right) + (m_i - y_i)\ln\left(\frac{1}{1+exp(\eta_i)}\right)\right]$$

de sorte à estimer $\hat{p}_i = \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)}$.

Pour évaluer la capacité du modèle à bien prédire, nous introduisons la matrice de confusion composée de quatre valeurs :

		Reality	
	Confusion matrix	**Negative : 0**	**Positive : 1**
Prediction	**Negative : 0**	**True Negative : TN**	**False Negative : FN**
	Positive : 1	**False Positive : FP**	**True Positive : TP**

Figure 24 : Matrice de confusion (Source : Kobia[70])

- Vrai Positif (TP) : les cas positifs réels qui ont été correctement prédits comme positifs par le modèle.
- Faux Positif (FP) : les cas négatifs réels qui ont été incorrectement prédits comme positifs par le modèle.
- Vrai Négatif (TN) : les cas négatifs réels qui ont été correctement prédits comme négatifs par le modèle.
- Faux Négatif (FN) : Les cas positifs réels qui ont été incorrectement prédits comme négatifs par le modèle.

Pour décrire la performance du modèle, nous utilisons donc les métriques basées sur la matrice de confusion suivantes :

- **L'exactitude (accuracy en anglais) :** Elle mesure le taux de prédictions correctes sur l'ensemble des prédictions :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

[70] Source : https://kobia.fr/classification-metrics-matrice-de-confusion/

- **Le rappel (recall en anglais) :** est adapté pour minimiser les faux négatifs, quand les conséquences de manquer des instances positives sont graves :

$$Recall = \frac{TP}{TP + FN}$$

- **La précision (precision en anglais):** elle permet de mesurer le coût des faux positifs, c'est-à-dire ceux détectés par erreur. Si l'on cherche à limiter les faux positifs, c'est cet indicateur que l'on va chercher à maximiser.

$$Precision = \frac{TP}{TP + FP}$$

- **F1-score :** est une moyenne harmonique de la précision et du rappel. Sa valeur est maximale lorsque le rappel et la précision sont équivalents.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Enfin, le F1-score a en commun avec l'accuracy de résumer la performance d'un modèle en un seul indicateur pour chaque seuil de classification. Il est plus complexe mais a l'avantage d'être robuste en présence de données déséquilibrées.

8.9 Limites d'un GLM

Bien que les GLM soient des outils utiles et polyvalents, leur utilisation nécessite des hypothèses sous-jacentes et des limitations pour obtenir des résultats précis et fiables. Les GLM supposent que les observations sont indépendantes les unes des autres, ce qui peut ne pas être vrai dans de nombreux cas. Les GLM ne capturent pas non plus la complexité dans les données. Il existe d'autres techniques d'apprentissage automatique (machine learning) qui peuvent être plus appropriées dans ces cas, telles que les arbres de régression et de classification - CART. La prochaine section se concentrera sur la présentation des bases théoriques des CART.

Chapitre 9 : Arbres de régression et de classification CART

9.1 Introduction

Le modèle d'apprentissage automatique que nous allons présenter dans ce mémoire est le modèle CART, signifiant Classification And Regression Trees. Cette méthode est introduite pour la première fois en 1984 par Breiman, Friedman, Olshen et Stone.

Le principe fondamental de CART repose sur une partition binaire itérative de l'espace des observations, suivie de la détermination de sous-partitions optimales pour prédire les données. La construction d'un arbre CART se déroule en deux phases distinctes. Tout d'abord, la construction d'un arbre maximal, qui établit une gamme de modèles parmi lesquels le meilleur sera sélectionné. Ensuite, une seconde phase, de « pruning » ou d'élagage, se concentre sur la création de sous-arbres optimaux à partir de l'arbre maximal. Détaillons chacune de ces étapes. Il existe de nombreuses références détaillées sur l'algorithme CART, notamment le livre originel de Breiman (1984).

9.2 Découpage de la base de données

Avant d'appliquer l'algorithme de création d'arbres de régression, la base de données initiale est séparée en deux sous-ensembles distincts :

- La base d'apprentissage : cette partie contient la plupart des observations et sert de support à l'algorithme de construction de l'arbre.
- La base de test : contrairement à la première, cette base n'est pas utilisée dans la création du modèle. Elle est réservée spécifiquement pour évaluer la capacité prédictive du modèle généré, ce qui permet de comparer les performances de différents modèles.

Il est à noter qu'un individu de la base de données initiale fait partie d'une et une seule des deux sous-bases.

9.3 Principe

L'algorithme des arbres de décision commence par segmenter la base de données en deux groupes. Pour ce faire, il choisit parmi les variables explicatives la variable et la valeur de cette variable qui va permettre d'avoir les deux groupes les mieux différenciés possible du point de vue de la variable à expliquer. Ces deux groupes nouvellement formés sont appelés des nœuds. Par la suite, l'algorithme réitère de manière récursive cette opération de segmentation dans les nœuds précédemment créés jusqu'à l'atteinte d'une condition d'arrêt. Cette condition peut être un nombre minimal d'individus dans un nœud ou un paramètre de complexité w qu'on définira après. Les nœuds terminaux, également connus sous le nom de feuilles, représentent les niveaux les plus bas de l'arbre et rassemblent des ensembles homogènes d'observations. La figure ci-dessous, donne un exemple d'un arbre CART[71].

[71] Source : Construction d'un arbre (Source : Institut des Actuaires - Atelier 100% Data Science « Quels algos pour quels usages en assurance auto ? »)

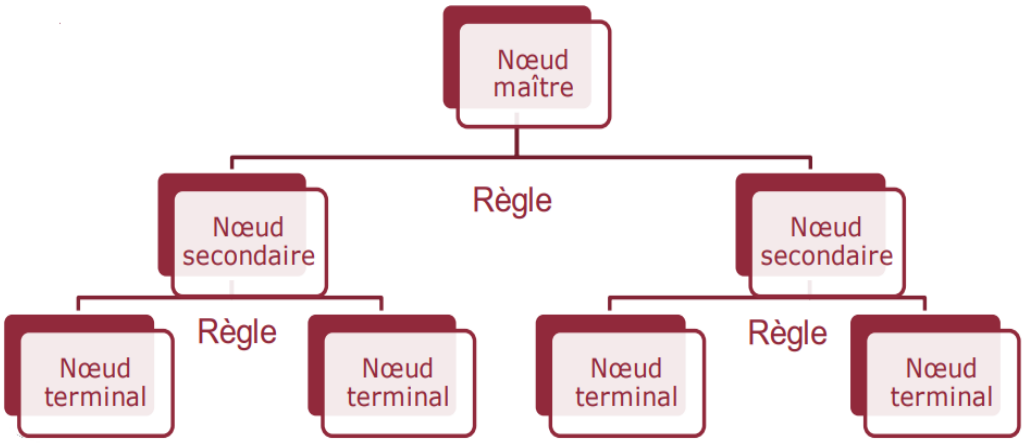

Figure 25 : Construction d'un arbre (Source : Institut des Actuaires - Atelier 100% Data Science « Quels algos pour quels usages en assurance auto ? »)

L'arbre de décision peut être utilisé en régression ou en classification. La distinction se fait au sujet des algorithmes. Elle distingue deux types de valeurs de sorties qu'on peut chercher à traiter. Dans le cadre d'un problème de classification, Y prend un nombre fini k de valeurs ($Y = \{1, \dots, k\}$). Tandis que dans un problème de régression, Y peut prendre une infinité de valeurs dans l'ensemble continu des réels.

9.4 Construction de l'arbre maximal

L'arbre maximal est l'arbre le plus grand que l'on peut former à partir d'une base de données, cependant, il est également le plus susceptible de présenter des problèmes de sur-apprentissage. À chaque étape de la partition, une portion de l'espace est divisée en deux sous-espaces distincts. On associe alors un arbre binaire à cette partition, où les nœuds de l'arbre correspondent aux éléments de la division. Par exemple, le nœud initial de l'arbre est lié à l'ensemble complet de l'espace d'entrée. Ses deux nœuds enfants sont associés aux deux sous-espaces générés après la première séparation, et ainsi de suite.

La règle de découpe varie selon le type de la variable explicative utilisée X^j :

- **Pour une variable explicative continue :** la division se fait en deux parties $\{X^j \leq d\} \cup \{X^j > d\}$, où $j \in \{1, \dots, p\}$ et $d \in R$. Cette découpe signifie que les observations ayant une valeur de la $j^{ème}$ variable inférieure à d sont dirigées vers le nœud fils de gauche, tandis que celles ayant une valeur supérieure à d vont dans le nœud fils de droite.

- **Pour une variable explicative catégorielle :** une division est de la forme $\{X^j \in d\} \cup \{X^j \in \bar{d}\}$, ou où d et $\bar{d}$ sont non vides et constituent une partition non vide de l'ensemble des modalités de la variable X^j.

L'algorithme recherche la meilleure découpe, c'est-à-dire le couple (j, d) qui minimise une fonction de coût spécifique :

- **En régression,** l'objectif est de minimiser la variance intra-groupe obtenue après la division d'un nœud t en 2 nœuds fils t_L et t_R. La variance d'un nœud t étant définie par $V(t) = \frac{1}{\#t}\sum_{i:x_i \in t}(y_i - \bar{y}_t)^2$ où $\bar{y}_t$ est la moyenne des y_i des observations présentes dans le nœud t et la fonction de coût est donc :

$$\frac{1}{n}\sum_{(x_i,y_i)\in t_L}(y_i - \bar{y}_{t_L})^2 + \frac{1}{n}\sum_{(x_i,y_i)\in t_R}(y_i - \bar{y}_{t_R})^2 = \frac{\#t_L}{n}V(t_L) + \frac{\#t_R}{n}V(t_R)$$

- **En classification,** généralement, l'indice de Gini est utilisé pour mesurer l'impureté des nœuds fils. L'indice de Gini d'un nœud t est défini par $\phi(t) = \sum_{c=1}^{L}\hat{p}_c(1 - \hat{p}_c)$, où $\hat{p}_c$ est la proportion d'observations de classe c dans le nœud t. L'objectif est de maximiser la réduction de l'indice de Gini après la division par rapport au nœud initial :

$$\phi(t) - \left(\frac{\#t_L}{\#t}\phi(t_L) + \frac{\#t_R}{\#t}\phi(t_R)\right)$$

L'algorithme continue à développer les arbres jusqu'à ce qu'une condition d'arrêt soit atteinte. Habituellement, cette condition consiste à éviter de diviser les nœuds contenant un nombre d'observations inférieur à un seuil défini. Les nœuds terminaux, qui ne sont plus divisés, sont appelés les feuilles de l'arbre. L'arbre maximal, noté T_{max}, est l'arbre pleinement développé.

9.5 Elagage de l'arbre maximal

La seconde étape de l'algorithme CART, appelée l'élagage, vise à chercher le meilleur sous-arbre élagué de l'arbre maximal. L'arbre maximal possède une variance élevée et un biais faible. L'élagage est une procédure de sélection de modèles, où les modèles sont les sous-arbres élagués de l'arbre maximal. Il s'agit de tous les sous-arbres binaires de T_{max} ayant la même racine que T_{max}. Entre T_{max}, le modèle de complexité maximale, qui conduit au surajustement aux données de l'échantillon d'apprentissage et l'arbre restreint à la racine qui est fortement biaisé, il s'agit de trouver l'arbre optimal parmi les admissibles.

Un élagage efficace vise à trouver un compromis entre la complexité de l'arbre et la précision de la prédiction. Cela implique de déterminer la valeur du paramètre de complexité c_p qui minimise une certaine erreur appelée erreur de validation croisée. L'arbre ayant la complexité la plus faible est celui qui prédit à tous les individus la même variable de sortie, correspondant à la moyenne des sorties de la base d'apprentissage. En revanche, l'arbre ayant la complexité la plus forte est l'arbre maximal. L'arbre optimal se trouve donc entre ces deux extrêmes. Cette valeur de c_p est ensuite utilisée comme règle d'arrêt pour former le nouvel arbre élagué.

9.6 Mesure de la qualité de la prédiction

L'arbre élagué est utilisé pour effectuer des prédictions sur la base de validation. Une métrique utilisée pour évaluer la qualité de prédiction, permettant également de comparer différents modèles de

prédiction, est l'erreur quadratique moyenne – RMSE, définie précédemment pour les cas de régression. Pour le cas de classification, les mêmes métriques que celles utilisées dans le cas du modèle de régression logistique, présentées dans la section précédente, seront utilisées, à savoir l'accurarcy et le F1-score.

9.7 Avantages et inconvénients de CART

L'algorithme CART présente plusieurs avantages, notamment :

- **Facilité de compréhension et d'interprétation :** les arbres de décision sont intuitifs à visualiser et à expliquer, car ils utilisent une logique simple basée sur des décisions successives.
- **Non-requête de normalisation :** contrairement à certains autres algorithmes, CART ne nécessite pas de normalisation des données avant son application.
- **Capacité à gérer de grandes quantités de données :** il peut traiter des ensembles de données volumineux, en travaillant avec des sous-ensembles de données plus petits.

Toutefois, malgré leurs nombreux avantages, les arbres de décision présentent des limites. Ils peuvent être sujets au surajustement dans les modèles non élagués, et ont tendance à créer des structures complexes et instables. Leur sensibilité accrue aux données d'apprentissage peut entraîner un ajustement excessif des modèles à ces données spécifiques, ce qui peut compromettre leur capacité à généraliser correctement sur de nouveaux ensembles de données.

Quatrième partie

Application aux données cyber

Cette partie a pour but de mettre en pratique les approches théoriques exposées dans la partie précédente, sur les sinistres cybers identifiés dans la base de données SAS OpRisk Global data. Il est à souligner que cette application s'appuie sur les données extraites à juillet 2023. Comme détaillé dans la section 6.2, nous avons pu identifier 1136 incidents cyber, engendrant une charge totale, indexée par le CPI[72], de 35,96 milliards de dollars.

Dans un premier temps, nous déterminons un seuil d'écrêtement des charges des sinistres en adoptons l'approche POT de la théorie des valeurs extrêmes. Nous procédons ensuite à la modélisation et à l'estimation des paramètres de la distribution des pertes extrêmes des sinistres cyber par la distribution généralisée de Pareto (GPD). La deuxième section de cette partie, se concentre sur la modélisation des charges des sinistres attritionnels à l'aide des GLM et arbres de régression CART. Enfin, nous modélisons la probabilité qu'un sinistre cyber soit grave, en utilisant les modèles de régression logistique et les arbres de classification.

Chapitre 10 : Application de la théorie des valeurs extrêmes

10.1 Choix du seuil d'écrêtement des charges des sinistres cyber

La partie cruciale de la méthode POT (Peaks-Over-Threshold) consiste à choisir le seuil au-delà duquel les données se comportent comme une GPD (Generalized Pareto Distribution). Toutes les méthodes présentées dans le chapitre 7 ont été exploitées pour déterminer un seuil. Un compromis entre le biais et la variance est soigneusement pris en considération pour sélectionner le seuil optimal.

Nous commençons par la méthode la plus rapide et pratique pour choisir un seuil. Comme illustré dans le tableau ci-dessous, les seuils sont définis en fonction des règles empiriques discutées dans la section 7.3.1. Pour chaque seuil, les paramètres d'échelle et de forme de la loi GPD calibrée sont estimés. Une disparité apparaît au niveau du nombre de dépassements du seuil pour le $90^{ème}$ percentile et le seuil $k = \sqrt{n}$ (où $n = 1136$ représente le nombre d'observations). Pour $k = \sqrt{n}$, on compte 148 dépassements, ce qui correspond approximativement au $87{,}05^{ème}$ percentile, par rapport aux 114 dépassements du $90^{ème}$ percentile. Cela entraîne un seuil plus élevé pour $k = 90^{ème}$ percentile avec une variance plus importante en raison du nombre limité de dépassements. Cependant, c'est le $90^{ème}$ percentile qui affiche les critères AIC et BIC les plus petits.

[72] Aux États-Unis, l'indice des prix à la consommation (CPI) mesure les variations dans les prix payés par les consommateurs pour un panier de biens et de services.

	$90^{ème}$ **percentile**	$k = \sqrt{n}$	$k = n^{2/3} / \log(\log(n))$
Seuil (en millions de dollars)	60,35	33,70	55,81
Nombre de dépassements	114	148	120
Paramètre d'échelle estimé	84,47	81,48	82,53
Paramètre de forme estimé	0,67	0,59	0,66
AIC	1 395,79	1 777,50	1 461,75
BIC	1 401,26	1 783,49	1 467,32

Tableau 18 : Valeurs des seuils, nombres d'observations au-delà du seuil, les estimateurs du ML pour les paramètres de forme et d'échelle et les critères AIC et BIC. Les paramètres sont estimés en utilisant le package « extrêmes » de R.

La Figure 25 présente le graphique Mean Residual Life (MRL), présenté dans la section 7.3.2, des charges des sinistres cyber, avec des intervalles de confiance de 95 %. Le graphique met en évidence la présence d'une pente positive, permettant ainsi de conclure à la présence d'une distribution des excès des charges des sinistres appartenant au domaine de GPD à queue épaisse. Il semble quelque peu linéaire jusqu'à un seuil d'environ 200 millions de dollars, où le comportement du graphique commence à changer.

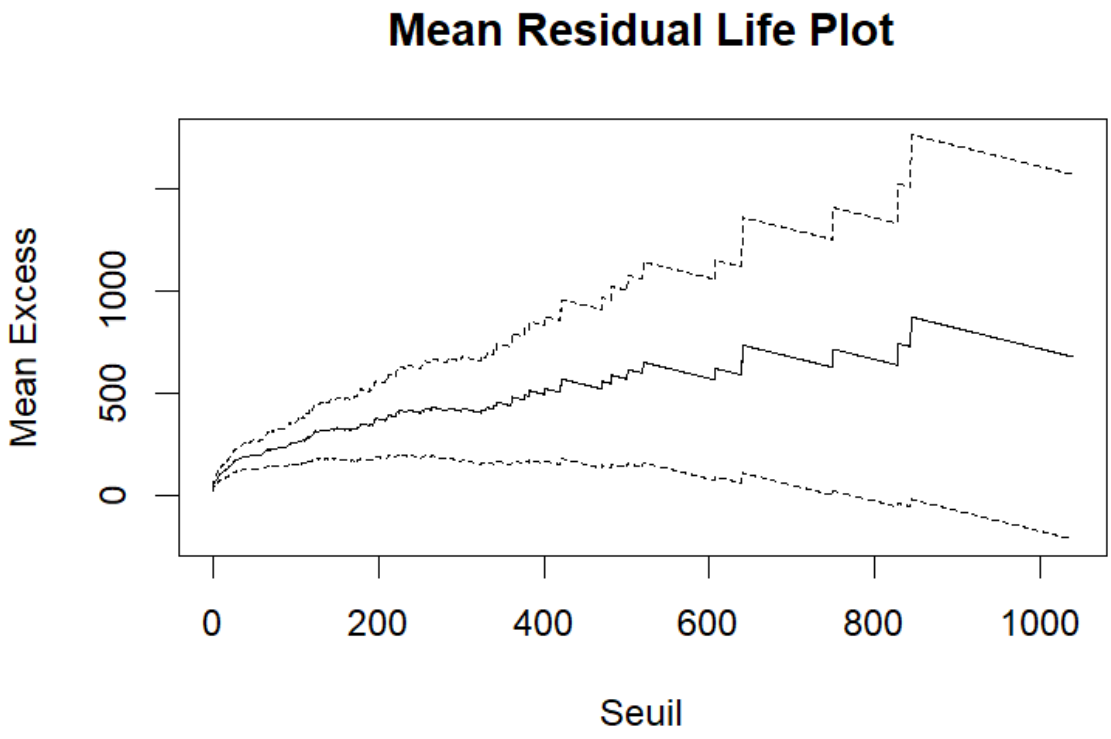

Figure 26 : Graphique de la moyenne des excès pour les charges des incidents cyber, tracé à l'aide de mrlplot du package « evd » dans R. La ligne pleine représente la MRL empirique, tandis que les lignes en pointillés représentent un intervalle de confiance à 95 %.

Comme les informations sur la durée de vie résiduelle (MRL) lorsque le seuil devient élevé n'apportent pas de conclusions claires du fait de la très forte volatilité de quelques observations extrêmes, nous examinons la Figure 18, où l'axe des x est limité à une plage plus restreinte de seuils. À partir de ce point, la théorie suggère qu'il devrait exister un point pour lequel, pour tous les seuils plus importants, l'excès moyen sera une fonction linéaire du seuil. Nous observons une certaine linéarité du graphe entre les seuils de 25 et 30 millions de dollars puis entre 55 et 60 millions de dollars.

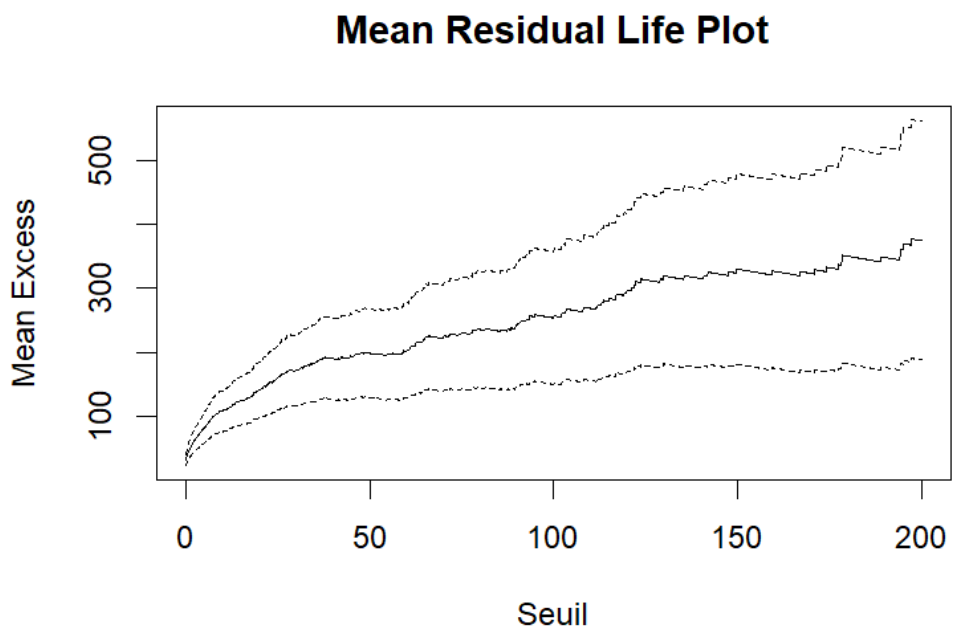

Figure 27 : Zoom sur le MRL des charges cyber tracé pour des valeurs de seuil inférieures à 200 millions de dollars

L'interprétation d'un MRL n'est pas toujours simple en pratique en raison de la subjectivité requise pour fixer un seuil en se basant uniquement sur une observation visuelle. Cela est également illustré dans l'exemple pratique présenté par Coles dans son livre « An Introduction to Statistical Modeling of Extreme Values ».

Les figures 27 et 28 présentent respectivement les graphiques de stabilité du paramètre d'échelle modifiée $\hat{\beta}_u - \hat{\xi}u$ et du paramètre de forme $\hat{\xi}$ pour différentes valeurs de seuils allant jusqu'à 100 millions de dollars. Comme déjà mentionné dans la section 7.3.2, notre objectif est de trouver un seuil pour lequel les estimateurs d'échelle modifiée et de forme sont approximativement constants. En observant le graphique d'échelle modifiée, il semble rester relativement constant entre les valeurs de seuils comprises entre 45 et 50 millions de dollars. Tout comme le graphique du paramètre d'échelle modifiée, le graphique de stabilité du paramètre de forme semble aussi se comporter comme une constante sur cet intervalle de seuils. Cependant, il est difficile d'obtenir une interprétation précise à partir de l'observation seule de la stabilité des paramètres.

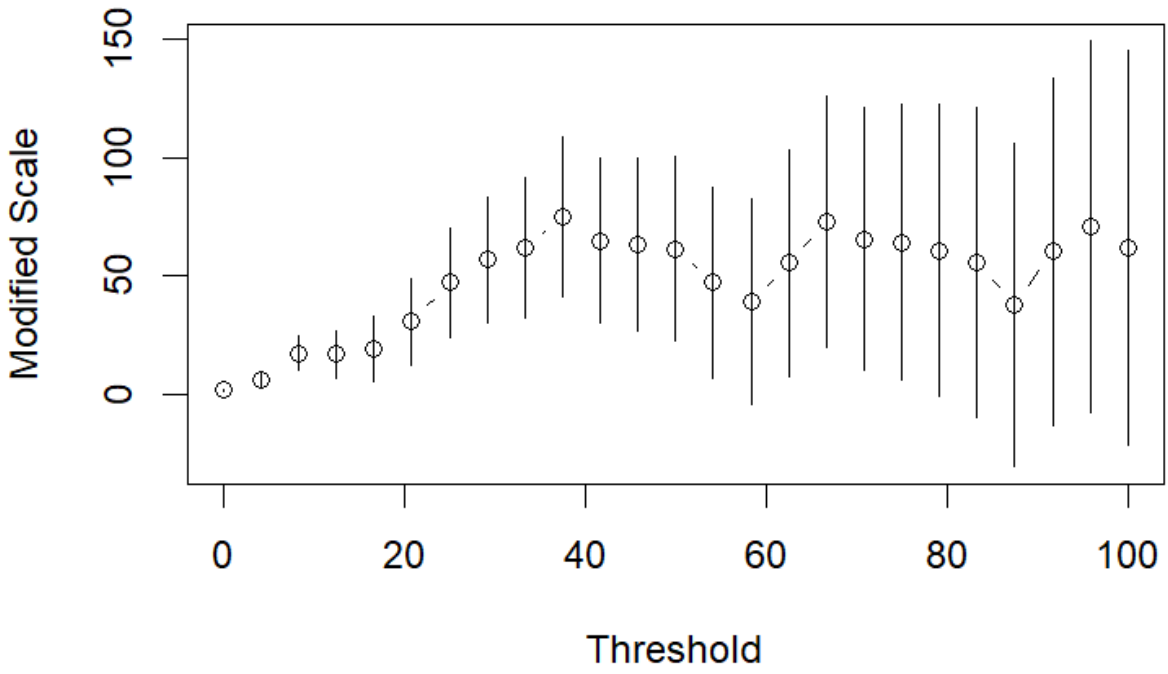

Figure 28 : Graphique de stabilité du paramètre d'échelle modifiée en fonction des seuils. Les ronds représentent l'estimation du paramètre d'échelle modifié pour un modèle GPD pour certains seuils. Les segments représentent un intervalle de confiance à 95 %. Le graphique est tracé avec tcplot du package « evd » dans R

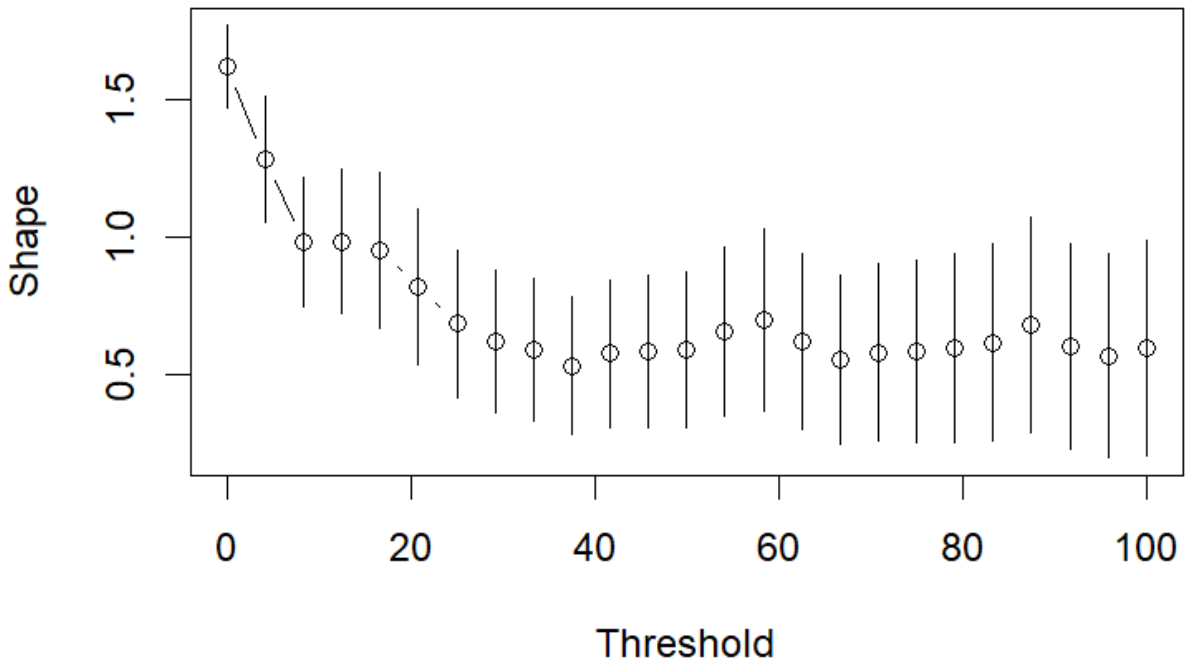

Figure 29 : Graphique de stabilité du paramètre de forme en fonction des seuils. Les ronds représentent l'estimation du paramètre de forme pour un modèle GPD pour certains seuils. Les segments représentent un intervalle de confiance à 95 %. Le graphique est tracé avec tcplot du package « evd » dans R

Ces méthodes graphiques fournissent un premier moyen pour réduire l'intervalle dans lequel peut se trouver le seuil approprié. Il semble se trouver aux alentours de 50 millions de dollars. Néanmoins, ces méthodes graphiques ne semblent pas suffisantes dans ce cas pour le déterminer plus précisément.

En complément, nous avons donc procédé à une analyse de sensibilité afin de déterminer le seuil permettant le bon compromis entre variance et biais. Nous estimons les paramètres de forme et d'échelle et les écarts types correspondants pour les seuils allant de 50 millions de dollars jusqu'au $90^{ème}$ percentile +0,5 millions de dollars, avec un pas de 0,5 millions de dollars. Les informations pour les différents seuils sont répertoriées dans le tableau suivant. Les deux lignes grisées correspondent respectivement aux seuils suivants :

- $90^{ème}$ percentile
- $n^{2/3} / \log(\log(n))$

Comme nous nous intéressons à la queue de distribution des charges, le paramètre de forme attire particulièrement notre attention plus que le paramètre d'échelle. Nous choisissons donc le seuil présentant la plus petite erreur standard du paramètre de forme estimé, soit 50 millions de dollars (en bleu dans le tableau). Il est à noter que ce seuil n'est pas très loin du seuil ayant la plus petite erreur du paramètre d'échelle estimé (54,4 millions de dollars), et qu'un seul dépassement supplémentaire existe entre les deux. De plus, le nombre d'observations est de manière très similaire autour de ces seuils et même si le seuil de 51 pouvait apparaître comme un meilleur compromis de minimum entre les deux paramètres, cela aboutissait au même nombre d'observations. Ainsi, le choix pour simplification de lecture a été de prendre 50 millions de dollars.

Dans ce qui suit, nous fixons le seuil des charges des sinistres cyber extrêmes à **50 millions de dollars**.

Seuil	Nombre de dépassements	Estimateur du paramètre d'échelle	Ecart type du paramètre d'échelle	Estimateur du paramètre de forme	Ecart type du paramètre de forme
50	123	91,8877	14,9586	0,5839	0,1450
50,5	123	90,5721	14,8507	0,5930	0,1466
51	123	89,2315	14,7407	0,6024	0,1482
51,5	123	87,8659	14,6289	0,6123	0,1500
52	122	88,8948	14,8186	0,6081	0,1499
52,5	122	87,5148	14,7041	0,6180	0,1517
53	122	86,1110	14,5897	0,6284	0,1536
53,5	122	84,6852	14,4750	0,6393	0,1556
54	122	82,9412	14,2991	0,6528	0,1579
54,5	122	81,4615	14,1796	0,6645	0,1600
55	121	82,4316	14,3743	0,6603	0,1601
55,5	120	83,4346	14,5744	0,6559	0,1602
55,81	120	82,5291	14,5040	0,6632	0,1616
56	120	81,9476	14,4594	0,6680	0,1625

56,5	119	82,9848	14,6654	0,6633	0,1625
57	119	81,4836	14,5520	0,6757	0,1649
57,5	119	79,9550	14,4397	0,6886	0,1675
58	119	78,3927	14,3280	0,7023	0,1702
58,5	118	79,4025	14,5425	0,6976	0,1703
59	117	80,4635	14,7627	0,6925	0,1704
59,5	115	84,3009	15,3284	0,6687	0,1676
60	114	85,5190	15,5653	0,6629	0,1675
60,35	114	84,4736	15,4905	0,6713	0,1692

Tableau 19 : Analyse des seuils compris entre 50 M$ et le 90ème percentile +0,5 M$, avec un pas de 0,5 M$. Les estimateurs de ML sont calculés à l'aide du package « extRemes » de R

10.2 Modélisation des charges des sinistres extrêmes par GPD

Une fois que le seuil d'écrêtement est défini, nous pouvons séparer les sinistres cyber de la base de données en sinistres extrêmes et sinistres attritionnels. Le résumé statistique des deux classes des sinistres est donné dans le tableau suivant:

	Nombre	Minimum	Q1	Médiane	Moyenne	Q3	Maximum
Sinistres attritionnels	1 013	0,11	0,47	1,54	5,36	5,84	48,50
Sinistres graves	123	51,92	77,99	120,00	248,19	244,11	3 505,35

Tableau 20: Analyse descriptive des charges des sinistres cyber attritionnelles et extrêmes (en millions de dollars)

Par la suite, les montants des sinistres au-delà du seuil sont modélisés via une loi GPD. Nous cherchons donc à calibrer une loi GPD sur les excès des charges des sinistres cyber au-delà de 50 millions de dollars. Dans cette étude, nous utilisons la méthode classique du maximum de vraisemblance (ML), présentée dans la section 7.4.1, pour estimer les paramètres de forme ξ et d'échelle β. Ces estimateurs sont présentés dans le tableau suivant. Le paramètre de forme estimé à 0,58 est positif, ce qui correspond à la loi de Fréchet. De plus, la valeur élevée de ξ (supérieure à 0,5) indique une probabilité relativement élevée d'observations extrêmes, ce qui signifie que les valeurs extrêmes sont plus probables par rapport à une distribution exponentielle. La valeur élevée de β (91,89) suggère une plus grande variabilité des charges des sinistres cyber extrêmes.

Seuil	50
Nombre de dépassement	123
Estimateur du paramètre d'échelle β	91,89
Estimateur du paramètre de forme ξ	0,58

Tableau 21 : Les estimateurs du ML des paramètres d'échelle et de forme de la GPD avec un seuil de dépassement de 50 millions de dollars

La fonction de densité empirique de la charge des sinistres au-delà du seuil et celle associée à la GPD estimée, dans la figure 29 semblent relativement proches, ce qui semble indiquer une calibration cohérente de la loi.

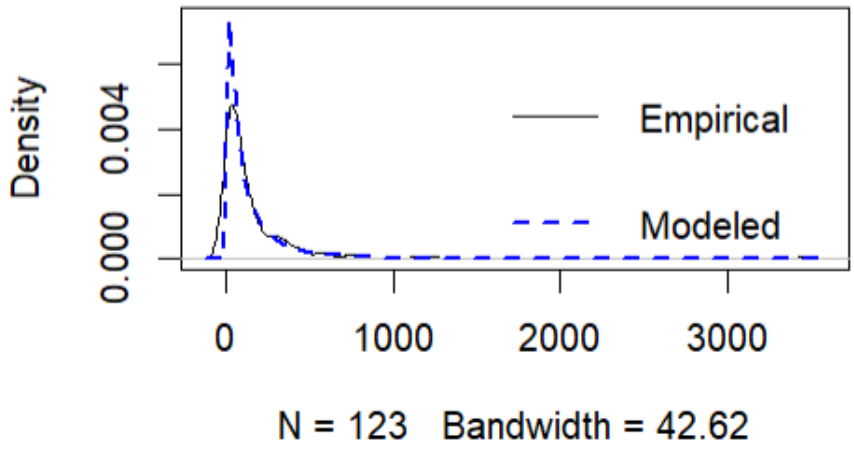

Figure 30 : Fonctions de répartition des données empiriques et de la GPD ajustée

Le graphique Quantile-Quantile (QQ-plot) est un outil qui permet d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique. Dans cette démarche, le QQ-plot représente les quantiles empiriques en fonction des quantiles théoriques afin d'étudier visuellement la linéarité entre ces deux quantités. L'axe des abscisses représente les quantiles empiriques et l'axe des ordonnées représente les quantiles théoriques de la loi considérée. Si les données suivent la loi théorique, alors les points sont alignés. La figure 30 présente le QQ-plot du modèle GPD calibré aux excès des coûts des sinistres cyber de la base au-delà du seuil de 50 millions de dollars. La relative linéarité de la courbe par rapport à la droite indique un bon ajustement entre la distribution empirique des charges extrêmes des sinistres cyber et la distribution théorique da la GPD. Ce qui nous conforte par rapport au choix du modèle GPD.

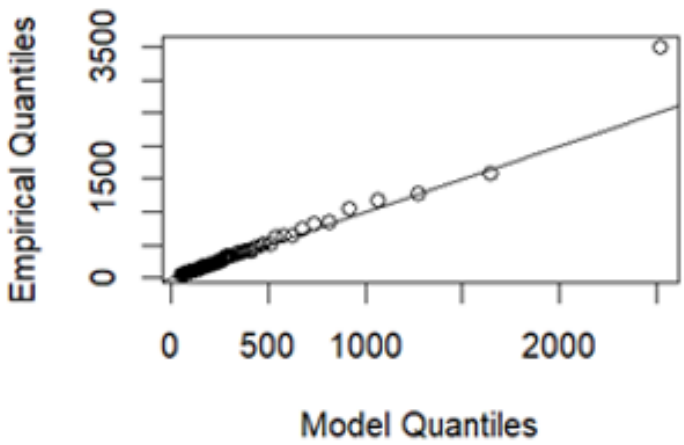

Figure 31 : QQ-plot de la loi GPD calibrée aux excès des charges des sinistres cyber avec le seuil de 50 millions de dollars

Chapitre 11 : Modélisation de la charge des sinistres cyber attritionnels

Ce chapitre est dédié à la modélisation de la charge des sinistres cyber en dessous du seuil des extrêmes de 50 millions de dollars identifié dans la section 10.1. L'objectif est d'évaluer le coût d'un sinistre cyber attritionnel en fonction des variables explicatives. Nous modéliserons d'abord ce coût à partir de modèles GLM puis via des arbres CART.

11.1 Séparation des données en base d'apprentissage et base de test

Une étape primordiale à la modélisation est la séparation des données en jeu d'apprentissage et jeu de test. Les données sont séparées avec les proportions suivantes : 70 % pour l'apprentissage et 30 % pour le test. Un échantillonnage aléatoire des risques cyber attritionnels nous permet d'assurer qu'il n'y a pas de biais à priori introduit lors de la construction de ces bases.

Ainsi, à partir des 1013 sinistres attritionnels, nous obtenons :

- 709 observations dans la base d'apprentissage ;
- 304 observations dans la base de test.

Les modèles de la charge des sinistres cyber seront construits sur la base d'apprentissage, puis testés en effectuant des prédictions sur la base de test. Les statistiques descriptives des coûts indexés des sinistres cyber dans les deux bases sont donnes dans le tableau 22. Elles sont globalement cohérentes.

Base	Nombre	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type
App	709	0,12	0,45	1,48	5,1	5,26	48,5	8,34
Test	304	0,11	0,48	1,90	5,97	6,27	47,31	9,32

Tableau 22 : Statistiques descriptives des coûts indexés des sinistres cyber attritionnels dans la base d'apprentissage et de test (en millions de dollars)

11.2 Modélisation de la charge par les GLM

Les coûts indexés par le CPI[73] des sinistres cyber en dessous du seuil défini seront d'abord modélisés à l'aide des GLM. Dans un premier temps, un regard sur la log vraisemblance et les critères AIC et BIC pour les lois classiques Exponentielle, Gamma, Log-normale et Weibull, largement utilisées pour modéliser les charges des sinistres en assurance non vie, nous permet de déterminer si les sinistres attritionnels sont proches d'une modélisation via l'une de ces distributions. A partir du tableau ci-après, parmi les quatre distributions, c'est la loi de Weibull qui donne les meilleurs résultats. Elle présente en effet, la plus grande log-vraisemblance et les plus petits critères AIC et BIC, et donc elle s'adapte le mieux à la distribution des charges indexées des sinistres cyber attritionnels. Elle est suivie de près par la loi de Gamma.

[73] Aux États-Unis, l'indice des prix à la consommation (CPI) mesure les variations dans les prix payés par les consommateurs pour un panier de biens et de services.

Modèle	Log-vraisemblance	AIC	BIC
Exponentielle	-1 863,93	3 729,86	3 734,42
Gamma	-1 755,47	3 514,94	3 524,07
Log-normale	-11 468,62	22 941,24	22 950,37
Weibull	**-1 725,79**	**3 455,57**	**3 464,70**

Tableau 23 : Analyse de l'adéquation de l'ajustement des lois classiques aux charges des sinistres cyber attritionnels

La représentation graphique ci-dessous, confirme le choix de Weibull et Gamma. Parmi les quatre lois examinées, elles s'adaptent le mieux aux données empiriques.

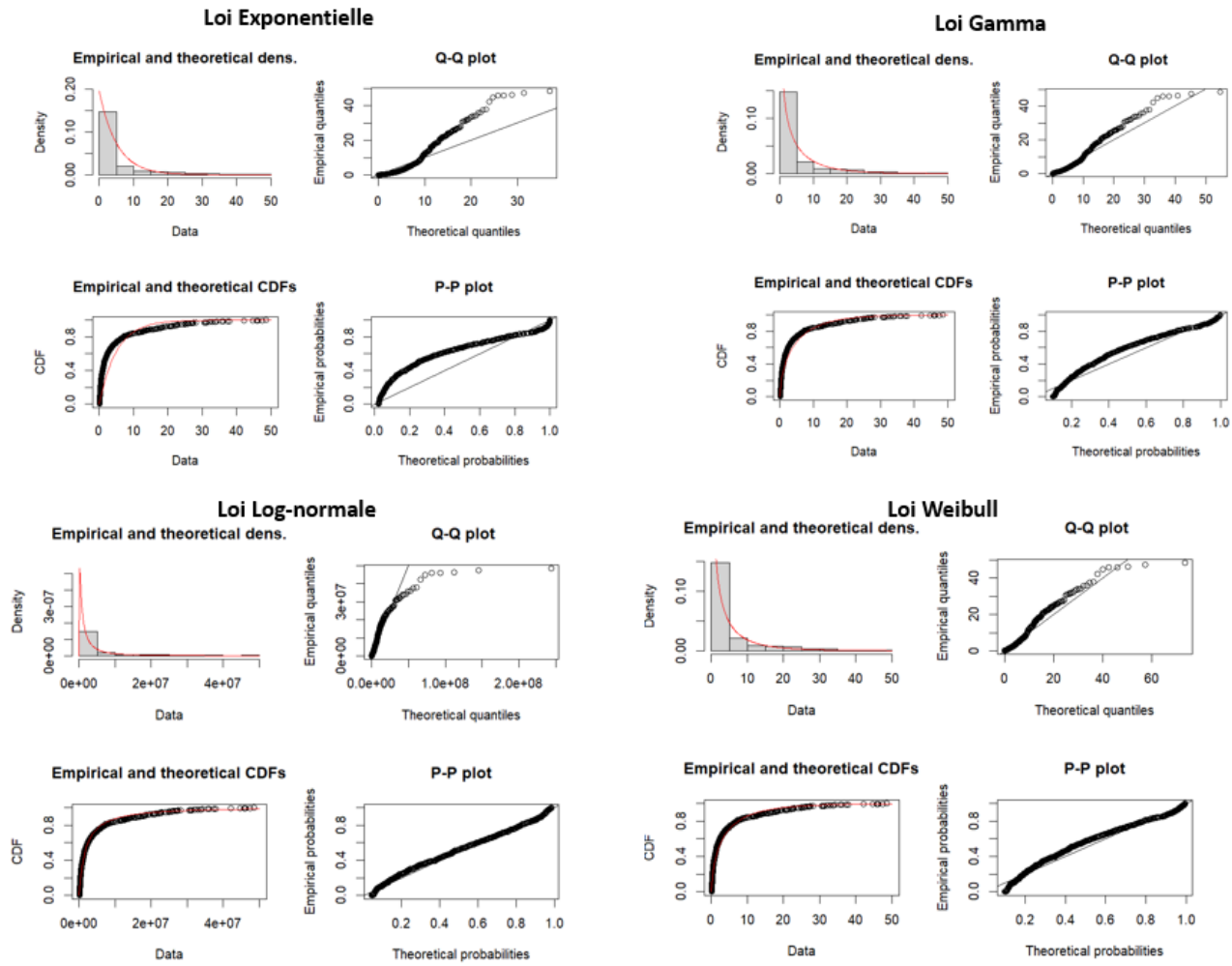

Figure 32: Analyse de l'adéquation de l'ajustement des lois classiques aux charges des sinistres cyber attritionnels

Ainsi, les lois de Weibull et gamma sont retenues pour la modélisation des charges des sinistres cyber attritionnels.

11.3 Choix des variables explicatives

Une question cruciale se pose lors de la modélisation des coûts : quelles variables utilisées pour expliquer les montants des sinistres ? Comme mentionné dans la présentation de la base de données (section 6.1), la base SAS OpRisk Global data comprend 49 variables, englobant celles liées à l'entreprise victime et d'autres associées au type d'incident. Cependant, toutes les variables ne sont pas pertinentes pour expliquer le coût d'une perte cyber.

La modélisation des coûts doit être cohérente en sélectionnant méthodiquement les variables pour ne retenir que les plus pertinentes. Dans un premier temps, ont ainsi été exclus les champs incomplets. De plus, la revue de la littérature et du monde réel nous permettent de faire une première sélection des variables explicatives potentielles.

Dans le chapitre 3 de la première partie, nous avons vu que la distribution des sinistres cyber, tant en nombres qu'en montants, présente des variations en fonction du business Line, du secteur industriel et de la zone géographique. La taille de l'entreprise elle-même n'est pas non plus sans effet sur les coûts liés à un sinistre cyber, avec des disparités notables entre les entreprises de différentes tailles.

La base de données fournit des informations sur trois niveaux de la ligne métier (business line). En effet, nous avions d'abord pensé à séparer la base des sinistres attritionnels entre secteur financier et non financier compte tenu des différences notables de la sinistralité cyber entre les entreprises de ces secteurs mais compte tenu de la taille réduite des populations que cela aurait induit, cette distinction n'a pas été retenue. En conséquence, nous avons choisi de travailler avec le niveau 2 de la ligne métier, appelé « Basel Business Line - Level 2 ». En effet, il apparait comme le plus pertinent pour avoir un bon niveau d'information, puisque le niveau 1 est renseigné uniquement pour le secteur financier et le niveau 3 est renseigné uniquement pour le secteur non financier.

En ce qui concerne le secteur industriel, nous utilisons le champ « Industry Sector Name », et créons une nouvelle variable appelée « Secteur », qui indique le type du secteur : Financier ou Non Financier.

Pour la zone géographique, la base fournit le pays du siège social et sa région, ainsi que le pays de survenance du sinistre. Pour une meilleure représentation, nous préférons travailler avec les régions géographiques plutôt qu'avec les pays. Nous disposons déjà de la région du siège social, notée « Region of Domicile », et créons à son instar, une nouvelle variable « Region of Incident », qui spécifie la région de survenance du sinistre.

Nous utilisons aussi la classification des événements opérationnels selon Bâle II, qui est renseignée dans le champ « Event Risk Category »[74].

[74] Voir la section 6.1.

Enfin, bien que la base de données fournisse le nombre de salariés, une trentaine d'entreprises présentent des informations manquantes. Pour remédier à cela, nous avons complété ces données, en consultant les sites internet de Wikipédia et Glassdoor. Le nombre de salariés est ensuite converti en une nouvelle variable, appelée « Taille Entreprise », selon la classification suivante :

TPE	PME	ETI	GE
Très petites entreprises	Petites et moyennes entreprises	Entreprise de taille intermédiaire	Grande entreprise
< 10 salariés	**Entre 10 et 250 salariés**	**Entre 250 et 5 000 salariés**	**> 5 000 salariés**

Tableau 24 : Taille de l'entreprise en fonction du nombre de salariés (Source : hubspot[75])

Après avoir effectué une première sélection des variables explicatives de la charge des sinistres, nous utilisons ensuite des tests statistiques pour affiner davantage cette sélection.

11.4 Analyse exploratoire et recodage des variables

Toutes les variables explicatives retenues sont catégorielles. Une analyse exploratoire sera réalisée sur l'ensemble de ces variables pour identifier celles à retenir. Cette analyse exploratoire se fera sur la base d'apprentissage.

11.4.1 Intensité des associations entre les variables explicatives

Afin de mesurer l'intensité des associations entre les variables explicatives candidates deux à deux, nous calculons le coefficient du V de Cramer présenté dans la section 8.6.2. La matrice suivante donne les coefficients obtenus :

V Cramer	Event Risk Category	Region of Domicile	Region of Incident	Taille entreprise	Industry Sector Name	Secteur
Basel Business Line - Level 2	0,31	0,25	0,26	0,23	0,98	0,97
Event Risk Category		0,06	0,11	0,08	0,16	0,25
Region of Domicile			0,82	0,16	0,14	0,21
Region of Incident				0,12	0,15	0,26
Taille entreprise					0,13	0,05
Industry Sector Name						1

Tableau 25 : Matrice des coefficients du V de Cramer calculés

[75] Source : https://blog.hubspot.fr/sales/taille-entreprise-classification

Nous rappelons la règle d'interprétation de cette mesure dans le tableau ci-dessous :

V de Cramer	Interprétation
$[0-0{,}2]$	Les champs ont un lien d'association faible
$]0{,}2-0{,}6]$	Les champs ont un lien d'association modéré
$]0{,}6-1]$	Les champs ont un lien d'association fort

Tableau 26 : Règle empirique d'interprétation du coefficient du V de Cramer

Ainsi, nous constatons sans surprise une forte association entre les deux champs « Industry Sector Name » et « Secteur », entre « Basel Business Line - Level 2 » et « Industry Sector Name », ainsi qu'entre « Basel Business Line - Level 2 » et « Secteur ». Ceci est logique, puisque toutes ces variables sont liées au secteur d'industrie, et donc apportent des informations en partie similaires. Le champ « Industry Sector Name » ne sera donc pas utilisé dans le modèle de prédiction du coût d'un incident cyber.

De manière similaire, une forte association est relevée entre les deux champs « Region of Domicile » et « Region of Incident ». Ces champs ne pourront donc pas forcément rentrer ensemble dans le modèle de prédiction du coût d'un incident cyber.

11.4.2 Détection des associations entre la variable du coût indexé et les variables explicatives

Nous cherchons à étudier la relation entre le coût et les différentes variables explicatives. Le test statistique non paramétrique de Kruskal-Wallis, décrit dans la section 8.6.2, permet d'évaluer l'influence de chacune des variables explicatives sur la variable cible « « Current Value of Loss ». Il est essentiel de rappeler que sous l'hypothèse nulle, pour une variable explicative donnée, tous les sous-groupes par modalités sont confondus, indiquant ainsi l'absence d'effet de cette variable explicative sur la variable à expliquer. L'hypothèse nulle est rejetée si la p-value du test est inférieure au seuil de signification, ici fixé à 0,05.

Le tableau ci-dessous, présente les résultats du test pour chaque variable explicative. Il est à souligner que seules les variables « Basel Business Line – Level 2 », « Event Risk Category », « Industry Sector Name » et « Secteur » ont un effet significatif sur le coût indexé des incidents cyber. L'influence du secteur d'industrie correspond à ce qui est observé en réalité (voir section 3.5). Par contre, contrairement à nos attentes, la taille de l'entreprise et la région géographique ne semble pas avoir de pouvoir discriminant sur le coût d'un incident cyber, dans le cas des données utilisées. Les champs « Taille entreprise », « Region of Incident » et « Region of Domicile » seront quand même utilisés dans la modélisation, même si le test les faisait apparaitre comme non significatifs. Il y a un priori fort à ce qu'ils ne soient pas utilisés dans la modélisation finale.

Variable	Statistique	p-value	Degré de liberté
Basel Business Line - Level 2	114,58	< 0,05	10
Event Risk Category	23,18	< 0,05	3
Industry Sector Name	58,36	< 0,05	8
Region of Domicile	8,84	0,0653	4
Region of Incident	4,46	0,2162	3
Secteur	52,77	< 0,05	1
Taille entreprise	4,65	0,0976	2

Tableau 27 : Test de Kruskall-Wallis appliqué aux variables explicatives et le coût des sinistres cyber

11.4.3 Recodage des variables explicatives

Plusieurs variables explicatives comportent un grand nombre de modalités, comme le montre le tableau suivant. Leur utilisation sans prétraitement conduirait à la création des modèles inutilement complexes. En effet, les variables qualitatives à k modalités sont souvent décomposées en une série de $k-1$ variables binaires dans le modèle. Le nombre de paramètres en jeu serait excessif et plusieurs modalités seraient non significatives du fait de leur faible importance.

Variable	Nombre de modalités
Basel Business Line - Level 2	37
Event Risk Category	6
Region of Domicile	6
Region of Incident	6
Secteur	2
Taille Entreprise	4

Tableau 28 : Nombres de modalités des variables explicatives

Nous procédons donc au regroupement des classes par fréquences. Les fréquences cumulées sont calculées et un seuil de 95% des observations est fixé. Les classes qui comptent moins de 5% des observations seront agrégées dans une nouvelle classe désignée « Others ».

A titre d'exemple, prenons la variable « Event Risk Category » qui comporte 6 classes. Le nombre de sinistres cyber et les pourcentages des fréquences cumulées sont présentés dans le tableau ci-après. En fixant le seuil des observations à 95%, les modalités « Clients, Products & Business Practices » et « Employment Practices and Workplace Safety », ne représentent que 4% des observations en effectif. Par conséquent, nous regroupons ces deux modalités dans une seule classe appelée « Others ».

Event Risk Category	Effectif	Pourcentage cumulé du nombre total des sinistres
External Fraud	362	51,06%
Internal Fraud	174	75,60%
Execution, Delivery & Process Management	102	89,99%
Business Disruption and System Failures	43	96,05%
Clients, Products & Business Practices	27	99,86%
Employment Practices and Workplace Safety	1	100,00%

Tableau 29 : Nombres de sinistres cyber et pourcentages des fréquences cumulées en fonction des classes de la variable « Event Risk Category »

Le regroupement des modalités qui ne sont pas significatives est appliqué pour toutes les autres variables explicatives.

Désormais, les variables explicatives retenues pour la modélisation des coûts des sinistres cyber attritionnels sont les suivantes:

- Basel Business Line - Level 2
- Event Risk Category
- Region of Incident
- Region of Domicile
- Taille entreprise
- Secteur

Nous sommes maintenant prêts à construire les modèles explicatifs des charges indexées des sinistres cyber attritionnels.

11.5 Modélisation du coût d'un incident cyber attritionnel

11.5.1 Modèles linéaires généralisés

Dans cette section, nous construisons des GLM, pour expliquer la variable cible « Current Value of Loss » à l'aide des six variables explicatives les plus pertinentes identifiées précédemment.

Comme observé dans la section 10.2, les distributions de Weibull et gamma ont montré les meilleures adaptations aux pertes des sinistres attritionnels. Par conséquent, nous utilisons d'abord la distribution de Weibull, suivie de la distribution gamma.

- **Modèle GLM de Weibull**

Nous commençons donc par modéliser les charges par la loi de Weibull. La fonction de lien canonique utilisée est la fonction logarithme. Nous commençons par construire un modèle complet : Full model, c'est-à-dire incluant toutes les variables explicatives présélectionnées. Ensuite, nous utilisons la régression Stepwise, et plus précisément la méthode Forward, expliquée dans la section 8.6.1, pour sélectionner les variables les plus significatives à intégrer dans le modèle. Cette méthode démarre avec

un modèle sans variables explicatives : Null model, et ajoute successivement la variable qui optimise le critère de sélection (AIC). Le processus s'arrête lorsque l'amélioration du modèle n'est pas significative, indépendamment de la variable ajoutée. Les variables retenues par l'algorithme sont :

- Basel Business Line - Level 2
- Event Risk Category
- Taille entreprise
- Secteur

La régression Stepwise Forward n'a pas donc retenu les deux régions : « Region of Incident » et « Region of Domicile », les jugeons non pertinentes conformément aux résultats précédents du test de Kruskall-Wallis (tableau 25), mais contrairement au test, elle a gardé le champ « Taille entreprise ».
Pour évaluer la qualité de l'ajustement et comparer les trois modèles GLM de Weibull construits, nous utilisons la déviance et le critère AIC, présentés dans le tableau suivant. Comme expliqué dans la section 8.5, le meilleur modèle est celui qui a la plus petite déviance et le plus petit citère AIC. Le modèle complet (Full model) présente la plus petite déviance, et le modèle Stepwise Forward a le plus petit AIC. Cependant, les écarts entre les deux modèles ne sont pas très prononcés, ce qui montre que les variables non retenues présentent un impact très peu significatif avec un bruit faible qu'elles soient intégrées ou non dans la modélisation.

Modèle	**Null model**	**Full model**	**Model Stepwise Forward**
Déviance	3 451,57	**3 285,19**	3 298,02
AIC	3 455,57	3 353,19	**3 352,02**

Tableau 30 : Déviances et AIC des modèles GLM de Weibull

Afin d'évaluer la qualité prédictive des modèles construits, nous les testons sur la base de test. La métrique utilisée est l'erreur quadratique moyenne (RMSE), présentée dans la section 8.5, et qui évalue la dispersion des écarts entre les coûts observés et les coûts prédits par le modèle. Il est important de noter que plus la RMSE est faible, plus les prédictions du modèle sont proches des valeurs réelles, indiquant ainsi une meilleure précision du modèle. Les valeurs RMSE des deux modèles GLM de Weibull, à savoir le modèle complet (Full model) et le modèle de la régression Stepwise Forward (Model Stepwise Forward), sont présentées dans le tableau suivant, exprimées dans la même unité que la variable à prédire, c'est à dire en millions de dollars. Au vu des résultats, il est déduit que l'utilisation de la régression Stepwise Forward plutôt que toutes les variables explicatives permet d'obtenir un modèle avec un pouvoir prédictif plus important, comme indiqué par une RMSE plus petite. Sachant que les sinistres cyber attritionnels de la base de test totalisent une charge de 1506,88 millions de dollars, la RMSE de 9,39 millions de dollars représente donc 0,62 % de cette charge totale. Le niveau d'erreur du modèle sur la base de test est relativement faible.

Modèle	**Full model**	**Model Stepwise Forward**
RMSE	9,53	**9,39**

Tableau 31: RMSE des modèles GLM de Weibull (en millions de dollars) sur la base de test

- **Modèle GLM gamma**

De la même manière que pour le modèle de régression de Weibull, nous procédons avec le modèle gamma. La fonction de lien canonique utilisée est la fonction logarithmique. Nous commençons par construire un modèle complet : Full model, c'est-à-dire comprenant toutes les variables explicatives présélectionnées. Nous utilisons ensuite la régression Stepwise Forward, en partant du « Null model », qui est le modèle vide équivalent à une constante, et qui sert de point de comparaison.

Les variables sélectionnées par la régression Stepwise Forward sont les mêmes que pour le modèle de Weibull, à savoir :

- Basel Business Line - Level 2
- Event Risk Category
- Taille entreprise
- Secteur

En examinant la déviance et le critère AIC relatifs aux trois modèles GLM gamma construits, présentés dans le tableau ci-après, nous constatons que le meilleur modèle est le modèle complet. Il affiche la plus petite déviance et le plus petit AIC.

Modèle	Null model	Full model	Model Stepwise Forward
Déviance	1 612,80	**1 308,30**	1 333,30
AIC	23 130,00	**23 004,00**	23 007,00

Tableau 32 : Déviances et AIC des modèles GLM gamma

En se basant sur la RMSE pour évaluer le meilleur modèle GLM gamma des coûts des sinistres cyber attritionnels, nous déduisons que le modèle Stepwise Forward est le meilleur. En effet, comme présenté dans le tableau suivant, le modèle Stepwise Forward affiche une RMSE plus faible que celle du modèle complet. Sachant que les sinistres cyber attritionnels de la base de test totalisent une charge de 1506,88 millions de dollars, la RMSE de 9,32 millions de dollars représente donc 0,62% de cette charge totale. Le niveau d'erreur du modèle sur la base de test est relativement faible. Nous déduisons donc que l'utilisation de la régression Stepwise Forward plutôt que toutes les variables explicatives permet d'obtenir un modèle avec un pouvoir prédictif plus important.

Modèle	Full model	Model Stepwise Forward
RMSE	9,60	**9,32**

Tableau 33: RMSE des modèles GLM gamma (en millions de dollars) sur la base de test

- **Conclusions**

Il est à conclure que pour les deux cas de Weibull et gamma, le modèle GLM le plus performant est celui construit avec les variables sélectionnées par la régression Stepwise Forward. De plus, les deux modèles de régression Stepwise Forward, arrivent à la même sélection de variables explicatives ce qui montre une cohérence dans le choix des variables pertinentes.

Les résultats présentés dans cette section, mettent aussi en lumière une meilleure qualité d'ajustement et de prédiction des GLM gamma par rapport aux GLM de Weibull, malgré des attentes contraires basées sur les résultats de la section 11.2, suggérant que la loi de Weibull s'ajuste légèrement mieux aux pertes attritionnelles que la loi gamma. Nous obtenons en effet, des niveaux de déviance, critère AIC et RMSE du GLM gamma plus petits que ceux du GLM de Weibull.

Pour remédier aux limites des GLM discutées dans la section 8.10, nous allons désormais mettre en œuvre les arbres de décision CART afin de mieux prendre en compte les spécificités des données et comparer les résultats de cet algorithme avec les résultats obtenus avec les GLM.

11.5.2 Modèle CART

Cette section présente les résultats de l'application des arbres de régression CART aux charges des sinistres cyber attritionnels, sous le logiciel de R. Comme pour les GLM, les modèles CART sont construits sur la base d'apprentissage et validés sur la base de test. Les résultats de cette section fourniront des informations sur la performance des modèles CART par rapport aux modèles GLM déjà présentés, permettant ainsi une comparaison des approches pour la modélisation des coûts des sinistres cyber attritionnels.

- **Construction de l'arbre maximal**

Lorsque l'on construit un arbre de décision, l'arbre maximal, également appelé "arbre saturé", est le plus grand arbre possible qui pourrait être formé à partir de la base d'apprentissage. Dans le contexte de notre étude, le package « rpart » est utilisé sous R pour construire des arbres de décision.

Nous parvenons à construire l'arbre maximal, mais celui-ci se révèle peu robuste, instable et souffre d'un sur apprentissage prononcé en raison de sa forte dépendance aux données d'apprentissage. Afin d'améliorer sa performance, une étape d'optimisation est requise, notamment par le biais de l'élagage.

- **Elagage de l'arbre maximal**

Nous cherchons à obtenir un arbre plus concis et donc plus robuste, en partant de l'arbre maximal. Pour ce faire, nous choisissions de tracer le graphique ci-après de la décroissance de l'erreur relative ou

encore l'erreur de validation croisée en fonction du paramètre de complexité. Il faut déterminer la valeur du paramètre de complexité c_p qui minimise l'erreur de validation croisée.

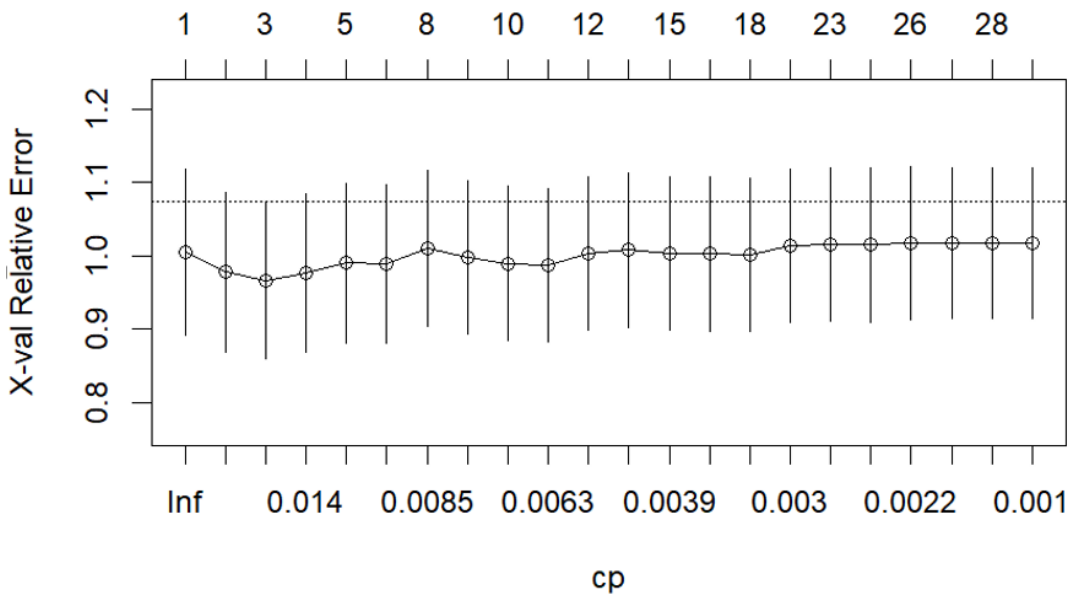

Figure 33 : Décroissance de l'erreur relative en fonction du paramètre de complexité de l'arbre

Cette valeur de c_p peut être aussi directement obtenue avec le code R. Nous trouvons une valeur de c_p de 0,01627446. Elle est ensuite utilisée comme règle d'arrêt pour construire le nouvel arbre élagué.

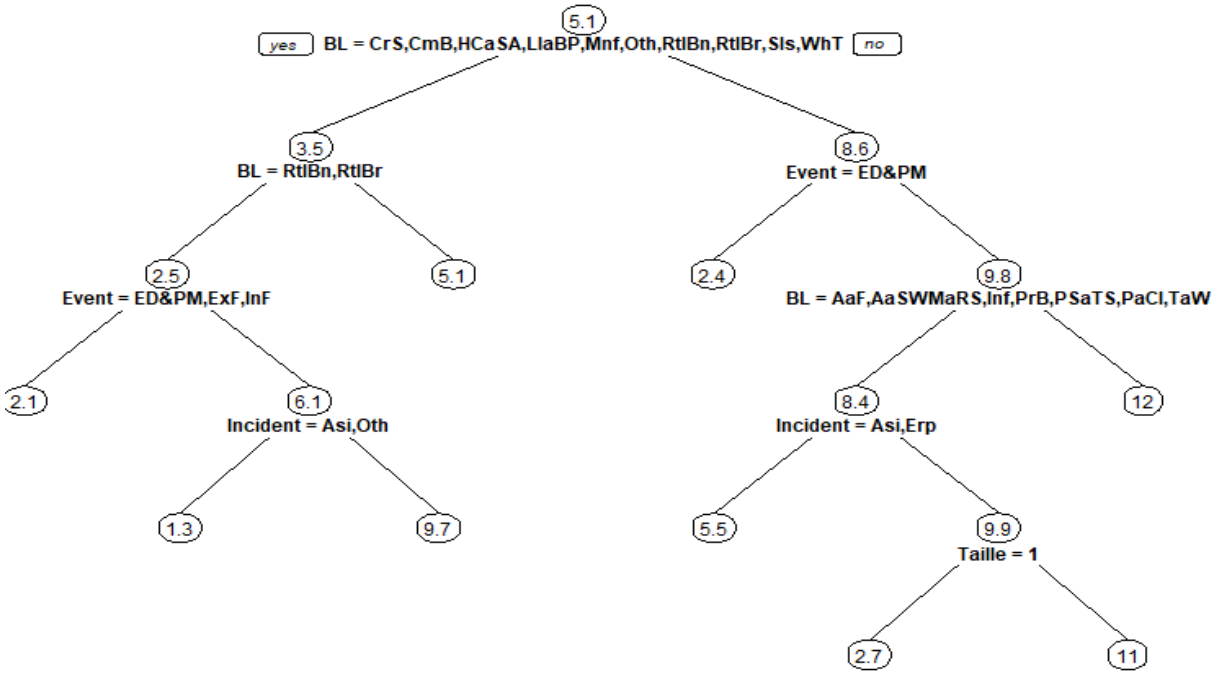

Figure 34 : Arbre CART élagué sur la base d'apprentissage[76]

Nous avons gagné en simplicité par rapport à l'arbre maximal. Nous constatons notamment l'importance de la variable du business line « Basel Business Line-Level 2», et de la variable de la classe d'événement opérationnel « Event Risk Category ». Cependant, contrairement à la régression stepwise forward, l'algorithme CART intègre la variable de la région de survenance du sinistre, et n'utilise pas la variable « Secteur ».

Nous appliquons l'arbre élagué aux données d'apprentissage et de test, et puis calculons les RMSE. Les résultats sont présentés dans le tableau suivant. Sachant que les sinistres cyber attritionnels de la base de test engendrent une charge totale de 1506,88 millions de dollars, les RMSE de 7,84 et 9,38 millions de dollars ne représentent respectivement que 0,52 % et 0,6 2% de la charge totale. Les niveaux d'erreur des CART sont relativement faibles, témoignant donc d'un bon pouvoir prédictif. De plus, l'écart entre les RMSE de l'arbre élagué appliqué aux données d'apprentissage et de test est relativement faible (1,54 millions de dollars), indiquant ainsi que le risque de sur-apprentissage n'est pas très fort.

Données	**RMSE**
Base apprentissage	7,84
Base test	9,38

Figure 35 : RMSE (en millions de dollars) de l'arbre CART élagué appliqué aux bases d'apprentissage et de test

[76] L'index des différents acronymes utilisés dans l'arbre est présenté dans l'annexe D.

L'algorithme CART permet aussi de classer les variables explicatives utilisées dans la construction de l'arbre par ordre d'importance. Les valeurs d'importance relatives des variables varient de 0 % à 100 %. Plus l'importance relative d'une variable est grande, plus elle est considérée importante dans la construction de l'arbre.

Le tableau ci-après met en évidence l'importance des variables « Basel Business Line - Level 2 », « Event Risk Category » et « Secteur », en accord avec les résultats de la régression Stepwise Forward. Nous observons aussi que la variable « Taille Entreprise », sélectionnée par la procédure Stepwise Forward, est également importante dans la construction de l'arbre. Néanmoins, selon l'algorithme CART, les deux variables « Region of Incident » et «Region of Domicile » sont plus importantes en termes de modélisation que la variable « Taille Entreprise ». Ce résultat diffère de la sélection effectuée par la régression Stepwise Forward, qui n'a pas retenu les deux variables relatives aux régions géographiques pour expliquer les pertes des sinistres cyber attritionnels.

Variable	Importance relative
Basel Business Line - Level 2	40,75%
Event Risk Category	20,70%
Secteur	12,64%
Region of Incident	11,07%
Region of Domicile	8,02%
Taille Entreprise	6,82%

Figure 36 : Classement des six variables explicatives par importance dans la construction de l'arbre

Il est à noter que les deux variables « Basel Business Line - Level 2 » et « Event Risk Category » sont les plus importantes selon l'algorithme CART. Il serait donc intéressant, de construire des modèles GLM de Weibull et gamma uniquement avec ces deux variables. Ces modèles serviront de point d'appui aux comparaisons avec les autres modèles déjà construits.

11.6 Conclusion

Le tableau suivant récapitule les performances des différents modèles GLM et arbres CART, testés sur la base de test. Les niveaux des RMSE sont globalement proches et relativement faibles. Cependant, **c'est le modèle GLM gamma, utilisant les deux variables explicatives considérées comme les plus importantes par CART, qui a la plus petite RMSE de 9,2 millions de dollars soit 0,61% de la charge totale des sinistres cyber de la base de test. Ce modèle est donc le plus performant et sera retenu.** Ainsi, la combinaison des deux approches GLM et CART dans notre étude, améliore le pouvoir prédictif du modèle de la charge du sinistre cyber attritionnel.

	Weibull			**Gamma**			**CART**
Modèle	Full model	Model Stepwise Forward	Modèle avec choix des variables par CART	Full model	Model Stepwise Forward	Modèle avec choix des variables par CART	
RMSE base test	9,53	9,39	9,32	9,60	9,32	**9,20**	9,38

Tableau 34 : RMSE (en millions de dollars), des différents modèles GLM et CART appliqués à la base de test

Nous souhaitons comprendre la source d'erreur du modèle GLM gamma retenu et détecter sur quels événements le modèle s'est le plus trompé. Nous commençons par analyser les sinistres dont le coût prévu par le modèle est surestimé par rapport au coût réellement observé. Le total de ces écarts entre les valeurs observées et les valeurs prédites surestimées par le modèle est de 323,81 millions de dollars. Si nous examinons de plus près les cas ayant les plus grands écarts, nous constatons qu'il s'agit de sinistres dont la variable « Event Risk Category » correspondante prend la valeur « Others ». Le modèle était donc incapable de bien prédire le coût et surestimait sa valeur quand la classification de l'événement n'est pas spécifiée, ce qui montre encore une fois l'importance de cette variable dans la prédiction du coût du sinistre cyber. Les caractéristiques des incidents surestimés ayant les plus grands écarts sont données dans le tableau ci-après.

Basel Business Line - Level 2	**Event Risk Category**	**Region of Domicile**	**Region of Incident**	**Taille Entreprise**	**Secteur**
Retail Trade	Others	North America	North America	2	Non-FS
External Clients	Others	North America	North America	2	FS
External Clients	Others	North America	North America	3	FS
Information	Others	Europe	North America	3	Non-FS

Tableau 35 : Caractéristiques des incidents surestimés par le modèle ayant les plus grands écarts (en négatif)

D'autre part, le total des écarts positifs est de 1002,72 millions de dollars. Un écart positif veut dire que la valeur prédite par le modèle est plus petite que la valeur réelle, autrement dit le modèle sous-estime le coût du sinistre. Si nous examinons les cas des incidents sous-estimés par le modèle, ayant les plus grands écarts, nous constatons qu'il s'agit de sinistres avec des coûts supérieurs à 20 millions de dollars. Ils sont 28 sinistres, avec un écart total de 684,46 millions de dollars, soit 68 % du total des écarts positifs. Ce sont donc des cas extrêmes et isolés, que le modèle n'a pas pu capter. Pour pallier à ce problème, une solution envisageable est de réduire un peu le seuil d'écrêtement des charges ou de prévoir un traitement particulier voire par jugement d'expert si le sinistre est identifié comme un cas extrême non adapté au modèle. Le reste des cas sous-estimés par le modèle, correspondent à des sinistres dont au moins une des variables « Basel Business Line- Level 2 », « Event Risk Category » n'est pas

spécifiée (valeur égale à « Others »). Ce qui montre encore une fois l'importance de la précision de ces variables dans la prédiction du coût du sinistre.

Maintenant, que nous avons construit un modèle combiné pour modéliser la charge des sinistres cyber : un modèle GPD pour les sinistres extrêmes et un modèle prédictif à variables explicatives pour les sinistres attritionnels, nous nous intéressons à la modélisation de la probabilité que le sinistre cyber soit grave ou non. Pour ce faire, nous étudierons d'abord l'option d'utiliser un modèle GLM logistique puis celle de l'arbre de classification CART.

Chapitre 12 : Modélisation de la probabilité que le sinistre cyber soit grave ou pas

Un sinistre cyber dont le coût est supérieur au seuil de 50 millions de dollars établi auparavant est considéré comme extrême. Ainsi, nous disposons maintenant de réalisations binaires qui classifient les sinistres cyber en tant qu'extrêmes ou non. Notre objectif est de modéliser la probabilité qu'un sinistre cyber soit grave en fonction des variables explicatives disponibles. Cela permettra en combinaison avec la modélisation de la charge attritionnelle et extrême, d'en déduire une modélisation de la charge totale. Cette section présente les résultats du modèle de régression logistique et de l'arbre de classification CART.

12.1 Modèle de régression logistique

Nous débutons en utilisant le GLM logistique pour prédire la probabilité qu'un sinistre cyber soit extrême ou non, reposant sur la fonction de lien logit. Les modèles sont construits sur la base d'apprentissage et testés sur la base de test, à l'instar de la modélisation attritionnelle. Les variables explicatives sélectionnées au préalable pour la modélisation des pertes des sinistres cyber, sont également employées pour cette modélisation, à savoir :

- Basel Business Line - Level 2
- Event Risk Category
- Region of Incident
- Region of Domicile
- Taille entreprise
- Secteur

Nous implémentons d'abord un modèle complet, le « Full Model », qui inclut les six variables explicatives. Ensuite, nous procédons à une régression Stepwise Forward à partir d'un modèle sans variables explicatives, le « Null model ». La régression Stepwise Forward est effectuée avec un code R, et les variables retenues par l'algorithme sont :

- « Basel Business Line - Level 2 »
- « Event Risk Category ».

Cependant, avant de pouvoir tester les modèles de régression logistique, il est nécessaire de choisir le seuil de classification. En effet, les résultats obtenus à partir d'une régression logistique se situent toujours entre 0 et 1. Si la valeur est proche de 0, la probabilité que l'événement arrive est faible, tandis que si la valeur est proche de 1, la probabilité est élevée. Un seuil de classification est donc utilisé pour mettre en correspondance les résultats de la régression logistique à une classification binaire. Par exemple, avec un seuil de classification de 0,5, les valeurs de la régression logistique supérieures à 0,5 sont classées comme « sinistre grave », et celles inférieures comme « sinistre non grave ».

Il est à noter que le seuil de classification peut être augmenté ou diminué selon des critères probabilistes. En fonction du déséquilibre observé dans la répartition entre le nombre d'observations prenant la valeur 1 ou 0, le seuil de 0,5 peut être inapproprié.

Un bon moyen pour déterminer le seuil de classification optimal, consiste à tracer la courbe de la métrique F1-score, présentée dans la section 8.9, en fonction de différentes valeurs de seuil. Nous rappelons que le F1-score évalue la capacité d'un modèle de classification à prédire efficacement les classes positives, en faisant un compromis entre la précision (precision) et le rappel (recall)[77]. Nous choisissons de travailler avec le F1-score car il est plus robuste en présence des données déséquilibrées, comme c'est le cas ici avec une proportion faible de sinistres graves comparativement aux sinistres attritionnels (11% vs 89 %). Le seuil de classification optimal est celui ayant le F1-score le plus élevé, qui est dans ce cas, de **0,157.**

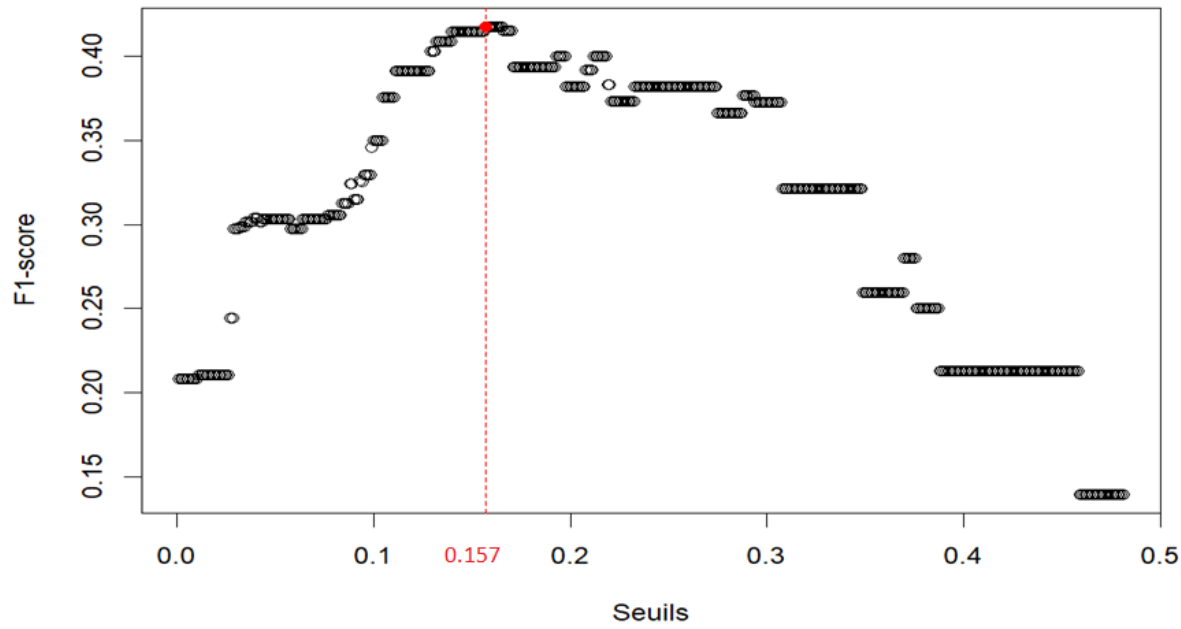

Figure 37 : F1-score en fonction des différentes valeurs du seuil de classification de la régression logistique

Nous testons maintenons les deux modèles de régressions logistiques : le modèle complet (Full Model) et le modèle avec la sélection des variables par la procédure Stepwise Forwrad (Modèle Stepwise Forward), sur la base de test. Pour mesurer la performance des classifications, nous examinons principalement l'accuracy et le F1-score[78]. Il est important de rappeler que plus ces deux métriques sont élevées, meilleure est la qualité prédictive du modèle. Les valeurs des différentes métriques des deux modèles GLM logistique appliqués à la base de test, sont présentées dans le tableau ci-après.

[77] Voir la section 8.9.
[78] Les deux métriques sont présentées dans la section 9.6.

Modèle	Full model	Modèle Stepwise Forward
Accuracy	0,768	0,771
Precision	0,29	0,30
Recall	0,65	0,70
F1-score	0,40	0,42

Tableau 36 : Métriques de performance des modèles GLM logistique appliqués à la base de test

Globalement, les niveaux des métriques de performances des deux modèles appliqués aux données de test sont très proches. Nous observons cependant, que la sélection des variables explicatives par la régression Stepwise Forward améliore légèrement le pouvoir prédictif par rapport au modèle complet, avec des valeurs d'accuracy et F1-score plus grandes.

Le modèle Stepwise Forward a une accuracy de 77,13%, ce qui signifie que 77,13% des prédictions du modèle sont correctes. Cependant, le F1-score associé est de 0,42, indiquant qu'il existe des erreurs dans les prédictions positives du modèle, autrement dit, le modèle a une tendance à prédire plus de sinistres graves comparé à ce qu'il en est en réalité.

Le modèle est à un fort rappel (70%) ce qui réduit les risques de ne pas détecter des sinistres cyber extrêmes. Inversement, la précision du modèle est faible (30%), ce qui augmente le risque de faux sinistres cyber extrêmes. Dans le contexte du risque cyber, ne pas détecter un risque extrême peut avoir des conséquences sérieuses. En effet, l'amplitude de la charge au-delà du seuil pourrait fortement induire une très importante charge de sinistres. Un fort rappel est donc plus adéquat.

12.2 Modèle CART de classification

Dans cette section, les résultats de l'application des arbres de classification CART aux sinistres cyber sont présentés, utilisant le langage R. Tout comme pour les GLM, les modèles CART sont construits sur la base d'apprentissage et validés sur la base de test.

12.2.1 Construction de l'arbre maximal

L'arbre maximal est le plus grand arbre que l'on peut former à partir de la base d'apprentissage et qui ne peut plus être segmenté. Il est construit sous R en utilisant le package « rpart », puis optimisé.

12.2.2 Elagage de l'arbre maximal

Nous cherchons à obtenir un arbre plus concis à partir de l'arbre maximal. La valeur du paramètre de complexité c_p qui minimise l'erreur de validation croisée est de 0,004016064. Elle est ensuite utilisée comme règle d'arrêt pour construite le nouvel arbre élagué dans la figure ci-après.

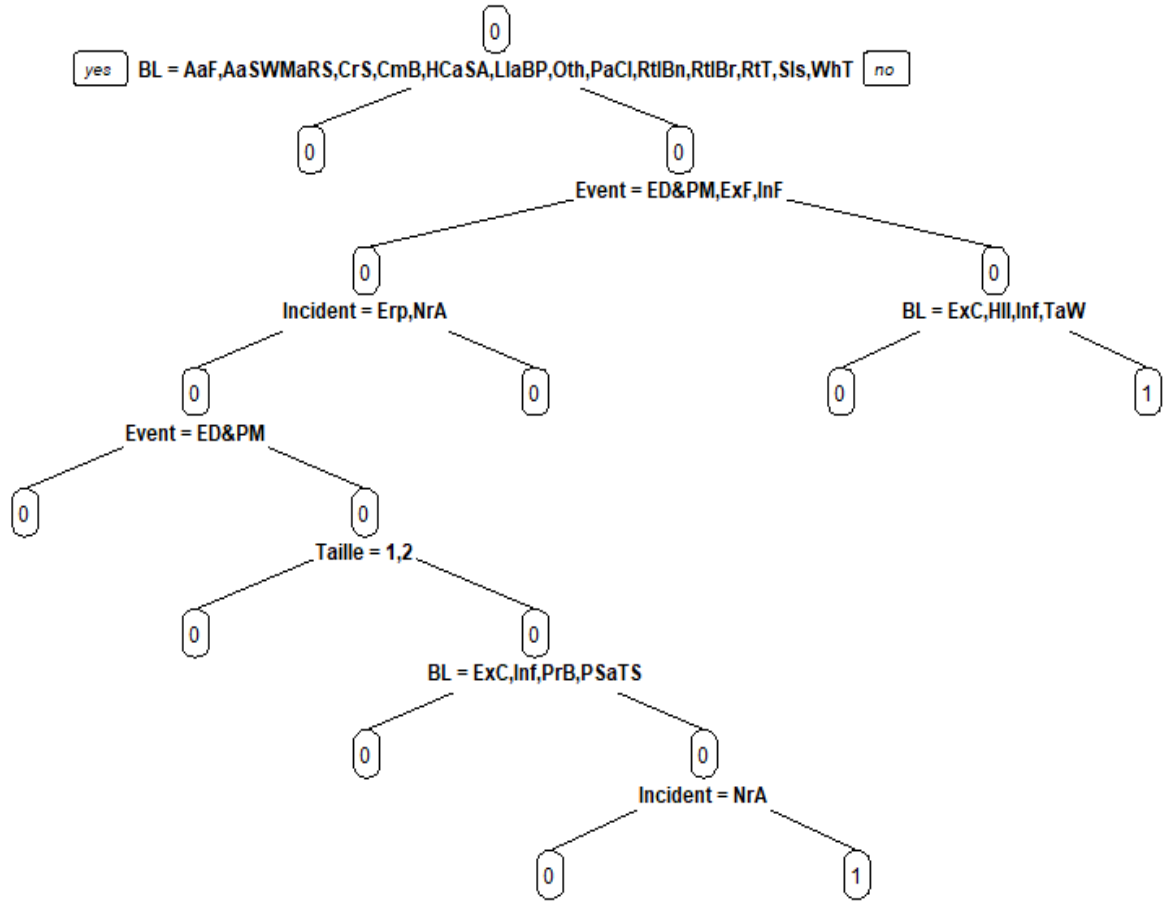

Figure 38 : Arbre CART élagué sur la base d'apprentissage[79]

L'arbre élagué reste identique à l'arbre maximal. Nous observons notamment l'importance de la variable du business line « Basel Business Line-Level 2», et de la variable de la classe d'événement opérationnel « Event Risk Category ».

Nous appliquons l'arbre de classification aux données d'apprentissage et de test, et calculons les métriques de performance. Les résultats sont donnés dans le tableau suivant :

Modèle	Base d'apprentissage	Base de test
Accuracy	0,90	0,88
Precision	0,57	0,45
Recall	0,16	0,13
F1-score	0,25	0,20

Tableau 37 : Métriques de performance de l'arbre de classification appliqué aux bases d'apprentissage et de test

[79] L'index des différents acronymes utilisés dans l'arbre est présenté dans l'annexe D.

Globalement, les niveaux des métriques de performance de l'arbre de classification appliqué aux deux bases sont proches, indiquant une stabilité de la performance du modèle quel que soit la base et un faible risque de sur-apprentissage.

Le pouvoir prédictif de l'arbre de classification proposé par CART est plus faible que celui du modèle de régression logistique, si nous nous basons sur le F1-score (0,2 vs 0,42). Certes l'accuracy du CART est plus grande que celle du GLM logistique (0,88 vs 0,77), mais étant donné que les classes dans la base de données sont déséquilibrées, le F1-score est à privilégier.

L'algorithme CART permet aussi de classer les variables explicatives utilisées dans la construction de l'arbre par ordre d'importance. Les valeurs d'importance relatives des variables varient de 0 % à 100 %. Plus l'importance relative d'une variable est grande, plus elle est importante.

Le tableau suivant met en lumière l'importance de la variable « Basel Business Line - Level 2 » en accord avec les résultats de la régression Stepwise Forward. Nous retrouvons la variable « Event Risk Category », sélectionnée aussi par la procédure Stepwise Forward. Néanmoins, selon l'algorithme CART, les variables « Secteur » et « Region of Incident » sont considérées comme plus importantes en termes de modélisation que la variable « Event Risk Category ». De plus, contrairement à l'algorithme CART, la régression Stepwise Forward, n'a pas sélectionné les variables des régions géographiques ni la taille de l'entreprise, ni le secteur pour expliquer la probabilité de classification des sinistres cyber.

Variable	**Importance relative**
Basel Business Line - Level 2	57,07%
Secteur	13,20%
Region of Incident	10,31%
Event Risk Category	8,73%
Region of Domicile	7,26%
Taille Entreprise	3,43%

Tableau 38 : Classement des six variables explicatives par importance dans la construction de l'arbre de classification CART

Il est à noter que l'algorithme CART, utilise les quatre variables « Basel Business Line - Level 2 », et « Event Risk Category », « Region of Incident » et « Taille Entreprise » dans la construction de l'arbre élagué. Il serait donc intéressant, de construire un modèle de régression logistique uniquement avec ces variables. Ce modèle servira de point d'appui aux comparaisons avec les autres modèles déjà construits.

12.3 Conclusion

Le tableau suivant résume les performances des différents modèles GLM logistique et de l'arbre de classification CART, testés sur la base de test.

Les métriques de performance des modèles GLM logistique sur la base de test sont très proches, avec une meilleure performance du modèle de la régression Stepwise Forward. Certes, l'arbre CART

présente une accuracy plus grande (0,88) mais un F1-score plus petit (0,2). Compte tenu du déséquilibre des classes dans la base de données, le F1-score est à privilégier par rapport à l'accuracy.

Encore une fois, la combinaison des deux approches GLM et CART améliore le pouvoir prédictif du modèle de la probabilité du sinistre cyber par rapport à l'utilisation de l'algorithme CART seul sur la base de test.

Parmi tous les modèles construits, le modèle de régression logistique de Stepwise Forward présente la meilleure performance sur la base de test, avec le plus grand F1-score (0,42). De plus, il affiche le plus grand rappel (0,7) ce qui est plus souhaitable dans le contexte du risque cyber. C'est donc le modèle que nous retiendrons.

	Modèle de régression logistique			**CART**
Modèle	Full model	Modèle Stepwise Forward	Modèle avec choix des variables par CART	
Accuracy	0,77	0,77	0,77	**0,88**
Precision	0,29	0,30	0,28	0,45
Recall	0,65	0,70	0,65	0,13
F1-score	0,40	**0,42**	0,39	0,20

Tableau 39 : Métriques de performance des modèles GLM logistique et arbre de classification CART appliqués à la base de test

Nous souhaitons détecter sur quels événements le modèle de régression logistique retenu se trompe. La matrice de confusion y correspondante retenu est donnée ci-après :

		Réalité	
		0 : Pas grave	1 : Grave
Prédiction	0 : Pas grave	TN = 235	FN = 12
	1 : Grave	FP = 66	TP = 28

Tableau 40 : Matrice de confusion du modèle de régression logistique retenu

Sur l'ensemble des 341 observations de la base de test, le modèle se trompe dans la classification de 78 sinistres, soit 22,87 % du total des observations, dont 66 sinistres non graves classés comme graves par le modèle. Ces cas faux positifs, correspondent à des sinistres dont la classification de l'événement « Event Risk category » est égale à « Others » (8 cas), ou les deux champs « Event Risk Category » et la ligne métier « Basel Business Line Level 2 » prennent la valeur « Others » (7 cas), ou même des sinistres dont ces champs sont connus (51 cas). Nous expliquons la source d'erreur pour ces derniers cas par la classification « Event Risk Category » qui n'est pas assez discriminante. Les modalités sont très larges et peuvent contenir des types de sinistres variés et qui n'ont pas nécessairement le même

profil de risque. La modalité « External Fraud » (fraude externe) par exemple, peut inclure des incidents comme des vols, l'hameçonnage, l'accès à des données à caractère personnel, les attaques d'ingénierie sociale, le ransomware et extorsion,..., des événements de types très différents et qui peuvent générer des coûts très variés.

Les 12 autres cas d'erreur sont des faux négatifs, c'est-à-dire que le modèle les prévoit comme sinistres graves alors qu'ils ne le sont pas en réalité. 5 cas parmi les faux négatifs ont l'un des deux champs « Basel Business Line – Level 2 » et « Event Risk Category » égal à la valeur « Others », ce qui montre l'importance de spécification de ces champs dans la détermination de la classification du sinistre. Pour les autres 7 cas des faux négatifs, malgré que les deux champs soient spécifiés, le modèle se trompe dans leur classification. Nous expliquons ces erreurs encore une fois par les catégories de la variable « Event Risk Category » très générales. Pour une meilleure gestion des risques, il est important d'adapter cette taxonomie des sinistres opérationnels très générale et d'opter pour une classification des sinistres plus fine et tenant compte des spécificités du risque cyber.

Chapitre 13 : Analyse des résultats et limites du modèle

13.1 Modèle de la charge totale d'un incident cyber

La distribution des charges des sinistres cyber peut être modélisée par deux distributions raccordées au seuil de 50 millions de dollars. Les charges extrêmes sont modélisées par la loi GPD et les charges attritionnelles sont quantifiées par un modèle prédictif en fonction de variables explicatives. Le modèle de la charge d'un sinistre cyber d'une entreprise proposé est ainsi la combinaison des meilleurs modèles testés auparavant :

- Le modèle GLM gamma, utilisant les deux variables explicatives « Basel Business Line - Level 2 » et « Event Risk Category », considérées comme les plus importantes par CART, prédictif de la charge attritionnelle ;

- La loi GPD calibrée aux excès de la charge au-delà du seuil des extrêmes de 50 millions de dollars, de paramètres de forme $\boldsymbol{\xi = 0,58}$ et d'échelle $\boldsymbol{\beta = 91,89}$;

- Le modèle de la régression logistique Stepwise Forward de la probabilité du risque extrême, utilisant les deux variables explicatives « Basel Business Line - Level 2 » et « Event Risk Category ».

La formule finale de prédiction du coût total d'un sinistre cyber, en millions de dollars, pour une entreprise *Y* de caractéristiques données dans le vecteur *X* peut être exprimée comme suit :

$$\boldsymbol{E[coût\ total\ du\ sinistre\ cyber/\ X]} = (\mathbf{1} - \boldsymbol{p_X}) \times \boldsymbol{E[coût\ attritionnel\ du\ sinistre\ cyber/\ X]} + \boldsymbol{p_X} \times (\mathbf{50}\boldsymbol{M}\$ + \boldsymbol{GPD_{\xi,\beta}})$$

Avec :
$\boldsymbol{X}$: est le vecteur des valeurs des variables explicatives pour l'entreprise $\boldsymbol{Y}$;
$\boldsymbol{p_X}$: est la probabilité que le sinistre cyber soit extrême sachant $\boldsymbol{X}$ pour l'entreprise $\boldsymbol{Y}$;
$\boldsymbol{E[coût\ attritionnel\ du\ sinistre\ cyber/\ X]}$: est la valeur du coût du sinistre attritionnel prédite par le GLM gamma sachant $\boldsymbol{X}$ pour l'entreprise $\boldsymbol{Y}$;
$\boldsymbol{GPD_{\xi,\beta}}$: est la valeur de l'excès de dépassement du coût du seuil des extrêmes de 50 millions de dollars donnée par la loi GPD estimée de paramètres de forme $\boldsymbol{\xi = 0,58}$ et d'échelle $\boldsymbol{\beta = 91,89}$. Cette valeur peut être la moyenne de la loi calibrée ou déterminée par l'expertise métier.

Ce modèle permettant de quantifier les pertes liées au risque cyber d'une organisation, peut être utilisé pour la détermination de la prime pure. Pour autant, il est à souligner que ce jeu de données ne permet pas de modéliser la fréquence de survenance des incidents cyber. Le modèle de la probabilité que le sinistre cyber soit extrême ou non peut aussi servir d'indicateur précieux au moment de la souscription d'une garantie cyber par exemple.

Le modèle de la charge construit dépend de la ligne métier (Business line) de l'organisation et de la classification des événements opérationnels selon le régime bâlois. Ainsi, grâce au modèle construit, il est possible de proposer diverses garanties selon la classification de l'éventuel sinistre cyber pour des organisations opérant dans tout domaine d'activité. Parmi les garanties possibles, nous mentionnons la garantie couvrant les défaillances de systèmes informatiques, que leur origine soit accidentelle ou criminelle, les fraudes externes comme l'hameçonnage, les fraudes internes comme la manipulation des comptes financières ou l'accès non autorisé à des informations à caractère personnel…

Le modèle permet de modéliser la charge des sinistres cyber en fonction de la garantie souscrite ainsi que de la ligne métier de l'organisation. Comme nous avons vu précédemment dans les sections 11.6 et 12.3, pour une meilleure gestion des risques, il est important d'adapter la taxonomie des sinistres opérationnels « Event Risk Category » très générale et d'opter pour une classification des sinistres plus fine et tenant compte des spécificités du risque cyber. A noter aussi, que les autres variables explicatives testées auparavant peuvent être gardés dans la modélisation finale de la charge du sinistre cyber, notamment que les niveaux de performance des modèles avec l'ensemble des variables explicatives et ceux avec uniquement les deux variables retenues par la régression Stepwise Forward, sur la base de test, sont proches. De plus, l'utilisation des variables du secteur, de la taille de l'entreprise et des régions géographiques permettra une segmentation plus fine de la modélisation de la charge proposée selon le profil de l'entreprise et une diversification dans le portefeuille de l'assureur. Un autre point à tenir en compte, est le seuil des extrêmes de 50 millions de dollars qui reste élevé, ce qui compromet parfois les modèles de régression à bien capter des cas extrêmes.

13.2. Limites et approfondissement du modèle

Dans cette dernière section, un regard critique est porté sur le modèle de la charge des sinistres cyber construit, et des extensions sont proposées pour l'affiner.

Il est à noter que la formule finale de prédiction du coût total d'un sinistre cyber proposée permet de synthétiser les différentes modélisations mises en place dans le cadre de cette étude mais n'a pas été testée en tant que telle et repose notamment sur sa partie de dépassement du seuil sur un niveau fixe de sinistralité, ce qui en constitue une limite forte. De plus, il est à noter que ce modèle, compte tenu des données en base, ne permet pas de modéliser la fréquence de survenance des sinistres cyber.

Les arbres de régression et de classification ont l'inconvénient d'être relativement instables. En effet, de légers changements peuvent influencer la structure de l'arbre. Tester des méthodes d'agrégation, pour augmenter la robustesse, peut être utile. Enfin, les réseaux de neurones, bien que complexes conceptuellement, peuvent également fournir de très bons résultats. Ces méthodes, ont été volontairement écartées du périmètre de l'étude pour se concentrer sur les arbres CART, puisqu'elles offrent une vision synthétique de la base de données et leurs interprétations aisées peuvent enrichir la réflexion autour de la stratégie tarifaire et de souscription des compagnies d'assurance.

Il est essentiel de souligner que les variables utilisées pour expliquer la charge des sinistres cyber dans cette étude ne sont pas directement liées au risque cyber de l'entreprise. Certes, l'utilisation des données

opérationnelles offre une piste de segmentation du processus de modélisation en vue de réaliser un véritable positionnement tarifaire, cependant la base de données est d'abord créée non spécifiquement pour les risques cyber, mais pour l'ensemble des risques opérationnels. Il a été mis en évidence la pertinence de la ligne métier et de la classification bâloise des événements opérationnels dans l'explication des charges des sinistres cyber. Néanmoins, la modélisation peut être améliorée, en incorporant des variables explicatives liées au risque cyber de l'entreprise, telles que, par exemple, l'organisation de l'informatique et de la sécurité, les mesures de prévention et de protection contre les risques cyber, le niveau d'interconnexion avec d'autres entreprises, les procédures de réaction en cas de survenance d'un risque cyber, le niveau d'interdépendance des systèmes informatiques de l'entreprise, à l'échelle de l'entreprise, mais également nationale puis mondiale. En rappelant que le risque cyber peut être systémique, la contagion entre les entreprises en cas de survenance d'un sinistre cyber peut être aussi prise en compte dans la modélisation. La forte interconnexion entre les entreprises laisse, en effet, le risque de contagion très élevé.

Le nombre des données reste aussi assez limité pour assurer la robustesse du modèle, et induit par conséquent une forte volatilité dans la modélisation établie dans l'étude. Nous avons souligné l'importance de modéliser séparément les charges extrêmes et attritionnelles, pour une prise en compte plus précise de la distribution des charges liées au risque cyber. Cependant, il serait judicieux de déterminer deux seuils d'écrêtement différents pour les charges des sinistres cyber dans les secteurs financiers et non financiers. Si le secteur financier est parmi les plus touchés, le coût des incidents y est plus faible, potentiellement expliqué par des niveaux d'investissement plus élevés en cybersécurité. Une tentative a été faite pour déterminer un seuil pour chaque secteur, mais le nombre limité d'observations extrêmes par secteur a empêché la calibration de deux lois GPD aux charges graves. Cette séparation n'était pas réalisable dans le cadre de ce mémoire, mais elle pourrait être envisagée en présence de plus de données.

En ce qui concerne les charges extrêmes, le nombre des sinistres extrêmes dans la base de données étant faible, leur modélisation peut être biaisée. Une pratique très courante dans le marché est de confier la tarification des sinistres extrêmes à des experts métiers. Les assureurs peuvent aussi définir des franchises pour transférer les sinistres extrêmes aux réassureurs.

Une autre limite du modèle, est sa dépendance à la graine aléatoire (seed). La randomisation de la graine aléatoire pour l'établissement des bases d'apprentissage et de test pourrait aboutir peut être à des résultats différents. Un moyen de statuer sur cette limite serait de procéder à la réutilisation des méthodes en place en faisant varier aléatoirement la graine aléatoire et étudier la stabilité du modèle. Par contrainte de temps, cette approche n'a pas pu être mise en place dans ce cadre.

Les arbres de régression et de classification ont l'inconvénient d'être relativement instables. En effet, de légers changements peuvent influencer la structure de l'arbre. Tester des méthodes d'agrégation, pour augmenter la robustesse, peut être utile. Enfin, les réseaux de neurones, bien que complexes conceptuellement, peuvent également fournir de très bons résultats. Ces méthodes, ont été volontairement écartées du périmètre de l'étude pour se concentrer sur les arbres CART, puisqu'elles

offrent une vision synthétique de la base de données et leurs interprétations aisées peuvent enrichir la réflexion autour de la stratégie tarifaire et de souscription des compagnies d'assurance.

Il est à souligner que la base de données utilisée dans ce mémoire reste assez restreinte, présente un biais géographique étant donné que la majorité des risques ont eu lieu aux Etats-Unis, et que suivant les régions, le coût associé au risque cyber pourrait être très différent. De plus, la base de données n'inclut pas les pertes liées à la réputation. S'ajoute à cela, le fait que la nature des sinistres est très variées et peut parfois s'écouler sur des périodes de temps très diverses. Aussi, l'alimentation de certaines variables par SAS est parfois obscure. Par conséquent, pour mieux mesurer l'apport de la méthodologie établie ainsi que l'efficacité des résultats, l'approche utilisée dans ce mémoire doit être appliquée à une base de données plus grande, avec plus de variables explicatives.

Une autre limite que présente ce modèle est qu'il suppose que le risque est constant dans le temps. En effet, vu le nombre limité des sinistres cyber dans la base de données, ils ont été considérés dans leur intégralité pour réaliser la modélisation. Il est proposé de donner plus de poids aux données récentes, pour tenir compte du caractère évolutif du risque cyber, voire d'étudier son évolution pour mieux l'anticiper.

Conclusion générale

Au regard de la rareté des données et le faible recul dans le temps sur la sinistralité liée au risque cyber dans le portefeuille des assureurs, l'utilisation de données externes peut se révéler comme un bon recours pour modéliser ce risque ou tout du moins comme un support à sa modélisation.

Le choix d'utiliser des données opérationnelles devient évident lorsqu'on considère que le risque cyber est le risque opérationnel lié à l'information et aux systèmes informatiques. Travailler avec ces données, permet aussi d'élargir la portée de l'étude, englobant diverses formes de risque cyber plutôt que de se concentrer uniquement sur un type spécifique d'incident.

Disposant de données relatives aux pertes opérationnelles, subies par des entreprises à l'international, nous avons proposé dans ce mémoire une approche de modélisation des pertes liées aux sinistres cyber de cette base. L'identification des sinistres cyber dans la base SAS OpRisk Global data a été réalisée à l'aide d'un algorithme de fouille de texte appliqué aux descriptions de chaque sinistre.

Pour ce faire, nous avons combiné des méthodes de la théorie des valeurs extrêmes, des modèles linéaires généralisés (GLM) et des méthodes d'apprentissage statistique de type CART. Compte tenu de la très forte amplitude de la charge de sinistre observée, la modélisation des sinistres graves ne peut être écartée et nécessite une modélisation dédiée.

Une modélisation des montants de sinistres cyber par deux distributions liées par un certain seuil de qualification comme sinistralité grave est proposée. Le choix du seuil d'écrêtement des charges peut être subjectif, mais il doit respecter un équilibre entre la variance et le biais. Il a été difficile d'obtenir un seuil précis à partir des méthodes graphiques de la théorie des valeurs extrêmes. Une analyse de sensibilité a été réalisée, choisissant le seuil avec la plus petite erreur standard du paramètre de forme estimé de la loi de Pareto Généralisée. Par la suite, nous avons calibré une loi GPD sur les excès des charges au-dessus du seuil défini.

Pour ce qui est de la modélisation de la charge attritionnelle, nous avons utilisé une base d'apprentissage pour la modélisation que l'on a ensuite appliquée à la base de test. La modélisation elle-même a reposé sur des méthodes que sont les GLM et les arbres de régression. Leur utilisation a nécessité de procéder à une sélection des variables explicatives.

Nous avons discuté du choix des variables explicatives de la charge. La revue de la littérature et les constats observés du monde réel ont guidé la sélection initiale des variables explicatives. Les tests statistiques ont affiné cette première sélection, ne conservant que les variables les plus pertinentes. Nous avons montré l'utilité de regrouper les modalités non significatives des variables explicatives catégorielles, car les laisser aurait conduit à des modèles trop lourds.

Les charges attritionnelles ont été d'abord modélisées à l'aide des GLM avec les lois de Weibull et gamma. Des modèles complets, comprenant toutes les variables explicatives présélectionnées, et des

modèles de régression stepwise forward ont été construits. Pour remédier aux limites des GLM et faire ressortir plus d'informations sur le risque cyber, nous avons mis en œuvre les arbres de décision CART. Etant donné que l'algorithme CART permet de classer les variables explicatives par ordre d'importance, des modèles GLM avec les variables choisies par CART ont également été construits.

Nous avons proposé de modéliser la probabilité qu'un sinistre cyber soit grave ou non, ce qui peut être très utile pour les souscripteurs et les actuaires travaillant sur la cyber-assurance. Les deux approches de régression et d'arbres de classification CART ont été utilisées. De plus, un modèle de régression logistique avec les variables utilisées par l'algorithme CART a été construit.

C'est sur la base des données de test que les modèles ont été comparés, au sens de l'erreur quadratique moyenne (RMSE) pour les modèles des charges attritionnelles et le F1-score pour les modèles prédisant la probabilité qu'un sinistre cyber soit extrême ou non. Les modèles GLM construits avec les variables explicatives sélectionnées par CART ont présenté les meilleurs résultats pour la charge des sinistres attritionnels. La combinaison des deux approches GLM et CART s'est avéré cruciale. Il est essentiel de noter que les conclusions tirées sont indubitablement liées à la base de données. Nous ne pouvons affirmer que coupler les GLM et les méthodes d'apprentissage est toujours meilleur.

Nous avons donc essayé de proposer une méthodologie de modélisation des charges des sinistres cyber, et tenté plusieurs modélisations pour voir la meilleure. Cette première approche, a permis à Deloitte de mieux s'approprier le risque cyber encore peu connu, et construire un modèle avec des données externes afin d'aider les organisations à construire et implémenter les solutions de la cyber-assurance. La modélisation proposée peut être aussi utile pour tarifier des couvertures cyber en fonction de la garantie souscrite ainsi que de la ligne métier de l'organisation.

Pour conclure, nous avons souligné l'importance de modéliser séparément les charges extrêmes et attritionnelles. Cette approche permet une prise en compte plus précise de la distribution des charges liées au risque cyber et en particulier les données utilisées ici qui présentaient une forte asymétrie sur les valeurs extrêmes de la distribution. Cependant, il serait judicieux de déterminer deux seuils d'écrêtement différents pour les charges des sinistres cyber dans les secteurs financier et non financier. Si le secteur financier est parmi les plus touchés, le coût des incidents y est plus faible, potentiellement expliqué par des niveaux d'investissement plus élevés en cybersécurité. Une tentative a été faite pour déterminer un seuil pour chaque secteur, mais le nombre limité d'observations extrêmes par secteur a empêché la calibration de deux lois GPD aux charges graves. Cette séparation n'était pas réalisable dans le cadre de ce mémoire, mais elle pourrait être envisagée en présence de plus de données.

Néanmoins, il est important de noter que la base de données utilisée dans ce mémoire reste assez restreinte, présente un biais géographique, la majorité des risques ayant eu lieu aux Etats-Unis et n'inclut pas les pertes liées à la réputation qui ne sont pas négligeables sur le véritable coût total d'un incident cyber. Par conséquent, pour mieux mesurer l'apport de la méthodologie établie ainsi que l'efficacité des résultats, l'approche utilisée dans ce mémoire doit être appliquée à une base de données plus grande et créée spécifiquement pour le risque cyber.

De plus, la base de données utilisée ne nous permet pas de modéliser les fréquences de survenance des sinistres cyber. En rappelant que le risque cyber peut être systémique, l'hypothèse d'indépendance des fréquences et sévérités des sinistres cyber n'est pas toujours vérifiée. Des approches alternatives de tarification sont la méthode Probabilité x Charge et les modèles par scénario.

De plus en ouverture, il est essentiel de souligner que les variables explicatives utilisées pour expliquer la charge des sinistres cyber dans cette étude ne sont pas directement liées au risque cyber de l'entreprise. Il a été mis en évidence la pertinence de la ligne métier et de la classification bâloise des événements opérationnels dans l'explication des charges des sinistres cyber. Certes, l'utilisation des données opérationnelles offre une piste de segmentation du processus de tarification, mais il est important d'adapter la taxonomie des sinistres opérationnels très générale et d'opter pour une classification des sinistres plus fine et tenant compte des spécificités du risque cyber. La modélisation peut être aussi améliorée, en incorporant des variables explicatives liées au risque cyber de l'entreprise, telles que l'organisation de l'informatique et de la sécurité, les mesures de prévention et de protection contre les risques cyber, le niveau d'interconnexion avec d'autres entreprises, les procédures de réaction en cas de survenance d'un risque cyber, le niveau d'interdépendance des systèmes informatiques de l'entreprise, à l'échelle de l'entreprise, mais également nationale puis mondiale.

Enfin, il convient de rappeler l'importance du partage des informations sur la sinistralité des contrats cyber entre les assureurs à l'échelle nationale et européenne, pour enrichir la connaissance des experts métiers sur le risque cyber et développer le marché de la cyber-assurance.

Bibliographie

Aldasoro I., Frost J., Gambacorta L. & Whyte D., *« Covid-19 and cyber risk in the financial sector »*, BIS Bulletin N°37. URL : https://www.bis.org/publ/bisbull37.pdf.

Allianz Global Corporate & Specialty (2023), *« Baromètre des risques Allianz 2023 »*, URL : https://newsroom.allianz.fr/actualites/barometre-des-risques-allianz-2023-a554-98cdb.html.

AMRAE (l'Association pour le Management des Risques et des Assurances de l'Entreprise) (2023), *« LUCY : LUmière sur la CYberassurance »*. URL : https://www.amrae.fr/bibliotheque-de-amrae/lucy-lumiere-sur-la-cyberassurance-amrae-mai-2023.

AMRAE (l'Association pour le Management des Risques et des Assurances de l'Entreprise) (2022), *« LUCY : LUmière sur la CYberassurance »*. URL : https://www.amrae.fr/bibliotheque-de-amrae/lucy-lumiere-sur-la-cyberassurance-amrae-juin-2022.

Aon (2017), rapport *« Global Risk Management Survey »*. URL : https://www.aon.com/2017-global-risk-management-survey/pdfs/2017-Aon-Global-Risk-Management-Survey-Full-Report-062617.pdf.

Asli M., *« Risque opérationnel bancaire : le point sur la réglementation prudentielle »*, Management & Avenir 2011/8 (n° 48), pages 225 à 238. URL : https://www.cairn.info/revue-management-et-avenir-2011-8-page-225.htm#:~:text=La%20d%C3%A9finition%20inclut%20%C3%A9galement%20le,et%20les%20risques%20de%20r%C3%A9putation.

Atlas Magazine (2022), *« Le marché mondial de la cyberassurance »*. URL : https://www.atlas-mag.net/category/tags/focus/le-marche-mondial-de-la-cyberassurance.

Baradel N., cours « Assurance dommage » (2022) de l'ENSAE.

Basel Committee on Banking Supervision (2006), *« International Convergence of Capital Measurement and Capital Standards A Revised Framework Comprehensive Version »*. URL : https://www.bis.org/publ/bcbs128.htm.

Bastien L. (2023), *« Cyber-Tempête : après le Covid, Davos prédit la future crise mondiale »*. URL : https://www.lebigdata.fr/cyber-tempete-davos.

Biener C., Eling M. & Wirfs J. (2015), *« Insurability of Cyber Risk: An Empirical Analysis »*, The Institute of Insurance Economics at the University of St. Gallen, Kirchlistrasse 2, St. Gallen 9010,

Switzerland. URL : https://www.internationalinsurance.org/sites/default/files/2018-03/Insurability%20of%20Cyber%20Risk.pdf.

Breiman L. (1984), *« Classification and Regression Trees »*.

Cebula et Young (2010), *« A Taxonomy of Operational Cyber Security Risks »*, Software Engineering Institute Technical Note CMU/SEI-2010-TN-028.

CERT-FR (2023), Rapport Menaces et Incident du CERT-FR, *« Panorama de la cybermenace 2022 »*. URL : https://www.cert.ssi.gouv.fr/cti/CERTFR-2023-CTI-001/.

Chavez-Demoulin, Embrechts & Hofert (2015), *« An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates »*, American Risk and Insurance Association, 2015. URL : https://www.jstor.org/stable/43998282.

CNIL (Commission Nationale de l'Informatique et des Libertés), « *Une donnée à caractère personnel, c'est quoi ?* ». URL : https://www.cnil.fr/fr/cnil-direct/question/une-donnee-caractere-personnel-cest-quoi.

Coles S. (2001), « An Introduction to Statistical Modeling of Extreme Values ».

CRO Forum (2014), *« Cyber resilience - The cyber risk challenge and the role of insurance »*. URL : https://www.thecroforum.org/cyber-resilience-cyber-risk-challenge-role-insurance/.

CRO Forum (2023), *« Emerging Risks Initiative Major Trends and Emerging Risk Radar 2023 »*. URL : https://www.thecroforum.org/emerging-risk-initiative-major-trends-and-emerging-risk-radar-2023/.

Cybercover, *« CYBERCRIMINALITÉ : NOTPETYA OU L'HISTOIRE D'UN FAUX RANÇONGICIEL »*. URL : https://www.cyber-cover.fr/cyber-documentation/cyber-criminalite/cybercriminalite-notpetya-le-malware-a-10-milliards-de-dollars#:~:text=L'ancien%20conseiller%20de%20la,%C3%A0%2010%20milliards%20de%20dollars.

Deloitte (2023), *« Cyber Insurance Cyber-pricing innovation and underwriting solutions— keeping pace with evolving threat landscapes »*, Actuarial and Insurance Solutions.

Doerr S., Gambacorta L., Leach T., Legros B. & Whyte D. (2022), « *Cyber risk in central banking* », BIS Working Papers No 1039,Monetary and Economic Department. URL : https://www.bis.org/publ/work1039.pdf.

Eling M. & Jung K. (2022), « Heterogeneity in cyber loss severity and its impact on cyber risk measurement ». URL : https://www.researchgate.net/publication/361117286_Heterogeneity_in_cyber_loss_severity_and_its_impact_on_cyber_risk_measurement.

Eling M., Wirfs J.H. *« What are the actual costs of cyber risk events? »*, European Journal of Operational Research, Volume 272, Issue 3, Pages 1109-1119. URL : https://www.sciencedirect.com/science/article/abs/pii/S037722171830626X.

Elinga M., Wirfs J.H. (2015), *« Modelling and Management of Cyber Risk »*, Institute of Insurance Economics, University of St. Gallen, Rosenbergstrasse 22, 9000 St. Gallen, Switzerland. URL : https://www.actuaries.org/oslo2015/papers/iaals-wirfs&eling.pdf.

ENISA (The European Union Agency for Cybersecurity) (2022), *« ENISA Threat Landscape 2022 »*. URL : https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022.

ENISA (The European Union Agency for Cybersecurity) (2023), *« ENISA Threat Landscape 2023 »*. URL : https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023.

Federal Bureau of Investigation (2021), *« Internet Crime Report 2021»*. URL : https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf/.

Financial Staility Board (2023), *« Cyber Lexicon Updated in 2023 »*. URL : https://www.fsb.org/wp-content/uploads/P130423-3.pdf.

Gericke & Raubenheimer (2020), *« Using SAS OpRisk Global Data to Improve Decision-Making at a Bank »*, Paper 5069-2020, Centre for Business Mathematics and Informatics, North-West University, SAS Global Forum 2020. URL : https://support.sas.com/resources/papers/proceedings20/5069-2020.pdf.

GOV.UK (site Web d'information du secteur public du Royaume-Uni), *« Official Statistics Cyber Security Breaches Survey 2022 Updated 11 July 2022 »*. URL : https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2022/cyber-security-breaches-survey-2022.

Greenberg A. (2018), *« THE UNTOLD STORY OF NOTPETYA, THE MOST DEVASTATING CYBERATTACK IN HISTORY »*.

Guilleux Y. (2023), *« Taille d'entreprise : les catégories à connaître pour segmenter ses ventes »*, HubSpot. *URL : https://blog.hubspot.fr/sales/taille-entreprise-classification*.

Hamilton A. (2020), *« Sopra Steria ransomware attack costs group €50m »*. URL : https://www.fintechfutures.com/2020/11/sopra-steria-ransomware-attack-costs-group-e50m/.

IBM (International Business Machines Corporation) (2023), *« Cost of a Data Breach Report 2023 »*.URL : https://www.ibm.com/reports/data-breach#:~:text=The%20global%20average%20cost%20of,15%25%20increase%20over%203%20year

s.&text=51%25%20of%20organizations%20are%20planning,threat%20detection%20and%20response%20tools.

Insuramore (2022), *« Gross Direct Premiums Written for Cyber Insurance, 2022: Top 100 Insurer Groups Worldwide »*. URL : https://www.insuramore.com/rankings/insurers/premiums-cyber/.
Le club des juristes (2018), Rapport *« ASSURER LE RISQUE CYBER »*, Commission ad hoc Cyber Risk, TOME 1. URL : https://think-tank.leclubdesjuristes.com/wp-content/uploads/2018/01/cdj_assurer-le-risque-cyber_janvier_2018_fr-2.pdf.

Le Conseil fédéral, Le portail du Gouvernement suisse (2023), *« Rencontre annuelle 2023 du Forum économique mondial »*. Lien : https://www.admin.ch/gov/fr/accueil/documentation/on-en-parle/wef.html.

MARTINEZ A. (2019), *« Modélisation assurantielle du risque cyber »*, Mémoire présenté devant le Conservatoire National des Arts et Métiers pour l'obtention du diplôme de la filière Actuariat et l'admission à l'Institut des Actuaires le 3 juillet 2019. URL : https://www.institutdesactuaires.com/docs/mem/d39dd709e4ba0bb10e621651a7c8fbcb.pdf.

McCullagh P. et Nelder J.A.(1983), *« Generalized Linear Models »*.

MEURANT S. et CARDON.R, Sénateurs (2021), Rapport d'information N° 678 fait au nom de la délégation aux entreprises relatif à la cybersécurité des entreprises, page 24. URL : https://www.senat.fr/rap/r20-678/r20-6781.pdf.

Microsoft, Learn, (2023), *« Health Insurance Portability and Accountability Act (HIPAA) & Health Information Technology for Economic and Clinical Health (HITECH) Act »*. URL : https://learn.microsoft.com/fr-fr/compliance/regulatory/offering-hipaa-hitech.

Morgan S. (2022), « *Top 10 Cybersecurity Predictions And Statistics For 2023* », Cybercrime magazine. URL : https://cybersecurityventures.com/stats/.

OpinionWay pour CESIN (2021), *« Baromètre de la cyber-sécurité des entreprises, Vague 6 – Janvier 2021 »*,page 39. URL : https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/marketing/internet-et-ntic.html.

P.McCullagh & J.A.Nelder (1989), *« Generalized Linear Models »*, Chapman and Hall. URL : *https://kobia.fr/classification-metrics-matrice-de-confusion/*.

Pirra M. (2022), *« CYBER RISK MANAGEMENT »*, Department of Economics, Statistics and Finance "Giovanni Anania" Università della Calabria. URL : https://www.actuaries.org/IAA/Documents/SECTIONS/JointColloquium2022/Day3_AFIRERM_Pirra.pdf.

Powell J. & Rauline N. (2021), *« La cybercriminalité, principal risque pour l'économie »*, LesEchos. URL : https://www.lesechos.fr/monde/etats-unis/la-cybercriminalite-principal-risque-pour-leconomie-selon-le-patron-de-la-fed-1306023.

Robert C.Y. (2023), Cours ENSAE-IPP *« Extreme Value Theory »*.

Rydman M. (2018), *« Application of the Peaks-Over-Threshold Method on Insurance Data »*, Department of Mathematics Uppsala University. URL : https://www.diva-portal.org/smash/get/diva2:1231783/FULLTEXT01.pdf.
Seabird, *« Risque Cyber – de quoi parle-t-on ? Lexique »*. URL :https://www.seabirdconseil.com/nos-decryptages/le-cyber-risque-de-quoi-parle-t-on-lexique-seabird/.

Specops (2020), *« The countries experiencing the most 'significant' cyber-attacks »*. URL : https://specopssoft.com/blog/countries-experiencing-significant-cyber-attacks/.

Strupczewsk G. (2019), *« What Is the Worst Scenario? Modeling Extreme Cyber Losses »*, Cracow University of Economics. URL : https://ideas.repec.org/h/spr/prbchp/978-3-030-16045-6_10.html

Stuart C. (2001), « *An Introduction to Statistical Modeling of Extreme Values* ».

The White House (2021), « Executive Order on Improving the Nation's Cybersecurity ». URL : https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/#:~:text=The%20Federal%20Government%20must%20improve,requires%20more%20than%20government%20action.

Thompson D.R. (2023), *« Extreme Value Theory ('EVT') & Extreme Value At Risk ('EVaR') The Practical Implementation of EVT to calculate EVaR and the Targeting of Risk in the Digital Asset Derivatives Market »*, LibertyRoad. URL : https://libertyroadcapital.com/wp-content/uploads/2023/07/LRC_EVT_5Feb23.pdf.

Wikipédia, Risque opérationnel (établissement financier). URL : https://fr.wikipedia.org/wiki/Risque_op%C3%A9rationnel_(%C3%A9tablissement_financier).

Annexes

A. Présentation des variables de la base SAS OpRisk Global data

Variables	Notes	Description
Ref ID Code	Informations descriptives	Numéro d'identification de référence
Parent Name		Nom de la société mère de l'entreprise
Firm Name		Nom de l'entreprise qui a subi la perte opérationnelle
Description of Event	Pertes monétaires (millions de dollars)	Description du sinistre
Loss Amount ($M)		Montant brut de la perte opérationnelle (en millions de dollars)
Current Value of Loss		Le montant de la perte opérationnelle (en millions de dollars) ajusté du CPI pour tenir compte de l'inflation.
Basel Business Line - Level 1	Détail du Business Line	Business Line - niveau 1 (basé sur les normes BIS)
Basel Business Line - Level 2		Business Line - niveau 2 (basé sur les normes BIS)
Business Unit		Activité de la Business Unit
Event Risk Category	Détails de la catégorie d'événement	Catégorie d'événement principal (niveau 1)
Sub Risk Category		Catégorie d'événement principal (niveau 2)
Activity		Catégorie d'activité (niveau 3)
Country of Legal Entity	Localisation	Pays du siège social
Country of Incident		Pays de survenance du sinistre
First Year of Event	Dates clés	L'année au cours de laquelle la perte a eu lieu ou la première année en cas de plusieurs années
Last Year of Event		L'année au cours de laquelle la perte a eu lieu ou la dernière année en cas de plusieurs années
Month & Year of Settlement		E moi et l'année au cours de laquelle la perte a été réalisée - si disponible, sinon l'année de l'article
Industry Sector Code	Informations sur le secteur d'industrie/la région	Code à 2 chiffres du Système de classification des industries de l'Amérique du Nord
Industry Sector Name		Nom du secteur industriel correspondant à l'entreprise qui a subi la perte
Region of Domicile		Région géographique du siège social de l'entreprise

Financial Information (F or P)	Détail de la taille de l'entreprise (millions de dollars)	Identifie si les informations financières concernent l'entreprise (F) ou la société mère (P)
Year of Financial Information		L'année au cours de laquelle les informations financières ont été utilisées
Source of Financial Information		La source des informations financières
Revenue ($M)		Revenus de l'exercice au cours duquel la perte a été subie (si disponible)
Current Value of Revenue ($M)		Revenus de l'exercice au cours duquel la perte a été subie ajustés par le CPI (si disponible)
Assets ($M)		Le montant des actifs en millions de dollars ajusté de l'inflation via le CPI
Shareholder Equity ($M)		Capitaux propres de l'exercice au cours duquel la perte a été subie (si disponible)
Number of Employees		Nombre de salariés pour l'exercice au cours duquel la perte a été subie (si disponible)
Net Income ($M)		Bénéfice net de l'exercice au cours duquel il a subi la perte (si disponible)
Legal Liability ($M)	Détail des pertes (en millions de dollars)	Jugements, règlements et autres frais juridiques
Regulatory Action ($M)		Amendes ou paiement direct de toute autre pénalité
Loss or Damage to Assets ($M)		Réduction directe de la valeur des actifs physiques
Restitution ($M)		Paiements à des tiers (y compris des clients) au titre de pertes opérationnelles dont l'entreprise est légalement responsable
Loss of Recourse ($M)		Pertes subies lorsqu'un tiers ne respecte pas ses obligations envers l'entreprise et qui sont imputables à un événement opérationnel
Write-downs ($M)		Réduction directe de la valeur des actifs en raison d'un vol, d'une fraude, d'une activité non autorisée ou de pertes de marché ou de crédit résultant d'événements opérationnels
Reported Loss Amount (M) in Local Currency	Admin I	Montant de la perte opérationnelle en millions (en monnaie locale), tel que déclaré
Currency Conversion		Taux utilisé pour convertir les pertes en monnaie locale en dollar - à compter de l'année de règlement

Rate (Reported: US)		
Currency Code		Abréviation de la devise
CPI Adjustment		Taux d'ajustement de l'inflation basé sur le CIP américain et l'année de règlement
Categorization Comments	Regroupement	Commentaires expliquant la catégorisation de l'événement
Multiple Firms Impacted Code		Le code unique identifiant toutes les pertes subies par un groupe d'entreprises en raison d'un seul événement
Single Event, Multiple Loss Code	Admin II	Le code unique identifiant les multiples pertes subies par une seule entreprise en raison d'un seul événement
Specific Fields Updated		Identifie les colonnes qui ont été mises à jour à partir de la version précédente. Utilisation de la fonction de commentaire pour fournir des détails spécifiques sur le contenu mis à jour dans chacun des champs identifiés.
Update Comments		Commentaires détaillant les modifications apportées au contenu
Data Set Comments		Commentaires détaillant la ségrégation de l'événement en différents ensembles de données
Data Set Classification		Identifie l'ensemble de données dans lequel se situe l'événement
Financial Loss Status		Statut actuel de la perte (Final, Estimatif, Provisoire)
Date of Entry		Date à laquelle l'observation a été publiée
Date of Revision		Dernière date à laquelle l'observation a été modifiée

Variables ajoutées au jeu de données :

- Taille entreprise : classification de l'entreprise selon le nombre de salariés (TPE, PME, ETI et GE).
- Region of Incident : spécifie la région de survenance du sinistre.
- Secteur : indique le type du secteur : Financier ou Non Financier.

B. Dictionnaire 1 de mots-clés liés au risque cyber pour l'identification des risques cyber

access
account
accounts
adware
application
applications
associate
associates
attack
attacks
availabillity
available
backdoor
bitcoin
bitcoins
blackmail
blackmails
bot
botnet
breache
breaches
brute force
click
clicked
compromise
computer
confidential
confidential information
contractor
control
criminals
cross site scripting
cryptocurrency
cryptojacking
cyber
cyber attack
cyber crime
cyber fraud
cyber hacking
cyber thieve
cyberattack
cyber-attack
cyberattacks
cybercrime
cyber-crime
cybercrimes
cybercriminal
cybercriminal
cybercriminals
cybercriminals
cyberfraud
cyber-fraud
cyberhacking
cyber-hacking
cyberthieve
cyber-thieve
cyberthieves
damage
damages
dark web
data
ddos
denial of service
destroy
destroying
disable
disabling
disrupt
disrupt
disrupting
disruption
dns
domain name system
dos
drive by attack
drive by download
drive-by attack
eavesdrop
eavesdropping
email
emails
employee
employees
espionage
exfiltrate
exfiltration
exploit
extortion
failure
failures
fileless malware
fraudulent
glitch
hack
hacked
hacker
hackers
hacking
harm
hijack
hijacked
hijacking
hostage
hostages
illegitimate
infect
infected
infected
infection
infiltrate
infiltrated
information
insecure
insider threat
internet of things
interruption
interruptions
iot
keylogger
keyloggers
leak
leaks
link
links
logic bomb
login
login
lose
loses
loss
lost
malicious
malicious
malware
malwares
man in the middle
mistake
mistakes
network
partner
partners
password
personal information
phish
phished
phishing
phone
poison
poisoning
ransom
ransoms
ransomware
record
records
remote
rootkit
sabotage
scam
scammer
scams
scareware
sensitive
session hijacking
shut down
shutdown
smishing
social engineering
software
spear
spies
spoofed
spoofing

spraying
spy
spyware
spywares
steal
stole
suspicious
system
system crash
theft
track
trojan
trojans
tunneling
unauthorized
unsecured
url
url manipulation
viruse
viruses
vishing
web
website
whale
whaling
worm
xss
zero day
zero-day

C. Dictionnaire 2 de mots-clés liés au risque cyber pour l'identification des incidents cyber

adware
attack
attacks
blackmail
blackmails
bot
botnet
breache
breaches
brute force
confidential
cross site scripting
cryptojacking
cyber
cyber attack
cyber crime
cyber fraud
cyber hacking
cyber thieve
cyberattack
cyber-attack
cyberattacks
cybercrime
cyber-crime
cybercrimes
cybercriminal
cybercriminals
cyberfraud
cyber-fraud
cyberhacking
cyber-hacking
cyberthieve
cyber-thieve
cyberthieves
data breach
data breaches
ddos
denial of service
dns
domain name system
dos
drive by download
drive-by attack
eavesdrop
eavesdropping
espionage
fileless malware
hack
hacked
hacker
hackers
hacking
hijack
hijacked
hijacking
infected
insider threat
internet of things
iot
keylogger
keyloggers
logic bomb
malware
malware
malwares
man in the middle
personal information
phish
phished
phishing
phishing
poisoning
ransom
ransoms
ransomware
rootkit
scam
scammer
scareware
smishing
spoofed
spoofing
spraying
spy
spyware
spywares
system crash
trojan
tunneling
url
viruse
viruses
vishing
xss
zero day

D. Index des acronymes utilisés dans les arbres CART

1. Basel Business Line Level 2

AF : Accommodation and Foodservices
ASWMR : Administrative and Support, Waste Management and Remediation Services
BL : Basel Business Line - Level 2
CB : Commercial Banking
CS : Card Services
EC : External Clients
HCSA : Health Care and Social Assistance
HI : Health Insurance
Inf : Information
LIBP : Life Insurance and Benefit Plans
Man : Manufacturing
PB : Private Banking
PCI : Property and Casualty Insurance
PSTS : Professional, Scientific and Technical Services
RBa : Retail Banking
RBr : Retail Brokerage
RT : Retail Trade
Sa : Sales
TW : Transportation and Warehousing
WT : Wholesale Trade

2. Event Risk Category

EDPM : Execution, Delivery & Process Management
EF : External Fraud
Event : Event Risk Category
IF : Internal Fraud

3. Region of Domicile et Region of Incident

Africa : Africa
Asia : Asia
Domicile : Region of Domicile
Eur : Europe
Incident : Region of incident
Nam : North America
Others : Others