Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

<table>
<tr><td colspan="2">Par : Madame ADEDJOUMA Yemissy</td></tr>
<tr><td colspan="2">Titre du mémoire : Construction d'un produit d'assurance cyber pour les TPE/PME : Zoom sur la garantie atteinte aux données</td></tr>
<tr><td colspan="2">Confidentialité : ☐ NON ☐OUI (Durée : ☐ 1 an ☐2 ans)</td></tr>
<tr><td colspan="2">Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.</td></tr>
<tr><td>Membres présents du jury de la filière :
Signature :</td><td>Entreprise : Moonshot Insurance

Nom : CISSE TIDIANE

Signature :

Directeur de mémoire en entreprise</td></tr>
<tr><td>Membres présents du jury de l'Institut des Actuaires :
Signature :</td><td>Nom : CHADI HANNA/ CISSE TIDIANE
Signature :
Hanna C.

Invité :
Nom :
Signature :</td></tr>
<tr><td></td><td>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 :

Signature du candidat :</td></tr>
</table>

RESUME

Les nouvelles technologies ont réformé les processus de production, de commercialisation, de communication et de traitement des données au sein des entreprises. En raison de cette digitalisation progressive de l'économie accentuée par la crise sanitaire actuelle, les entreprises sont de plus en plus exposées aux attaques informatiques. Ainsi, s'observe une intensification des risques numériques tels que le risque cyber.

Face à l'expansion de ce risque, les assureurs proposent des contrats d'assurance afin de permettre aux entreprises d'y faire face plus efficacement. Cependant, cette tâche peut parfois s'avérer complexe du fait de l'existence d'un nombre limité de données sur le sujet. Cela est principalement dû à la complexité et au caractère évolutif du cyber-risque.

Le principal objectif de ce mémoire est de présenter la construction d'un produit d'assurance cyber à destination des PME/TPE. Pour y parvenir, une présentation détaillée du risque ainsi que du marché de la cyber assurance a été faite afin de recenser les caractéristiques et les spécificités de la menace étudiée. Le présent mémoire comporte également, une description du processus de création d'un produit d'assurance ainsi qu'une fiche produit de l'offre construite.

La modélisation du risque d'atteinte aux données a été basée sur l'utilisation de la base Privacy Clearing House (PRC), qui a permis d'estimer la sévérité des incidents cyber grâce à l'utilisation de lois théoriques et de construire des modèles linéaires généralisés pour approcher la fréquence des attaques.

Par ailleurs, dans le souci d'explorer d'autres aspects de la modélisation du risque de cyber attaque, une estimation de la probabilité de perte d'exploitation d'une entreprise suite à un incident cyber a également été faite à l'aide de méthodes bayésiennes appliquées aux données de la base VERIS (Vocabulary for Event Recording and Sharing).

Mots clés : Digitalisation, Privacy Clearing House, Tarification, Cyber-attaques, Assurabilité, Modèle linéaire généralisé, Données personnelles, Risque cyber, Prévention, Création de produit, Loi de poisson zéro-tronquée, Loi binomiale négative zéro-tronquée, Perte d'exploitation, Inférence bayésienne, Veris, Elicitation de l'avis d'expert, loi a priori, loi a posteriori.

ABSTRACT

New technologies have changed the production, marketing, communication and data processing in companies. As a result of this progressive digitalisation of the economy, accentuated by the current health crisis, companies are increasingly exposed to computer attacks. This situation lead to an increase of digital threats such as cyber risk.

In response to the expansion of this risk, insurers are offering insurance policies to enable companies to deal with it more effectively. However, this can sometimes be a complex task due to the lack of available data on the subject. This is mainly caused by the complexity and evolving nature of cyber risk.

The main objective of this paper is to present the construction of a cyber insurance product for SMEs. To achieve this, a detailed presentation of the risk and the cyber insurance market has been made in order to identify the characteristics and the specificities of the studied threat. This report also includes a description of the process of developing an insurance product and a product card for the offer constructed.

The risk modelling of data breach was based on the use of the Privacy Clearing House database, which made it possible to estimate the severity of cyber incidents through the use of theoretical laws and to construct generalized linear models to approximate the frequency of attacks.

In addition, in order to explore other aspects of the risk modelling of a cyber attack, an estimation of the probability of a business interruption following a cyber incident was also carried out by using Bayesian methods applied to data from the VERIS (Vocabulary for Event Recording and Sharing) database.

Key words : Digitalisation, Privacy Clearing House, Pricing, Cyber-attacks, Insurability, Generalized linear model, Personal data, Cyber risk, Prevention, Insurance offer, Zero-Truncated Poisson distribution, Zero-Truncated Negative Binomial distribution, Business interruption, Bayesian Inference, Veris, Elicitation of expert opinion, Prior probabillity, Posterior probabillity

NOTE DE SYNTHESE

1 Contexte de l'étude et problématique étudiée

Les nouvelles technologies font aujourd'hui partie intégrante du quotidien de nombreuses entreprises. Cette digitalisation progressive renforcée par la crise sanitaire actuelle, facilite cependant leur exposition aux attaques informatiques. Ainsi, s'observe au sein des entreprises, une intensification des risques numériques tels que le risque cyber.

Afin de permettre à ces dernières d'affronter efficacement les conséquences de ces attaques informatiques, plusieurs contrats d'assurance sont proposés par les assureurs du marché. Cependant, la construction d'une offre d'assurance cyber peut parfois s'avérer complexe du fait de l'existence d'un nombre limité de données sur le sujet. Cela est principalement dû à la complexité et au caractère évolutif du risque. Ainsi, la modélisation et la quantification du risque cyber devrait passer avant tout par une analyse approfondie de ce risque et de ses caractéristiques et spécificités. Cela devant permettre une meilleure appréhension du risque et faciliter la construction d'hypothèses fiables nécessaires à sa quantification.

Ce mémoire aura pour principal objectif de réaliser une étude détaillée du risque cyber en vue de la construction d'un produit d'assurance cyber à destination des PME/TPE.

2 Etude du risque cyber

Le risque numérique se caractérise principalement par sa complexité. Il est en effet assez complexe d'obtenir une définition claire et détaillée de ce risque ainsi que de mesurer son ampleur. Cela pourrait s'expliquer en partie par l'existence de plusieurs types de cyber attaques, ce qui rend difficile l'analyse des différentes conséquences qui pourraient découler de ces incidents. Au nombre des types de cyber attaques, figurent : les attaques par malware, l'hameçonnage, le téléchargement furtif, les attaques par déni de service et bien encore.

De même, le risque cyber est caractérisé d'évolutif, étant donné que les types d'attaques et les moyens d'infiltration sont en constante évolution. Chaque nouvelle attaque enregistrée peut être totalement différente de la précédente. Ainsi, du fait de cette caractéristique, aucun programme ou système de prévention ne peut garantir une protection efficace à 100% face à la menace cyber. Néanmoins, ce phénomène est pris en compte par les assureurs qui adaptent leurs offres aux nouvelles formes d'attaques enregistrées.

Par ailleurs, la menace cyber peut être aussi qualifiée de systémique et contagieuse. En effet, du fait de l'interdépendance des systèmes informatiques, les probabilités de propagation des incidents cyber sont élevées. Un même virus informatique, peut se propager en s'autorépliquant dans un programme et infecter presque instantanément des dizaines de milliers d'ordinateurs.

3 Cyber assurance

La cyber assurance figure parmi les solutions apportées pour aider les entreprises à faire face aux désastreuses conséquences des cybers attaques. Ainsi, face à l'expansion de la menace cyber, les entreprises sont de plus en plus conscientes de la nécessité d'investir dans un contrat d'assurance cyber. Le marché de la cyber assurance connait alors un essor notoire et cela se manifeste par la création de nombreuses offres d'assurance cyber à destination de différentes catégories d'entreprises de secteurs variés. D'ailleurs, selon une étude de l'AMRAE, le volume de primes d'assurance cyber a augmenté de 49 % entre 2019 et 2020 : il est passé de 87 millions d'euros à 130 millions d'euros. (AMRAE, 2021)

Les garanties pouvant être incluses dans les offres proposées sont : atteinte aux données clients, cyber extorsion ou encore perte d'exploitation. Dans certaines offres d'assurance, la prise en charge de ces garanties par l'assureur est conditionnée à la mise en place au sein de l'entreprise assurée d'un mécanisme de cyber sécurité efficace, principal moyen de réduction de risque. L'instauration d'un mécanisme de cyber sécurité efficient peut se faire à l'aide de nombreux services de prévention proposés sur le marché tels que : les antivirus, le patch manager ou encore des gestionnaires de mots de passe.

Toutefois, il faut noter que malgré la croissance du marché de l'assurance cyber, plusieurs problèmes restent encore à résoudre. En effet, les couvertures d'assurance disponibles sont inégalement réparties sur le marché. Le marché est en grande partie porté par les grandes entreprises et peine encore à décoller en ce qui concerne les PME/TPE. De même, face à la dégradation de leurs ratios Sinistre/primes, les assureurs ont pour la plupart durci les conditions de souscriptions et élevé le niveau des primes. Cela a entrainé une baisse globale des taux de couvertures. Certaines entreprises renoncent même à souscrire à une assurance cyber et se tournent vers diverses solutions qui pourraient être moins couteuses telles que l'auto-assurance.

Par ailleurs, malgré l'existence de multiples contrats d'assurance cyber, plusieurs débats sont encore à mener sur la question de l'assurabilité de ce risque.

Conception d'une offre d'assurance cyber pour les PME/TPE

Un produit d'assurance sert principalement à se couvrir contre un éventuel danger inhérent à une activité ou à une situation et qui pourrait engendrer des coûts très importants. Ainsi, la conception d'un produit d'assurance nécessite minutie et rigueur afin de pouvoir analyser au mieux toutes les caractéristiques du risque à couvrir et de proposer les solutions les plus efficaces pour y faire face. Le processus de création d'une offre d'assurance se fait principalement en 3 grandes étapes :

- **Réflexion** : La phase de réflexion consiste principalement à s'interroger sur la pertinence de l'offre d'assurance proposée et à réaliser une étude de marché afin d'évaluer les différents aspects des offres similaires disponibles.

- **Conception du produit** : La phase de conception du produit est l'une des plus importante de ce processus car c'est au cours de cette étape, qu'une analyse approfondie du risque à couvrir est réalisée. A cette étape seront ainsi définies les caractéristiques (garanties, exclusions…) de l'offre à construire.

- **Calibration du risque** : Il s'agit là de la phase finale du processus. Elle consiste à concrétiser toutes les idées soulevées en phase 1 et 2. Elle implique de nombreux de aller-retours entre les équipes marketing et actuariat afin de définir le tarif des différentes garanties à inclure dans l'offre.

L'offre d'assurance proposée dans le cadre du mémoire est à destination des PME/TPE de moins de 20 salariés et ayant un chiffre d'affaires inférieur à 10 millions d'euros.

Sur la base d'une étude du marché et de l'analyse des besoins des PME/TPE en matière de couverture cyber, les garanties qui sont retenues dans le contrat d'assurance proposé sont :

- **Atteinte aux données** : Comme son nom l'indique, la présente garantie peut être déclenchée en cas de destruction, perte, altération, divulgation ou d'accès non autorisé aux données personnelles confidentielles de l'assuré ou de celles détenues pour un tiers et causée par une erreur humaine ou une cyberattaque.

- **Cyber vol** : Cette garantie est déclenchée en cas de virement frauduleux des comptes de l'assuré vers un compte tiers lors d'une attaque cyber et sans collaboration de celle-ci.

- **Cyber extorsion** : Cette garantie intervient en cas de menace d'extorsion par un cyber-pirate, aux fins d'obtenir le paiement d'une rançon suite à l'endommagement, la destruction, la modification ou la corruption du système informatique de la société assurée.

- **Responsabilité civile**

- **Perte d'exploitation**

La prise en charge de ces garanties par l'assureur est conditionnée à la mise en place au sein des entreprises assurées de services de prévention afin de permettre une réduction efficace du risque. Les services de prévention retenus sont :

Formation des équipes	Authentification multifacteur	Campagnes de phishing factices
Antivirus/Antimalware	Mise à disposition d'experts en cyber sécurité	Gestionnaire de mot de passe

Figure 1 : Services de préventions retenus

4 Tarification de la garantie « Atteinte aux données »

La méthode de tarification utilisée est la méthode de **fréquence x coût**.

Présentation des données

Mesurer l'exposition d'une entreprise au risque cyber n'est pas chose aisée. En effet, le caractère complexe et évolutif du risque rend épineux la recherche de bases de données complètes et fiables.

Certains assureurs présents sur le marché de l'assurance cyber, disposent de base de données internes. Ces dernières peuvent être utiles pour mesurer l'exposition réelle de leur portefeuille à la menace mais peuvent être limitées pour mesurer l'exposition globale au risque.

Dans le cadre de ce mémoire, plusieurs bases de données ont été explorées. Moonshot insurance étant nouveau sur le périmètre de l'assurance cyber pour les professionnels, l'assurtech ne possède pas de données internes.

Après une analyse des différentes bases publiques disponibles sur le risque cyber, la base Privacy Rights ClearingHouse (PRC) est celle qui a été retenue pour la modélisation du risque d'atteinte aux données. Il s'agit d'une base américaine répertoriant depuis 2005 des incidents de fuites de données affectant des entreprises américaines. Dans cette base de données, plusieurs informations sur les caractéristiques des cyber événements reportés (localisation, date…) sont recensées. Plus de détails sur ces caractéristiques seront fournis dans la suite du mémoire.

Modélisation de la sévérité

Sévérité en nombre de données compromises

Dans la détermination de la sévérité d'un incident de violation de données, deux principaux éléments sont à prendre en compte : la taille de la violation (nombre de données compromises) ainsi que la sensibilité de la violation (coût de la donnée perdue). Dans la base PRC étudiée, la variable ***Total Records*** fournit des informations sur la taille des violations enregistrées. Cela permettra ainsi d'estimer la sévérité en nombre de données compromises des incidents cyber.

Afin d'obtenir un tarif final adapté aux caractéristiques de chaque type d'entreprise, il est plus prudent pour la suite, d'évaluer le nombre moyen de données compromises pour chaque secteurs pris séparément. Les graphiques ci-dessous représentent les distributions en densité des tailles des violations pour chaque type d'organisation (BSO, BSR, BSF). Afin de faciliter l'analyse et la lisibilité des graphes, l'étude réalisée dans ce mémoire se portera sur le logarithme des tailles de violations.

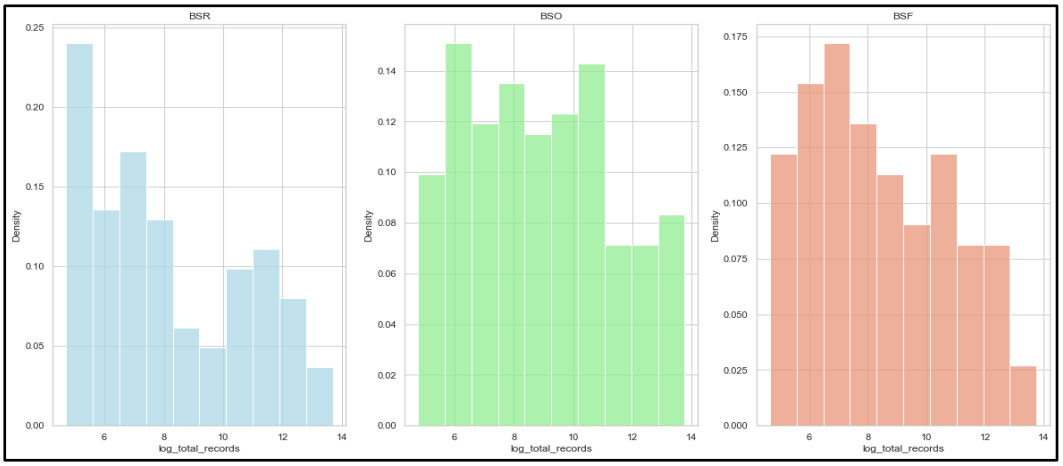


Figure 2 : Densité des tailles des violations pour chaque type d'organisation

A première vue, les distributions des tailles des violations sont assez différentes pour chaque secteur d'entreprise. Ce qui conduit à penser que la répartition des tailles de violations varie en fonction du secteur de l'entreprise. La segmentation des tarifs de l'offre d'assurance en par secteurs d'activités semble donc être une bonne idée. Ainsi, une étude approfondie des spécificités des incidents affectant chaque type d'organisation afin de déterminer un coût moyen de sinistre pour chacune d'elles sera effectuée.

Pour chacun des secteurs d'activités retenus, la sévérité en nombre de données des incidents de violations de données sera évaluée à l'aide de l'ajustement de loi de probabilité aux données. **Cependant, étant donné la diversité des sources d'informations approvisionnant la base PRC, une étude approfondie des incidents recensés par chaque source a été effectuée en amont de l'ajustement de la loi. Cela devant permettre de limiter le biais que pourrait engendrer la multitude des sources d'alimentation de la base.** Ci-dessous la distribution du logarithme des tailles de violations par source d'information dans chacun des secteurs d'activités considérés :

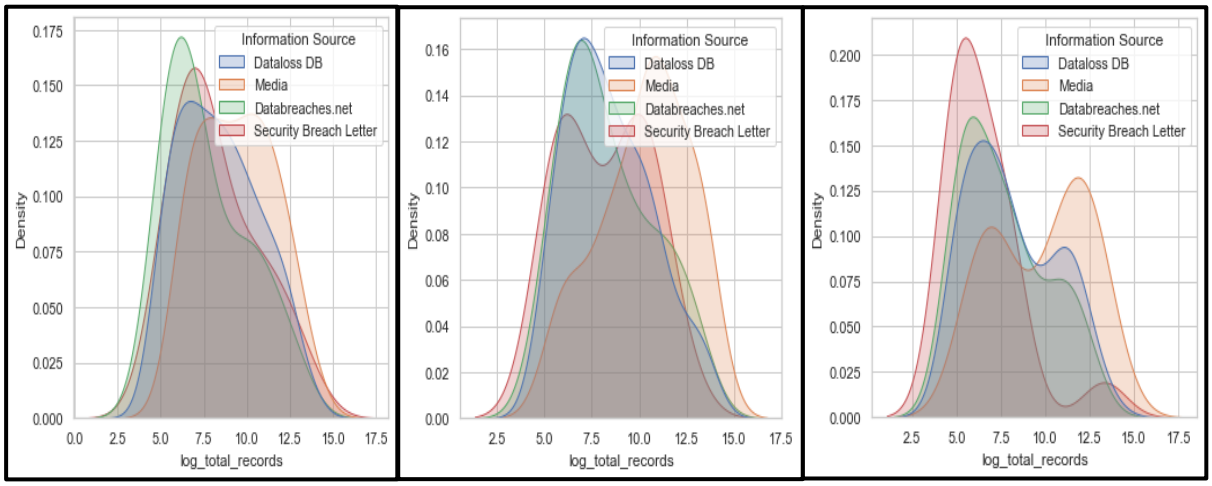


Figure 3 : BSF

Figure 4 : BSO

Figure 5 : BSR

Une franchise à 100 données compromises ainsi qu'un plafond d'indemnisation à 50 000 données ont également été appliqué aux données étudiées.

Sévérité en coût

Afin d'associer aux nombres moyens de données violées dans chaque secteur un coût, il a fallu dans un premier temps, estimer le coût unitaire d'une donnée. L'étude *« Cost of data breach »* publiée par le Ponemon institute, examine le montant des coûts engendrés par la perte ou le vol de données personnelles. Dans l'édition 2019 de cette étude, il a été possible d'extraire le coût unitaire par secteur d'une donnée violée. L'estimation de ce coût prend en compte :

- Les frais de notifications de l'incident aux personnes concernées

- Les frais de détection et de gestion de l'incident

- Les coûts post violation

- Les coûts liés à une interruption de l'activité de l'entreprise

Il a été mentionné dans ce rapport que, les coûts liés à une interruption de l'activité de l'entreprise représentent 38% du coût moyen total d'un incident cyber. Etant donné que la perte d'exploitation fait l'objet d'une garantie à part entière dans l'offre d'assurance proposée dans ce mémoire, la part liée aux coûts engendrés par une interruption de l'activité de l'entreprise sera déduite du coût moyen obtenu.

Le coût moyen de sinistre pour la garantie atteinte aux données a été calculé à l'aide de la formule ci-dessous :

$$Coût_{moyen} = Nb_{moy} * unit_{moy}$$

Avec :
Nb_{moy} : le nombre moyen de données compromises
$Unit_{moy}$: le coût moyen unitaire d'une donnée violée

Etant donné que le coût unitaire de la donnée disponible date de l'édition 2019 du rapport, il faudrait l'actualiser afin d'obtenir une vision actuelle de la sévérité des incidents de violations de données. Diverses approches peuvent être utilisées pour actualiser le coût unitaire d'une donnée. Une première idée serait de se baser sur les taux d'inflations annuels entre 2019 et 2022 ou encore de faire une simple translation entre les coûts moyens totaux des incidents de violations de données fournis dans les rapports Ponemon de 2019 et 2022.

Pour l'étude réalisée ici, la seconde approche sera privilégiée car elle semble plus pertinente. En effet, le calcul du taux d'inflation annuel se base sur des facteurs tels que l'indice des prix à la consommation ou encore les dépenses moyennes des ménages. Ces éléments paraissent tous assez décorrélés de l'évolution du coût moyen d'une données. De plus, plusieurs éléments, autre que l'inflation, peuvent influer sur la variation du coût moyen unitaire d'une donnée compromise. Ce peut être par exemple, l'évolution des techniques de sécurité mise en place ou encore l'amélioration des dispositifs de réponse aux incidents cyber.

Modélisation de la fréquence

Certaines raisons laissent penser que la base PRC n'est pas adéquate pour modéliser rigoureusement la fréquence de sinistre. En effet, la diversité des sources d'information alimentant la base ne permet pas de faire la distinction entre l'évolution de la sinistralité causée par une réelle augmentation du risque et celle venant de l'instabilité de de ces sources d'informations. Le graphique suivant montre l'évolution du nombre d'incidents rapportés par les différentes sources d'informations dans la base :

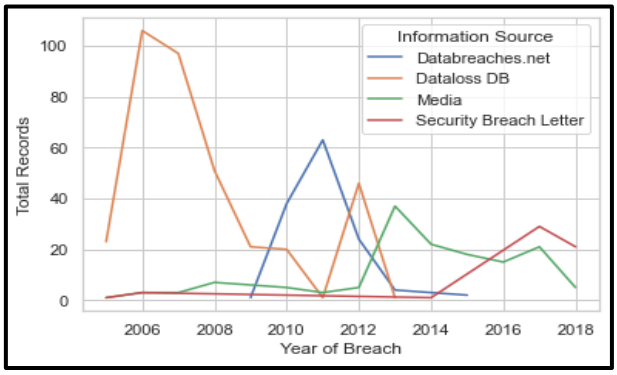


Figure 6 : Evolution du nombre d'incidents enregistrés par source d'information par année

Face au déséquilibre important observé sur les données de la base PRC, on admettra que ces dernières sont zéro-tronquées. Cela reviendrait à considérer que les entreprises n'ayant pas eu de sinistre cyber ne sont pas observées dans la base. Ensuite, pour estimer la fréquence de sinistre, un modèle linéaire généralisé (GLM) prenant pour variables explicatives les secteurs d'entreprises (BSR, BSO, BSF) sera ajustée aux données. Cela permettra de prendre en compte les spécificités de chaque type d'organisation et ainsi d'obtenir une fréquence plus adaptée à chaque secteur. Trois types de GLM seront testés sur les données : un GLM basé sur une loi de poisson 0- tronquée, un sur une loi géométrique 0-tronquée et un dernier sur une loi binomiale négative 0-tronquée.

Des critères statistiques tels que la déviance et l'AIC permettront de départager ces trois modèles et de définir le plus prédictif.

Dans un second temps, afin d'améliorer la précision des estimations faites l'utilisation d'un modèle multinomial permettra de quantifier l'impact des différents types d'attaques sur la fréquence de sinistres.

5 Estimation de la fréquence de sinistre d'une garantie de « perte d'exploitation »

La garantie ***perte d'exploitation*** couvre les pertes liées à la baisse ou à l'arrêt de l'activité économique d'une entreprise à la suite d'un événement donné. Ici, la garantie prend en charge la perte d'activité survenant suite à une attaque cyber. Ainsi, pour cette garantie, la fréquence de sinistre sera approchée par le pourcentage d'entreprises qui subissent une interruption ou une baisse d'activité causée par une cyber attaque.

Base de données

La base utilisée pour la résolution de cette problématique est la base Veris. VERIS (Vocabulary for Event Recording and Sharing) désigne un ensemble de métriques conçues dans le but de fournir un langage commun permettant de décrire les incidents cyber de façon claire, structurée et reproductible. Cet outil permet de répondre efficacement au manque d'informations fiables disponibles en matière de risque cyber et aide les organisations à mieux mesurer et gérer ce risque. La base de données mise à disposition par ce dispositif est assez riche avec plus de 8000 incidents individuels recensés. Elle fournit plusieurs informations utiles sur les caractéristiques des incidents survenus (acteurs, type d'attaque rencontré, perte financière suite à l'incident, délai de réponses etc). De nombreuses caractéristiques relatives aux victimes de l'incident (pays, nombre d'employés, revenu annuel etc) figurent également dans cette base de données.

Modélisation

La modélisation de la fréquence de sinistre de la garantie ***perte d'exploitation*** se fera à l'aide de méthodes bayésiennes. Ces dernières permettent globalement de déduire la probabilité d'un événement en se basant sur des probabilités d'autres événements déjà estimées : on cherche à exprimer ce que l'on sait sur les inconnus du problème sachant ce qui est connu (ou assumé comme tel). Ces méthodes reposent sur la notion de probabilité personnelle et leur cohérence est garantie par des règles mathématiques du calcul des probabilités. Le principe de l'approche bayésienne peut être décrit à l'aide du schéma ci-dessous :

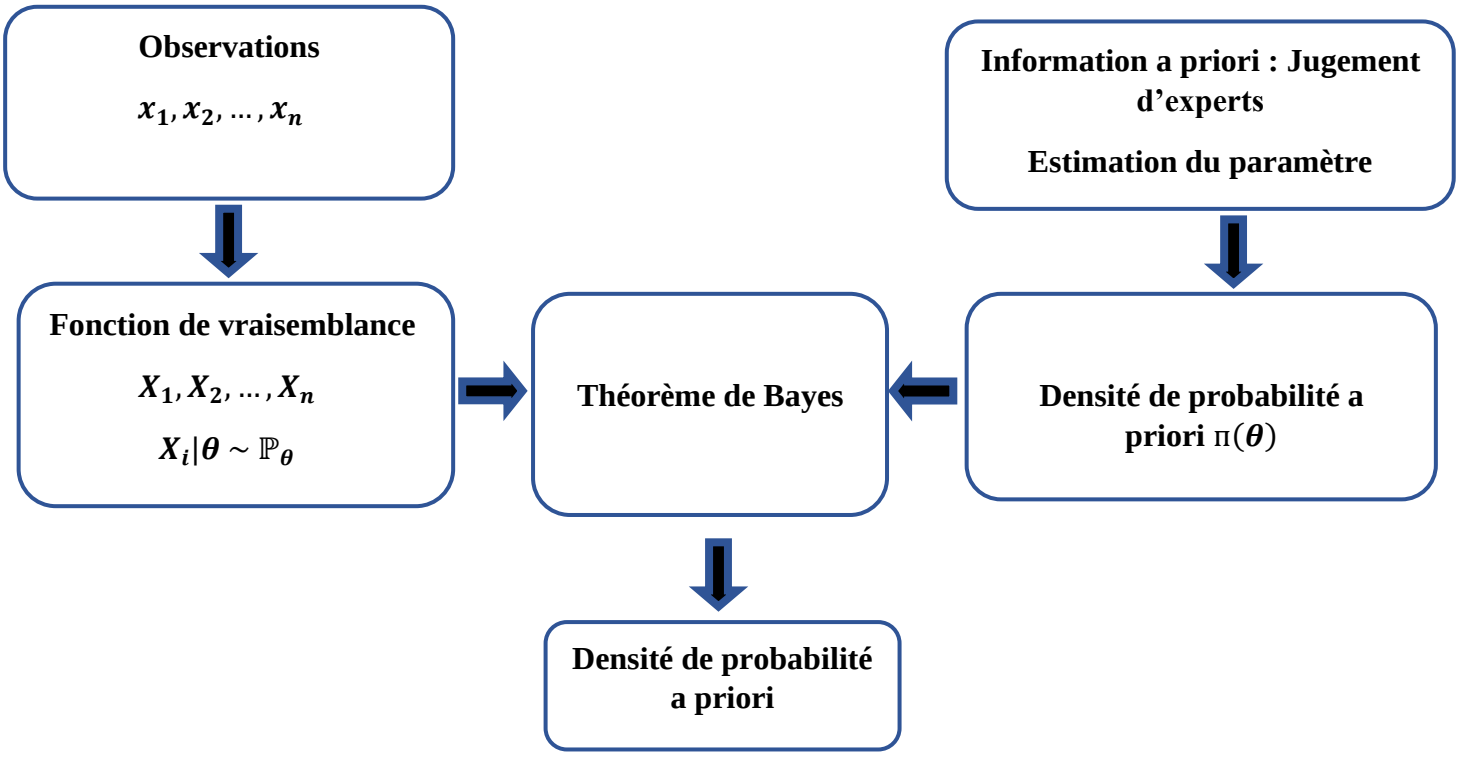


Figure 7 : Principe de l'approche bayésienne

Approche 1 – Utilisation d'une loi a priori informative

Le premier modèle sera basé sur l'utilisation d'une loi a priori informative. Une distribution a priori est dite informative si elle est non dominée par la vraisemblance et a un impact sur la distribution a posteriori. Les lois a priori informatives doivent être définies avec précaution dans la pratique et privilégient certaines valeurs de la quantité d'intérêt. Ces préférences sont généralement basées sur des études antérieures.

Soit θ la probabilité qu'une entreprise victime de cyber attaque subisse une perte d'exploitation. L'estimation du paramètre θ se fera à l'aide des données d'observations de la base Veris. Soient des variables aléatoires $X_{i,i=1\ldots n}$ (n étant la taille de la base étudiée) désignant les entreprises de la base Veris. Les réalisations de ces variables aléatoires seront notées x_i ; $x_i \in \{0,1\}$ étant l'état de l'entreprise i après une attaque cyber où $x_i = 1$ si l'entreprise est victime d'une interruption d'activité et $x_i = 0$ sinon. Alors : $X_i \sim$ Bernoulli (θ)

Choix de la loi a priori

Puisqu'il existe peu d'informations a priori, une option pour le choix de la loi a priori est d'utiliser une loi a priori conjuguée. Dans le cadre de cette étude, $X_i|\theta \sim$ Bernoulli(θ), i=1,…n avec n la taille de la base étudiée. La conjuguée naturelle de la loi de Bernoulli étant la loi bêta, on a : $\theta \sim Be(\alpha, \beta)$. Alors la densité de la loi a priori ici est :

$$\boxed{\pi(\theta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\,1_{[0,1]}(\theta)}$$

D'après la formule de Bayes :

$$\boxed{\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}}$$

Par conséquent, la densité de la loi a posteriori est :

$$\boxed{\pi(\theta|x) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}1_{[0,1]}(\theta)}$$

Afin de déterminer cet estimateur il faudrait inclure l'opinion de l'expert considéré.

Elicitation de l'avis d'expert

Les experts auxquels l'on pourrait se référer dans le cadre de la problématique actuelle sont généralement des professionnels de la cyber sécurité. L'opinion de ces experts est requise dans le but d'obtenir des informations statistiques (moyenne, variance…) sur le paramètre d'intérêt θ.

N'ayant malheureusement pas pu obtenir des rendez-vous avec des experts en cyber sécurité capables de fournir des informations sur le paramètre θ, l'expertise sera ici basée sur les résultats obtenus dans le baromètre de la cybersécurité des entreprises réalisé par Opinionway

pour le CESIN[1]. De cette étude sera tirée l'espérance de θ ainsi qu'un intervalle de confiance à 95% pour cette valeur.

La combinaison de toutes ces informations permettra ainsi d'obtenir les paramètres des lois a priori et a posteriori et ainsi de donner l'estimateur bayésien du paramètre d'intérêt θ. Cet estimateur est défini comme la moyenne de la loi a posteriori. Une étude de sensibilité de la réponse bayésienne à la loi a priori sera ensuite réalisée.

Approche 2 – Utilisation d'une loi a priori non informative

Cette approche se base sur l'utilisation d'une loi a priori non informative. La loi a priori obtenue est ensuite mise à jour avec les données d'observations de la base Veris dans un premier temps, ce qui permet d'obtenir une première distribution a posteriori. Dans un second temps, la loi a posteriori obtenue sera elle aussi mise à jour avec des informations fournie par un expert. Cela permettra ainsi d'obtenir une deuxième loi a posteriori.

1ière étape :

Ne disposant pas d'informations précises sur le paramètre d'intérêt θ, une loi a priori non informative sera utilisée. La loi non informative retenue ici est la loi uniforme sur [0,1]. Ainsi, θ ~ U (0,1). Cela permet d'obtenir une loi bêta comme loi a posteriori.

2ième étape

A cette étape, la loi a posteriori obtenue précédemment devient une loi a priori. Cette loi est mise à jour avec l'information apportée par l'expert.

Elicitation de l'avis d'expert

Pour cette méthode également on imaginera que les experts qui seront, généralement des professionnels de la cyber sécurité disposent de connaissances statistiques (moyenne, variance…) sur le paramètre d'intérêt θ. L'expert contacté peut soit être sûr à 100% de son estimation de la valeur de θ ou alors définir une distribution de probabilité pour θ qui prendra en compte son incertitude sa la valeur du paramètre estimé.

De façon générale, l'information recherchée est obtenue en réalisant des sondages ou des interviews auprès de plusieurs experts. En raison du manque de temps et de la complexité du risque étudié, des intervalles de confiance à 95% sur la valeur du paramètre θ seront simulés aléatoirement et serviront d'avis d'expert. Les informations statistiques recueillies auprès de ces experts fictifs permettront ensuite de mettre à jour la loi a priori pour obtenir une nouvelle loi a posteriori. Une fois la loi a posteriori explicitée, l'estimation finale de θ se fera alors à l'aide d'une moyenne effectuée sur les estimateurs de Bayes obtenus grâce aux informations de chaque expert.

[1] Club des Experts de la Sécurité de l'Information et du Numérique

EXECUTIVE SUMMARY

1 Study Framework and Issues

New technologies are now an integral part of the daily life of many companies. However, this progressive digitalization, reinforced by the current health crisis, makes it easier for them to be exposed to computer attacks. Thus, there is in companies an intensification of digital risks such as the cyber risk.

In order to ensure that companies can deal effectively with the consequences of these computer attacks, several insurance policies are created on the market. However, the construction of a cyber insurance offer can sometimes be complex task because of the existence of a limited amount of data on the subject. This is mainly due to the complexity and evolving nature of the risk. Thus, the modelling and quantification of the cyber risk should be based on an in-depth analysis of this risk and its characteristics and specificities. This should allow a better understanding of the risk and facilitate the construction of reliable hypotheses necessary for its quantification.

The main objective of this thesis will be to carry out a detailed study of the cyber risk with a view to building a cyber insurance product for SMEs.

2 Cyber risk study

The main characteristic of digital risk is its complexity. It is indeed quite complex to obtain a clear and detailed definition of this risk and to measure its magnitude. This could be partly explained by the existence of several types of cyber-attacks, which makes it difficult to analyze the different consequences that could arise from these incidents. Types of cyber-attacks include : malware attacks, phishing, stealth downloads, denial of service attacks and more.

Similarly, cyber risk is characterized as evolutionary, as the types of attacks and means of infiltration are constantly changing. Each new attack recorded may be completely different from the previous one. Thus, because of this characteristic, no prevention program or system can guarantee 100% effective protection against the cyber threat. Nevertheless, this phenomenon is considered by insurers who adapt their offers to the new forms of attacks recorded.

Moreover, the cyber threat can also be described as systemic and contagious. Indeed, due to the interdependence of computer systems, the probability of propagation of cyber incidents is high. A single computer virus can spread by self-replicating in a program and infect tens of thousands of computers almost instantaneously.

3 Cyber insurance

Cyber insurance is one of the solutions provided to help companies cope with the disastrous consequences of cyber-attacks. As the cyber threat grows, companies are increasingly aware of the need to invest in cyber insurance. The market for cyber insurance is therefore booming and this is reflected in the creation of numerous cyber insurance offers for different categories of companies in various sectors. Moreover, according to a study by AMRAE, the volume of cyber insurance premiums increased by 49% between 2019 and 2020: it rose from 87 million euros to 130 million euros. (AMRAE, 2021)

The coverages that can be included in the proposed offers are : breach of customer data, cyber extortion or business interruption. In some insurance offers, the insurer's coverage of these guarantees is conditional on the implementation of an effective cyber security mechanism within the insured company, which is the main means of risk reduction. The implementation of an efficient cyber security mechanism can be done with the help of numerous prevention services offered on the market such as: antivirus, patch manager or password managers.

However, it should be noted that despite the growth of the cyber insurance market, several problems remain to be solved. Indeed, the insurance cover available is unevenly distributed on the market. The market is largely driven by large companies and is still struggling to take off regarding SMEs. Similarly, faced with the deterioration of their loss/premium ratios, most insurers have tightened their underwriting conditions and raised the level of premiums. This has led to an overall drop in coverage rates. Some companies are even giving up on cyber insurance and turning to various solutions that may be less costly, such as self-insurance.

Moreover, despite the existence of multiple cyber insurance contracts, several debates are still to be held on the question of the insurability of this risk.

Creation of a cyber insurance offer for SMEs

The main purpose of an insurance product is to cover against a possible danger inherent in an activity or situation that could result in very high costs. Thus, the design of an insurance product requires meticulousness and rigor in order to be able to analyze all the characteristics of the risk to be covered and to propose the most effective solutions to deal with it. The process of creating an insurance offer consists mainly of 3 main stages:

- Reflection: The reflection phase mainly consists of questioning the relevance of the proposed insurance offer and carrying out a market study in order to evaluate the different aspects of similar offers available.

- Product design: The product design phase is one of the most important in this process because it is during this stage that an in-depth analysis of the risk to be covered is carried out. At this stage, the characteristics (guarantees, exclusions, etc.) of the offer to be built are defined.

- Risk calibration: This is the final phase of the process. It consists of putting into practice all the ideas raised in phases 1 and 2. It involves a lot of back and forth between the marketing and actuarial teams in order to define the tariffs for the various guarantees to be included in the offer.

The insurance offer proposed in the framework of the brief is intended for SMEs with less than 20 employees and an annual revenue of less than 10 million euros.

Based on market research and an analysis of the needs of SMEs in terms of cyber coverage, the guarantees that are retained in the proposed insurance contract are

- Data breach: As the name suggests, this cover can be triggered in the event of destruction, loss, alteration, disclosure, or unauthorized access to the insured's confidential personal data or that held for a third party caused by human error or a cyber-attack.

- Cyber theft: This cover is triggered in the event of a fraudulent transfer from the insured's accounts to a third-party account during a cyber-attack and without their cooperation.

- Cyber extortion: This cover is triggered in the event of a threat of extortion by a cyber hacker, for the purpose of obtaining payment of a ransom following the damage, destruction, modification or corruption of the insured company's computer system.

- Civil liability

- Business interruption

The insurer's coverage of these guarantees is conditional on the implementation of preventive services within the insured companies in order to allow an effective reduction of the risk. The prevention services selected are :

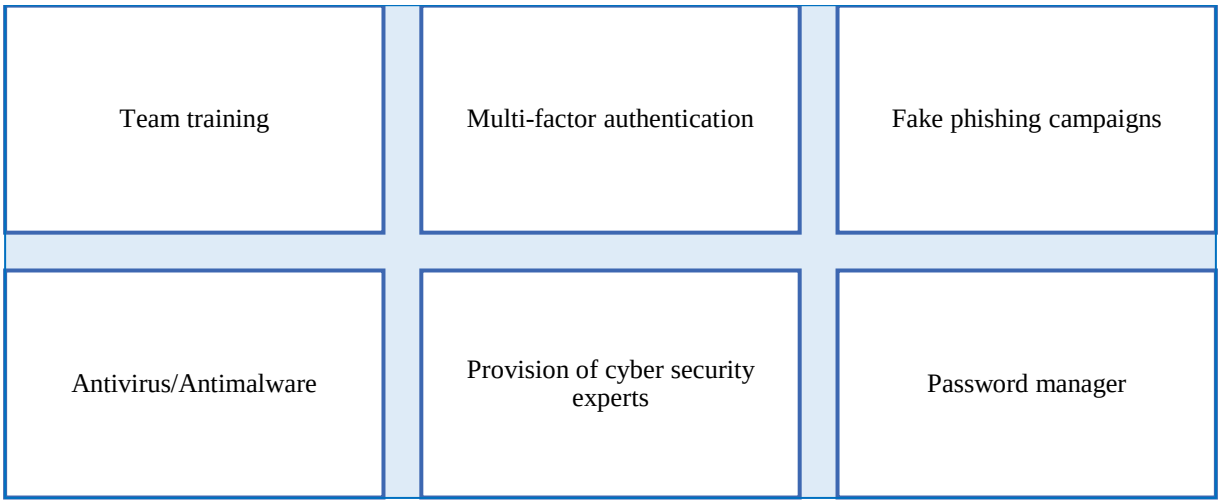


Figure 8 : Selected prevention services

4 Pricing of a "Data Breach" cover

The pricing method used is the frequency x cost method.

Data presentation

Measuring a company's exposure to cyber risk is not an easy task. Indeed, the complex and evolving nature of the risk makes it difficult to find complete and reliable databases.

Some insurers in the cyber insurance market have internal databases. These can be useful for measuring the actual threat exposure of their portfolio but may be limited for measuring overall risk exposure.

In the context of this thesis, several databases were explored. As Moonshot Insurance is new to the field of cyber insurance for professionals, the insurtech does not have any internal data.

After an analysis of the various public databases available on cyber risk, the Privacy Rights ClearingHouse (PRC) database was chosen for modelling the risk of a data breach. This is an American database listing data breaches incident affecting American companies since 2005. In this database, several pieces of information on the characteristics of the reported cyber events (location, date, etc.) are listed. More details on these characteristics will be provided later in the report.

Modelling of severity

Severity in number of compromised data

When determining the severity of a data breach incident, there are two main elements to consider: the size of the breach (number of compromised data) and the sensitivity of the breach (cost of the lost data). In the PRC database studied, the variable Total Records provides information on the size of the recorded breaches. This will allow the estimation of the severity in number of compromised data of cyber incidents.

To obtain a final rate adapted to the characteristics of each type of company, it is more cautious to evaluate the average number of compromised data for each sector taken separately. The graphs below represent the density distributions of breach sizes for each type of organization (BSO, BSR, BSF). In order to facilitate the analysis and readability of the graphs, the study carried out in this thesis will focus on the logarithm of the breach sizes.

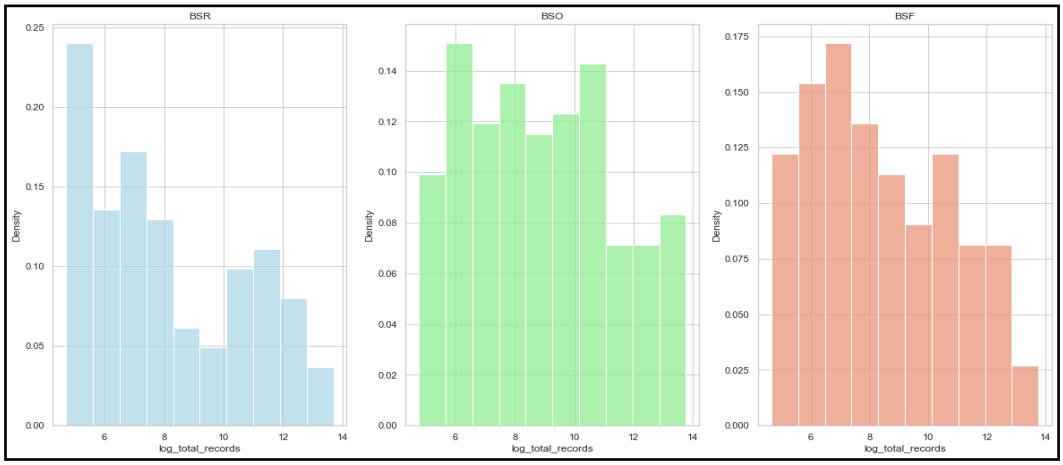


Figure 9 : Density of violation sizes for each type of organization

At first sight, the distributions of breach sizes are quite different for each company sector, suggesting that the distribution of breach sizes varies according to the sector of the company. The segmentation of insurance prices by sector of activity therefore seems to be a good idea. Thus, an in-depth study of the specificities of the incidents affecting each type of organization in order to determine an average cost of loss for each of them will be carried out.

For each of the selected sectors of activity, the severity in terms of the number of data breaches will be assessed using a probability distribution fitted to the data. **However, given the diversity of sources of information supplying the PRC database, an in-depth study of the incidents recorded by each source was carried out prior to the adjustment of the law. This should make it possible to limit the bias that could be generated by the multitude of sources feeding the database.** Below is the distribution of the logarithm of the sizes of violations by source of information in each of the sectors of activity considered:

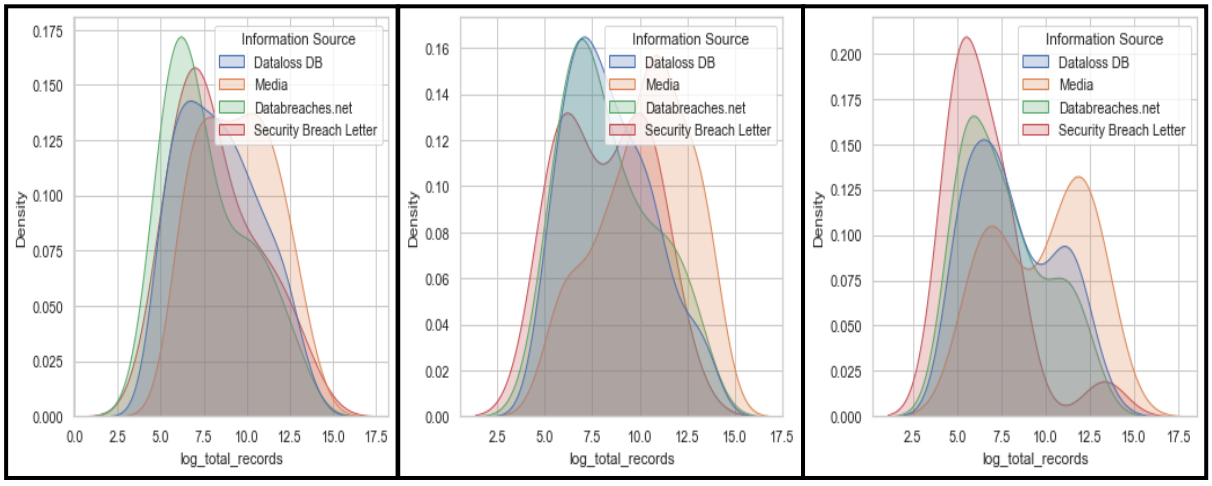


Figure 10 : BSF *Figure 11 : BSO* *Figure 12 : BSR*

A limit of 100 compromised data and a compensation cap of 50,000 data were also applied to the data studied.

Severity in cost

In order to associate the average number of data breaches in each sector with a cost, it was first necessary to estimate the unit cost of data. The study "The cost of data breach" published by the Ponemon Institute, examines the amount of expenses generated by the loss or theft of personal data. In the 2019 edition of this study, it was possible to extract the unit cost of a data breach per sector. The estimate of this cost takes into account:

- The cost of notifying the persons concerned of the incident

- The cost of detecting and managing the incident

- Post-breach costs

- Costs associated with business interruption

It was mentioned in this report that business interruption costs represent 38% of the average total cost of a cyber incident. As business interruption is a separate cover in the insurance package offered in this thesis, the business interruption costs will be deducted from the average cost obtained.

The average cost of a claim for data breach cover has been calculated using the formula below:

$$Cost_{\,average} = Nb_{moy} * unit_{moy}$$

With:
Nb_{moy}: the average number of compromised data
$Unit_{moy}$: the average unit cost of a data breach

As the available cost per data is from the 2019 edition of the report, it should be updated to provide a current view of the severity of data breach incidents. Various approaches can be used to update the unit cost of data. A first idea would be to use the annual inflation rates between 2019 and 2022 or to do a simple translation between the total average costs of data breach incidents provided in the 2019 and 2022 Ponemon reports.

For the study carried out here, the second approach will be preferred as it seems more relevant. Indeed, the calculation of the annual inflation rate is based on factors such as the consumer price index or the average household expenditure. These elements all appear to be fairly uncorrelated with the evolution of the average cost of data. In addition, several factors other than inflation can influence the variation in the average unit cost of a compromised item. For example, the evolution of the security techniques implemented or the improvement of the response mechanisms to cyber incidents.

Frequency modelling

There are reasons to believe that the PRC database is not adequate to model loss frequency rigorously. Indeed, the diversity of the information sources feeding the database does not allow to distinguish between the evolution of claims caused by a real increase of the risk and that coming from the instability of these information sources. The following graph shows the evolution of the number of incidents reported by the different information sources in the database:

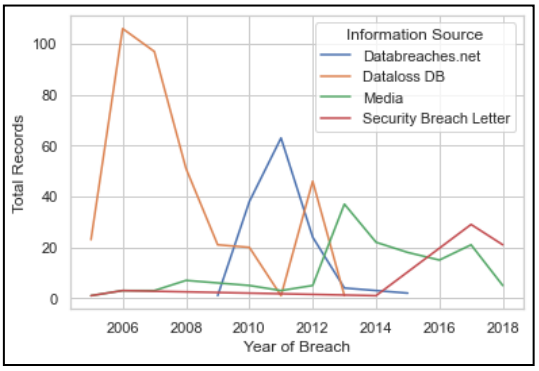


Figure 13 : Evolution of the number of recorded incidents by information source per year

In view of the significant imbalance observed in the PRC database, it can be assumed that the data is zero-trended. This would mean that companies that have not had a cyber loss are not observed in the database. Then, to estimate the frequency of loss, a generalized linear model (GLM) taking the business sectors (BSR, BSO, BSF) as explanatory variables will be fitted to the data. This will make it possible to consider the specificities of each type of organization and thus to obtain a frequency more adapted to each sector. Three types of GLM will be tested on the data: a GLM based on a 0-truncated Poisson distribution, a 0-truncated geometric distribution, and a 0-truncated negative binomial distribution.

Statistical criteria such as deviance and AIC will be used to distinguish between these three models and define the most predictive.

In a second time, in order to improve the accuracy of the estimates made, the use of a multinomial model will allow to quantify the impact of the different types of attacks on the frequency of claims.

5 Estimating the loss frequency of a "business interruption" policy

Business interruption insurance covers losses related to the decline or cessation of a company's economic activity as a result of a given event. In this case, the cover is for loss of business following a cyber-attack. Thus, for this coverage, the loss frequency will be approximated by the percentage of businesses that experience a business interruption or decline due to a cyber-attack.

Database

The data used for solving this problem is the Veris database. VERIS (Vocabulary for Event Recording and Sharing) is a set of metrics designed to provide a common language for describing cyber incidents in a clear, structured and repeatable way. This tool effectively addresses the lack of reliable information available on cyber risk and helps organisations to better measure and manage this risk. The database made available by this tool is quite rich with more than 8000 individual incidents recorded. It provides useful information on the characteristics of the incidents that occurred (actors, type of attack encountered, financial loss following the incident, response time, etc.). Numerous characteristics relating to the victims of the incident (country, number of employees, annual income etc.) are also included in this database.

Modelling

The modelling of the loss frequency of the business interruption cover will be done using Bayesian methods. These methods generally allow the probability of an event to be deduced based on the probabilities of other events that have already been estimated: the aim is to express what is known about the unknowns of the problem, knowing what is known (or assumed to be known). These methods are based on the notion of personal probability and their consistency is guaranteed by the mathematical rules of probability calculus. The principle of the Bayesian approach can be described using the diagram below

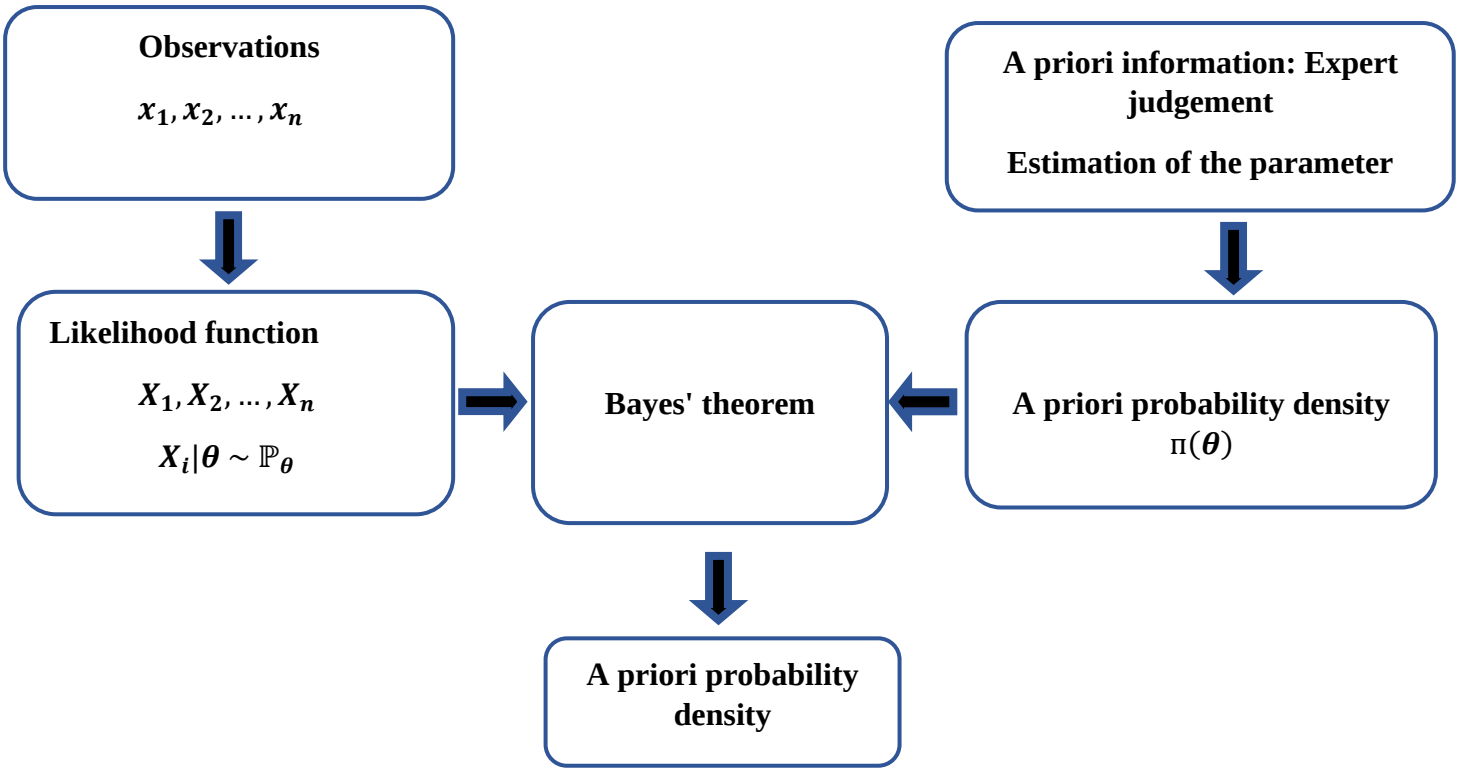


Figure 14 : Principle of the Bayesian approach

Approach 1 – Using an informative prior distribution

The first model will be based on an informative a priori distribution. An a priori distribution is said to be informative if it is not dominated by the likelihood and has an impact on the a posteriori distribution. Informative prior distributions need to be defined carefully in practice and prefer certain values of the quantity of interest. These preferences are usually based on previous studies.

Let θ be the probability that a company victim of a cyber-attack will suffer a business interruption. The estimation of the parameter θ will be done using the observation data of the Veris database. Let be the random variables $X_{i,i=1\ldots n}$ (n being the size of the database studied) designating the companies in the Veris database. The realizations of these random variables will be noted as x_i ; $x_i \in \{0,1\}$ being the state of company i after a cyber attack where $x_i = 1$ if the company suffers a business interruption et $x_i = 0$ if not. So : $X_i \sim$ Bernoulli (θ)

Choice of the prior distribution

Since there is little prior information available, one option for choosing the prior distribution is to use a conjugate prior distribution. As part of this study, $X_i|\theta \sim$ Bernoulli(θ), i=1,…n with n the length of the data set. The natural conjugate of Bernoulli's law being the Beta law, we have: $\theta \sim Be(\alpha, \beta)$. Then the density of the prior distribution is here:

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\,1_{[0,1]}(\theta)$$

From Bayes' formula :

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}$$

Therefore, the density of the a posteriori law is:

$$\pi(\theta|x) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}1_{[0,1]}(\theta)$$

In order to determine this estimator, the opinion of the expert's review should be included.

Elicitation of expert opinion

The experts that could be referred to in the context of the current problem are generally cyber security professionals. The opinion of these experts is required in order to obtain statistical information (mean, variance, etc.) on the parameter of interest θ.

Unfortunately, I was unable to obtain appointments with cyber security experts who could provide information on the θ parameter, the expertise will be based on the results obtained in the barometer of the cybersecurity of companies carried out by Opinionway for the CESIN[2]. From this study the expectation of θ and a 95% confidence interval for this value will be derived.

[2] Club des Experts de la Sécurité de l'Information et du Numérique

The combination of all this information will thus make it possible to obtain the parameters of the prior and posterior distribution and thus to give the Bayesian estimator of the parameter of interest θ. This estimator is defined as the mean of the posterior distribution. A sensitivity study of the Bayesian response to the prior distribution will then be performed.

Approach 2 - Using an uninformative prior distribution

This approach is based on the use of an uninformative prior distribution. The prior distribution obtained is then updated with observation data from the Veris database in a first step, which makes it possible to obtain a first posterior distribution. In a second step, the posterior distribution obtained will also be updated with information provided by an expert. This will make it possible to obtain a second posterior distribution.

1st step:

As we do not have precise information on the parameter of interest θ, an a uninformative prior distribution will be used. The non-informative law chosen here is the uniform law on [0,1]. Yet, $\theta \sim U(0,1)$. This makes it possible to obtain a beta law as an a posteriori law.

2nd step:

At this stage, the posterior distribution obtained previously becomes an prior distribution. This law is updated with the information provided by the expert.

Elicitation of expert opinion

For this method as well, it is assumed that the experts, who are generally cyber security professionals, have statistical knowledge (mean, variance, etc.) of the parameter of interest θ. The contacted expert can either be 100% sure of his estimate of the value of θ or define a probability distribution for θ which will consider his uncertainty about the value of the estimated parameter.

In general, the information sought is obtained by conducting surveys or interviews with several experts. Due to time constraints and the complexity of the risk under study, 95% confidence intervals on the value of the parameter θ will be randomly simulated and used as expert opinions. The statistical information collected from these fictitious experts will then be used to update the a prior distribution to obtain a new posterior distribution. Once the posterior distribution has been made explicit, the final estimate of θ will then be made using an average of the Bayes estimators obtained from the information of each expert.

Remerciements

J’adresse mes plus sincères remerciements à ma famille et à mes amis qui m’ont aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire et pendant mes années d’études.

Je remercie particulièrement Tidiane Cisse et Chadi Hanna, mes encadrants, pour leurs nombreux conseils ainsi que leurs suivis au cours de la réalisation de cette étude.

J’adresse également mes remerciements à tous mes collègues de Moonshot Insurance en particulier les équipes Marketing et Tech pour leur disponibilité et leur sympathie.

Enfin, je tiens à remercier l’ensemble des enseignants de l’ISUP, pour ces années très enrichissantes passées à leurs côtés, et à Olivier Lopez, pour ses précieux conseils et son expertise.

TABLE DES MATIERES

GLOSSAIRE

ACPR : Autorité de Contrôle Prudentiel et de Régulation

ANSSI : Agence Nationale de Sécurité des Systèmes d'Information

AMRAE : Association pour le Management des Risques et des Assurances de l'Entreprise

CESIN : Club des Experts de la Sécurité de l'Information et du Numérique

CNIL : Comission Nationale de l'Informatique et des Libertés

DDoS : Distributed Denial of Service

DoS : Denial of Service

EIOPA : Autorité Européennes des assurances et des pensions professionnelles

GLM : Generalized Linear Model

LUCY : LUmière sur la Cyber assurance

ONG : Organisation Non Gouvernementale

PCI DSS : Payment Card Industry Data Security Standard

PRC : Privacy ClearingHouse

PRO : Professionnel

PE : Perte d'exploitation

RC : Responsabilité Civile

RCP : Responsabilité Civile Professionnels

RGPD : Règlement Général sur la Protection des Données

RSSI : Responsable de la Sécurité des Systèmes d'Information

UE : Union Européenne

VCDB : VERIS Community Database

VERIS : Vocabulary for Event Recording and Sharing

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Depuis plusieurs années, la digitalisation est au cœur de la société. L'innovation technologique, les réseaux sociaux, la data, le cloud, l'intelligence artificielle ou encore la réalité virtuelle accompagnent aujourd'hui de nombreuses entreprises dans leur tâches quotidiennes. De plus, avec l'avènement de la crise sanitaire, elles ont pour la plupart, intensifié leurs projets de numérisation et dans certains cas entamé une transformation totale de leurs services et processus. Selon une récente étude du cabinet McKinsey, la crise sanitaire a accéléré la transformation digitale des entreprises d'environ 7 ans. (McKinsey & Company, 2020). Ainsi, les entreprises sont très dépendantes des nouvelles technologies informatiques et de l'internet.

Cette numérisation croissante les expose cependant à différentes formes de cyber-attaques. Le risque cyber se définit comme étant tout risque émanant de l'utilisation des données électroniques et de leur transmission, y compris des outils technologiques comme internet et les réseaux de communication (CRO Forum, 2014). Les attaques informatiques connaissent aujourd'hui une expansion notoire et sont difficiles à maitriser par les entreprises. A titre d'exemple, le nombre de cyberattaques a été multiplié par 4 en 2020 par rapport à 2019 (ANSSI, 2020).
De plus, le Rapport d'Information de Sébastien Meurant et Rémi Cardon tenu au Sénat le 10 juin 2021 révèle que 43 % des PME françaises ont constaté au moins un incident cyber au cours de l'année 2020 (MEURANT & CARDON, 2021). Le risque numérique est donc plus présent et alarmant au sein des PME/TPE.

Ce déferlement des cybercriminels sur les PME provient, du manque de ressources financières et humaines suffisantes, consacrées à la cybersécurité au sein de ces entités. En réalité, ces dernières préfèrent le plus souvent concentrer leurs fonds sur leur cœur de métier et reléguer la sécurité informatique au second plan. Une étude menée par l'Ifop pour la société de conseil en cybersécurité F-Secure révèle que, pour 60% des PME, le budget pour la protection de leur système informatique ne devrait pas excéder 1.000 euros par an. (LESAFFRE, 2021).

De même, de nombreuses PME vendent leurs services à des grandes entreprises. Les pirates informatiques utilisent alors fréquemment ces fournisseurs tiers comme un tremplin vers leurs principales cibles. A titre d'illustration, fin 2013, la chaîne américaine de magasins Target avait été piratée par le biais d'un sous-traitant en charge de la climatisation.

Plus exposées et vulnérables face aux attaques cyber, les PME sont aussi beaucoup plus démunies pour en affronter les conséquences. Bien souvent, les cyberattaques sont désastreuses pour ces petites structures et permettent de constater l'absence de backup fiable ou de plan de reprise. Elles sont pour la plupart paralysées et subissent d'importantes pertes. D'ailleurs, aux Etats unis, 60% des PME victimes de cyber attaques déposent le bilan dans les six mois. (Shepherd, 2020).

Face à de telles vulnérabilités, il devient de plus en plus essentiel pour les PME exposées d'investir dans leur cyber sécurité. Cependant, d'après le dernier rapport de l'AMRAE, malgré leur exposition aux attaques cyber, seules 0,2 % des TPE et PME sont couvertes face à de tels risques en France contre 84 % pour les grandes entreprises.

Il existe aujourd'hui sur le marché plusieurs solutions pour protéger les PME contre les attaques cybers, parmi lesquelles figure la cyber assurance. Afin de renforcer sa proposition de valeur, Moonshot Insurance a entamée des réflexions sur la création d'une police cyber-pro vis-à-vis de ses partenaires et plus généralement des TPE/PME. Ainsi, la construction d'un produit d'assurance cyber à destination des PME/TPE sera le sujet abordé dans le cadre de ce mémoire.

Quelles sont alors les principales étapes de la construction d'une offre d'assurance cyber ? Comment modéliser ce risque pour les PME/TPE étant donné les contraintes existantes ? (Manque de données, complexité du risque…).

Le présent mémoire tâchera de répondre aux problématiques évoquées ci-dessus en trois principales étapes :

La première partie sera consacrée à la présentation du risque cyber sous toute ses formes : définition, enjeux, type d'attaques cyber et exemple d'attaques cyber survenus. Un état des lieux du marché de l'assurance cyber en France sera également réalisé dans cette partie : garanties proposées, limites et taux de couverture.

La seconde partie du mémoire sera quant à elle, dédiée à la conception de l'offre proposée. Après la présentation d'un benchmark sur les différents acteurs présents sur le marché de l'assurance cyber, les caractéristiques de l'offre, les garanties retenues ainsi que les services de prévention à inclure seront présentés.

Enfin, en troisième partie, sera réalisée, après présentation du concept de tarification en assurance non-vie, la tarification de la garantie *« atteinte aux données »* incluse dans l'offre d'assurance proposée. Cette tarification a été basée sur la méthode classique de fréquence coût utilisé en assurance non-vie.
Dans cette partie sera également présentée la première étape de la tarification d'une garantie de « *perte d'exploitation* ». La fréquence de sinistre liée à une garantie d'interruption d'activité survenant à la suite d'une attaque cyber sera estimée à l'aide de deux méthodes d'inférences bayésiennes.

1 MOONSHOT INSURANCE

Initié au cœur d'un projet d'intrapreneuriat au sein de Société Générale Assurances, Moonshot Insurance voit le jour en 2017 avec pour objectif de construire un nouveau modèle d'assurance simple et créateur de valeur. Grâce à l'innovation technologique, l'insurtech propose des offres d'assurances basées sur l'utilisation d'API (interface de programmation d'application), incluant une souscription fluide et transparente ainsi qu'une couverture produit innovante.

Co-construite en quelques semaines seulement, les solutions de Moonshot Insurance sont nativement digitales et se différencient grâce à des modalités d'indemnisation automatique, en s'appuyant sur une gestion personnalisée et sécurisée des données.

En 5 ans, l'insurtech a bien grandi : de 6 collaborateurs à ses débuts, elle en compte désormais une trentaine. Moonshot Insurance se positionne comme un des principaux acteurs européens de l'Insurance-As-A-Service : ses solutions d'intégration en marque blanche en font un partenaire-clé des distributeurs souhaitant répondre aux besoins de protection de leurs clients, encore accrus par la crise du Covid. Ses produits sont majoritairement distribués en France et en Allemagne grâce à ses partenaires, et bientôt en Italie preuve de l'accélération du développement de ce modèle unique.

Et le succès est au rendez-vous ! En 2021, le cap du million de contrats a été franchi, illustrant la pertinence de l'assurance contextuelle : une proposition de valeur qui associe une expérience client intuitive basée sur l'immédiateté. L'insurtech de Société Générale Assurances continue de déployer de nouvelles solutions sur des gammes de plus en plus larges : l'événementiel, le voyage, les services financiers, la mobilité et l'e- commerce.

Moonshot Insurance est à son origine la pionnière de l'assurance contextuelle en Europe. C'est désormais une référence de l'assurance 2.0. et sa signature **« Tomorrow's Insurance, Today »** affirme la volonté d'offrir dès aujourd'hui l'assurance de demain. C'est ainsi qu'en 2021, Moonshot Insurance a développé la toute 1ère assurance préventive contre le cyber-harcèlement ainsi que l'une des premières assurances dédiées à la Multi-mobilité. Ces innovations ont été grandement récompensée. Cette année, c'est sur le secteur du tourisme que l'insurtech a décidé de frapper un grand coup en annonçant à la rentrée la 1ère assurance voyage 100% automatique.

2 LE RISQUE CYBER

2.1 CONTEXTE

La digitalisation peut être définie comme un procédé de transformation des processus traditionnels, des objets, des outils ou des professions par un recours accru aux nouvelles technologies. (LAROUSSE, 2021)

Ces dernières décennies, avec le développement de l'intelligence artificielle et la découverte de nombreuses innovations technologiques, la digitalisation occupe une place de plus en plus importante dans la société. Aujourd'hui, que ce soit dans les ménages, les écoles ou dans la vie quotidienne, les outils numériques sont très présents et permettent une simplification des tâches. Dorénavant, plus besoin de se déplacer au lycée ou au collège pour obtenir ses résultats à un contrôle. Au sein de 99,4 % des collèges et lycées, les élèves ont la possibilité d'accéder à leurs notes via une plateforme numérique accessible avec internet. (INSEE, 2019)

Ce mouvement de numérisation croissante est également présent au sein des entreprises et conduit à transformer la manière dont celles-ci sont gérées, se créent et se développent. Ces dernières se servent aujourd'hui des nouvelles avancées et mutations technologiques dans leurs tâches quotidiennes, afin de gagner en performance, en rapidité et en efficacité. Désormais, l'échange, la collecte, l'analyse et le traitement des données au sein des entreprises sont fréquemment automatisés grâce à de nombreux outils digitaux.

De plus, des dispositions législatives comme la loi PACTE viennent encourager et encadrer légalement la mutation technologique des entreprises. L'une des principales mesures du Plan d'Action pour la Croissance et la Transformation des Entreprises est l'installation d'un guichet unique électronique. Ce dispositif permettra aux entreprises de réaliser de façon digitalisée et accélérée des formalités d'enregistrement et d'immatriculation. De même, le PACTE facilitera l'accès des entreprises à des financements diversifiés, leur donnant les moyens d'innover et d'activer leur transformation numérique.

La crise sanitaire actuelle vient, elle aussi, renforcer cet élan de digitalisation. En effet, avec le développement du télé travail, l'automatisation des tâches, ainsi que l'utilisation des services à distance, les entreprises ont pour la plupart accéléré leurs projets de transformation numérique et dans certains cas entamé une transformation totale de leurs services et processus. D'après une étude du cabinet McKinsey, la crise sanitaire a accéléré la transformation digitale des entreprises d'environ 7 ans. (McKinsey & Company, 2020)

Face à tous ces facteurs, il est difficile de nier aujourd'hui que les entreprises sont dépendantes du digital. Cependant, cette mutation technologique vient avec certains inconvénients. En effet, l'utilisation de nouveaux outils numériques pourrait entrainer des pertes d'informations et/ou de données au cours des échanges, du fait de la multiplication des canaux de transmission ou de la négligence des employés. Ainsi, il est judicieux au sein des entreprises, de structurer les échanges et de définir des canaux à privilégier pour les informations essentielles. De plus, le numérique expose les entreprises à des actes de malveillances qui peuvent être aussi bien internes qu'externes. A titre d'exemple, un employé qui a accès au réseau interne d'une entreprise peut éventuellement s'en servir pour altérer,

modifier ou voler les données confidentielles détenues par celle-ci. De ce fait, bien qu'étant bénéfique pour les entreprises, la digitalisation augmente malheureusement les risques d'attaques informatiques. En conséquence, le niveau d'exposition des entreprises au risque cyber ne cesse de croître.

Qu'est-ce que le risque cyber et quelles en sont les conséquences ?

2.2 DEFINITION

Une étude de l'**APREF** (Association des Professionnels de la Réassurance en France) définit le risque cyber comme étant « pour toute personne morale ou physique, ci-après désignée comme « l'entité », toutes atteintes à :

- des systèmes électroniques et/ou informatiques de production, d'exploitation, de gestion d'informations et de télécommunication sous le contrôle de l'entité ou de ses prestataires et/ou

- des données informatisées (personnelles, confidentielles ou d'exploitation) appartenant à ou sous le contrôle de l'entité, qu'elles soient transférées ou stockées chez elle ou chez ses prestataires consécutives à :

- un acte malveillant ou de terrorisme

- une erreur humaine, une panne ou des problèmes techniques

- un évènement naturel ou accidentel

Ayant pour conséquences :

- des dommages corporels, matériels, et/ou immatériels (frais ou pertes financières), subis par l'entité et/ou ses employés

- une mobilisation de ressources internes ou externes

- des dommages corporels, matériels, et/ou immatériels, frais ou pertes financières causés par l'entité à des tiers (y compris chaines logistiques / sous-traitants)

- une atteinte à la marque et/ou à la réputation de l'entité » (APREF, 2016)

Aujourd'hui, avec la digitalisation progressive et l'avènement de la crise sanitaire, la fréquence et la sévérité des attaques cyber ont considérablement augmenté. Le nombre de cyberattaques a été multiplié par 4 en 2020 par rapport à 2019 (ANSSI, 2020). A titre d'illustration, la plateforme d'assistance ***cybermalveillance.gouv.fr***, ayant pour but d'assister et d'informer les victimes de cyberattaques, a vu sa fréquentation augmenter de 155% en 2020. Cela représente plus de 10 000 entreprises venues y chercher de l'assistance à la suite d'une attaque informatique. (Cybermalveillance, 2021)

Le risque cyber connait ainsi un fort développement et devient de plus en plus difficile à maitriser du fait de son évolution continue. De nombreuses organisations considèrent d'ailleurs ce risque comme étant leur première menace. D'après la cartographie des risques réalisée par France Assureurs en 2022, les cyber-attaques majeures se placent au premier rang du classement des risques émergents à horizon 5 ans pour les sociétés d'assurance et de

réassurance (figure 1). Cette opinion est globalement partagée par les dirigeants du secteur depuis la première édition du baromètre en 2016.

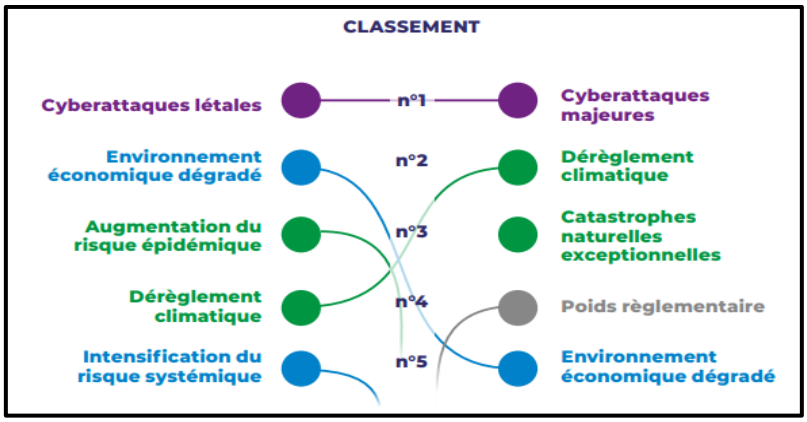


Figure 15 : Classement des risques émergents par France assureurs 2022, Source : (France Assureurs, 2022)

Si de nombreuses organisations sont craintives face à la menace cyber, c'est en grande partie dû à la complexité de celle-ci. En effet, il existe plusieurs types d'attaques cyber, ce qui rend difficile la mesure de l'ampleur de ce type de risque. De même, le risque cyber étant évolutif, aucun programme ou système de prévention ne peut garantir une protection efficace à 100% face à cette menace. Les types d'attaques et les moyens d'infiltration des systèmes informatiques sont en constante évolution. Chaque nouvelle attaque peut être totalement différente de la précédente.

2.2.1 Quelques types d'attaques cyber

Afin de mieux comprendre l'ampleur de ce risque, il est important de l'analyser sous toutes ses formes. Ci-dessous, une description de 5 différents types d'attaques cyber courants.

Attaque par déni de service (DoS) et attaque par déni de service distribué (DDoS)

Ces deux types d'attaques consistent à rendre partiellement ou totalement indisponible un serveur ou une infrastructure. Cela peut se faire soit par l'envoi de multiples requêtes jusqu'à épuisement de la bande passante ou par l'exploitation d'une faille de sécurité afin de provoquer une panne du serveur. Les hackers peuvent également utiliser plusieurs périphériques compromis pour lancer ce type d'attaque : ce sont les DDoS.

(Cybermalveilance.gouv, 2019)

Programme malveillant (malware)

Un ***malware*** est un logiciel indésirable installé dans le système d'un utilisateur sans son consentement. Le logiciel installé, peut se cacher dans un code légitime ou dans des applications. Un exemple de *malware* assez fréquent est le **Rançongiciel (ransomware)**. Comme son nom l'indique, le **Rançongiciel** est un code malveillant qui prend en otage les données ou les fichiers d'un utilisateur dans l'attente du paiement d'une rançon. En général, les hackers chiffrent totalement ou partiellement les données de l'utilisateur de façon à les rendre inexploitables, puis menacent de les supprimer ou les divulguer. L'appareil peut être infecté de différentes façons :

- Après l'ouverture d'une pièce-jointe frauduleuse ou d'un lien malveillant reçu par courriel ;
- Lors d'une navigation sur des sites compromis ;
- À la suite d'une intrusion informatique dans le système de l'entreprise victime. (Cybermalveillance.gouv, 2022)

Hammeçonnage (phishing)

L'hameçonnage est une technique frauduleuse consistant à envoyer des mails à un utilisateur en se faisant passer pour une source de confiance dans le but de collecter des données personnelles (comptes d'accès, mots de passe…) ou de l'inciter à une action. Ce type d'attaque peut se dissimuler dans une pièce jointe de mail, ou bien utiliser un lien pointant vers un site web illégitime pour inciter les utilisateurs à télécharger des logiciels malveillants ou transmettre certaines données personnelles. (Cybermalveillance.gouv, 2020)

Téléchargement furtif (Drive by Download)

Il s'agit d'une méthode populaire de diffusion de logiciels malveillants. Les cyberattaquants hackent des sites web non sécurisés en insérant un script dans le code http ou PHP d'une des pages web. Le but étant d'installer des logiciels malveillants directement sur l'ordinateur d'un visiteur du site, via un téléchargement furtif. Ce dernier peut se faire à l'insu du visiteur du site ou bien avec son consentement mais sans qu'il n'ait compris les conséquences : téléchargement de programmes malveillants ou simplement non désirés. (Netwrix, 2022)

Attaque de l'homme au milieu (MitM)

Il s'agit d'une technique de piratage visant à intercepter des échanges cryptés entre deux personnes ou deux ordinateurs pour en décoder le contenu. Le hacker doit ainsi réceptionner les messages des deux parties et répondre à chacune se faisant passer pour l'autre. De façon générale, le chiffrement et l'utilisation de certificats numériques permettent une protection efficace contre envoyer ce genre d'attaques. (Netwrix, 2022)

Il faut alors retenir que les différentes techniques d'attaques cyber sont toujours plus créatives et variées. Cependant, certaines sont plus fréquemment rencontrées que d'autres. D'après un sondage de « opinionway » réalisé auprès de 228 entreprises membres du CESIN[3] en 2022, le **Phishing** représente le premier vecteur d'attaque des entreprises. En effet, 73 % des entreprises déclarent avoir constaté une attaque de type **Phishing** au cours des 12 derniers mois. (Opinionway pour CESIN, 2022) (figure 2)

[3] CESIN : Club des Experts de la Sécurité de l'Information et du Numérique

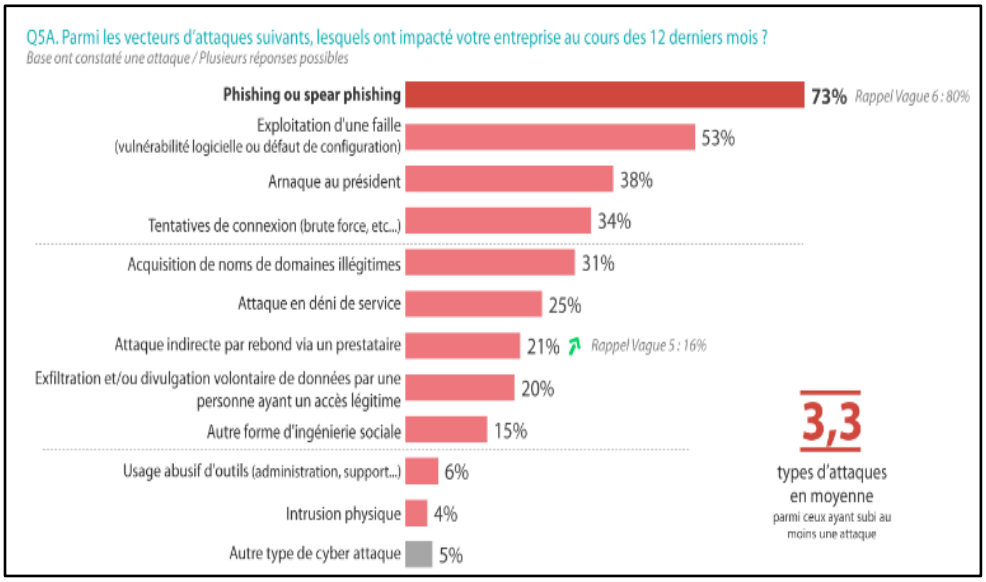


Figure 16 : Classement des différents vecteurs d'attaques des entreprises.
Source : (Opinionway pour CESIN, 2022)

Les conséquences de ces attaques informatiques sont de divers ordres. Comme mentionné dans la définition du risque cyber plus haut, une attaque cyber peut générer à la fois des dégâts sur le plan financier, réputationnel ou juridique. D'après l'étude réalisée par « opinionway », la principale conséquence des cybers attaques est l'**usurpation d'identité** juste avant le **vol de données** (figure 3)

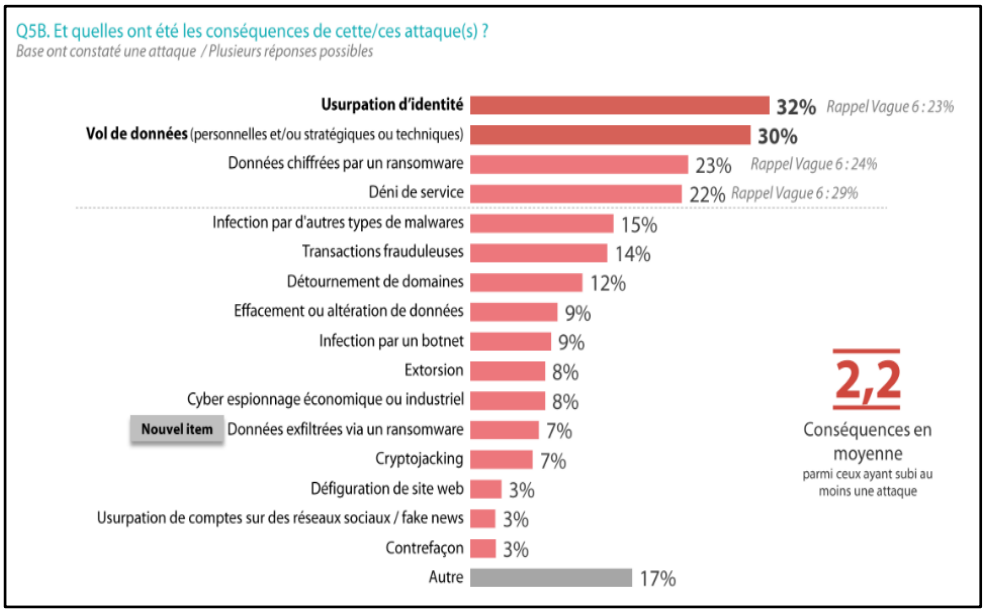


Figure 17 : Classement des conséquences des cyber attaques.
Source : (Opinionway pour CESIN, 2022)

A ce jour, plusieurs attaques informatiques ayant visé à la fois des grandes et des petites entreprises sont recensées. Ci-dessous quelques exemples d'évènements cyber enregistrés dans le monde :

2.2.2 Exemples d'événements cyber

Target (décembre 2013)

Fin 2013, la chaîne de magasin américaines Target a été piratée par le biais d'un sous-traitant. En effet, plus de 40 millions de clients du géant américain de la grande distribution se sont fait dérober les numéros et cryptogrammes de leurs cartes bancaires. Pour parvenir à leurs fins, les hackers avaient piraté un sous-traitant chargé des systèmes de chauffage et de climatisation. Lequel avait accès à un système de facturation lié au réseau interne du distributeur. Une fois dans la place, ils n'avaient plus qu'à installer leurs outils, faire quelques tests, puis attendre les fêtes de fin d'année, quand l'affluence leur assurait un maximum de victimes.

Wannacry (mai 2017)

Il s'agit ici d'une attaque de type ransomware. Cette attaque a infecté environ 300 000 appareils informatiques dans 150 pays et ses victimes se voyaient réclamer des rançons comprises entre 300 $ USD et 600 $ USD à régler en bitcoins. Les hackers ont exploité ici une faille de sécurité présente sur des systèmes non mis à jour, opérant sur des versions de Windows antérieures à Windows 10. Cette faille aurait été utilisé en premier par la NSA (National Security Agency) avant de tomber aux mains des hackers. Les premières infections auraient eu lieu en Espagne ou au Royaume-Uni avant de se propager dans le reste du monde en seulement quelques heures. A ce jour, le mode de transmission exact du virus n'est pas connu. Certaines hypothèses stipulent que le virus se serait auto répliqué. Cette attaque a eu de nombreux dégâts dans le monde. Au Royaume-Uni par exemple, dans le système de santé, plusieurs rendez-vous ont été annulés, des ambulances déviées, et quelques opérations annulées. De même, de grandes entreprises et organisations telles que Vodafone, FedEx, Renault, Telefonica et bien d'autres ont été lourdement affectées. La perte totale engendrée par cette attaque est estimée à 8 milliards de dollars tandis que le gain pour les auteurs de WannaCry est inférieur à 150 000 $ USD. Ce qui semble peu comparé au trouble provoqué dans l'économie.

AXA (mai 2021)

En mai 2021, Axa Partners, filiale d'assistance du géant de l'assurance AXA, a été victime d'une cyberattaque de type ransomware par le groupe de hackers Avaddon. Cette attaque a perturbé les opérations informatiques du groupe en Thaïlande, en Malaisie, à Hong Kong et aux Philippines. Plusieurs données sensibles telles que des passeports, des cartes d'identité, des contrats, des informations de comptes bancaires auraient été collectées par les pirates informatiques. Le groupe AXA aurait finalement payé la rançon réclamée par les hackers.

Ces différents exemples montrent bien que les attaques cyber sont de natures très diverses et peuvent provenir d'acteurs différents. Ainsi, il est assez complexe de quantifier les dégâts causés par ces dernières et d'y trouver des solutions efficaces. C'est pourquoi, il est important pour les entreprises de mettre en place des mesures de prévention afin de limiter les risques

numériques. Un moyen de prévention souvent utilisé de nos jours au sein des entreprises est la cyber sécurité. Qu'est-ce que la cyber sécurité ? Est-elle réellement efficace contre les attaques cyber ?

2.3 CYBERSECURITE ET PME

La cybersécurité est un néologisme qui désigne l'ensemble des outils et des processus de sécurité utilisés pour la protection de l'environnement numérique. Elle protège à la fois les personnes, les idées et les données. (l'internaute, 2021)

Il est recommandé au sein d'une entreprise de disposer d'un processus de cybersécurité fiable et performant. Cela permettrait en effet, de limiter les risques d'atteintes aux données, d'usurpation d'identité ou encore de piratage par rançongiciel. Intégrer un mécanisme de cybersécurité contribue également à la gestion de risques dans l'entreprise. En réalité, lorsqu'une organisation met en place une politique de sécurité réseau solide cumulée à un plan de réponse aux incidents efficace, elle est plus apte à prévenir les cyberattaques ou à en atténuer les conséquences.

Outre le fait de prévenir les cyberattaques, la cybersécurité permet aussi la mise en place auprès des collaborateurs, de processus d'instauration de bonnes pratiques. En effet, les erreurs humaines sont des sources réelles de fuites de données. Selon une étude de Verizon, 85 % des violations de la cybersécurité sont causées par une erreur humaine. (Verizon, 2021). Ainsi, la sensibilisation des équipes aux différentes problématiques de phishing ou d'usurpation d'identité est une composante importante d'un système de sécurité informatique efficace.

Il existe différents mécanismes de cyber sécurité parmi lesquels figurent :

- Les processus d'identification,
- Le chiffrement des données et des connexions,
- Les processus pour le contrôle et la mesure des mécanismes mis en place,
- La mise en place de dispositifs permettant la récupération rapide des données sensibles en cas de problèmes techniques, etc

Aujourd'hui, face à la recrudescence et à l'évolution des attaques cyber, les entreprises se soucient de plus en plus de leur sécurité informatique. Ainsi, elles sont nombreuses à investir dans leur cybersécurité. Le marché de la cybersécurité qui était estimé à 176.5 milliards de dollars en 2020 devrait atteindre d''ici 2027, une valeur de 403 milliards de dollars avec un TCAC (Taux de croissance Annuel Composé) de 12.5%. (CEPro, 2021).

Cependant, si la plupart des grandes entreprises ont conscience de l'importance d'investir dans leur cybersécurité, les PME quant à elles, restent encore en retrait. D'après une enquête réalisée par la CPME (Confédération des petites et moyennes entreprises), seules 38% des PME de moins de 50 salariés ont nommé un référent interne à la société en charge de la sécurité informatique. (CPME, 2019). De plus, une étude de BullGard, compagnie spécialisée en cybersécurité, révèle que 43 % des propriétaires de PME américaines et anglaises ne disposent d'aucun plan de défense en matière de cybersécurité. (Cision PRWeb, 2020)

Ces résultats ne sont pas surprenants dans la mesure où de nombreuses PME/TPE ont aujourd'hui une forte méconnaissance des risques encourus et sous-estiment leur attractivité pour les cybers attaquants. Elles pensent à tort que seules les grandes entreprises sont visées par les attaques informatiques.

Elles préfèrent alors pour la plupart, consacrer leurs budgets à leur cœur de métier et à d'autres tâches qu'elles jugent prioritaires. Au sein des PME/TPE, la sécurité informatique est donc le plus souvent reléguée au second plan principalement par manque de fonds. Un tiers des entreprises de moins de 50 employés déclarent utiliser une cybersécurité gratuite, de qualité grand public. (Cision PRWeb, 2020)

Ainsi, puisqu'elles ne bénéficient pas de protections adéquates, les PME représentent l'une des principales cibles des attaques informatiques. Le Rapport d'Information de Sébastien Meurant et Rémi Cardon tenu au Sénat le 10 juin 2021 illustre cette situation en révélant que 43 % des PME françaises ont constaté au moins un incident au cours de l'année 2020 (MEURANT & CARDON, 2021). Le risque numérique est donc bien présent au sein des PME/TPE. Plusieurs raisons pourraient également justifier ce déferlement des cybercriminels sur les PME. En effet, ces entités manquent pour la plupart d'expérience et d'accompagnement. Elles seront donc plus susceptibles de céder à un paiement de rançon ou à une cyber extorsion, ce qui fait d'elles des cibles attractives pour les hackers.

De plus, de nombreuses PME vendent des services à de grandes entreprises, les pirates informatiques utilisent donc fréquemment ces fournisseurs tiers comme un tremplin vers leurs cibles principales. L'exemple de Target, détaillé plus haut, illustre bien cette situation.

Plus exposées et vulnérables face aux attaques cyber, les PME sont aussi beaucoup plus démunies pour en affronter les conséquences. Bien souvent, les cyberattaques sont désastreuses pour ces petites structures et permettent de constater l'absence de backup fiable ou de plan de reprise. Elles sont pour la plupart paralysées et subissent d'importantes pertes. Selon le Ponemon Institute, les cyberattaques coûtent en moyenne plus de 2,2 millions de dollars aux PME aux Etats Unis (CyberCover). De plus, 60% des PME américaines victimes de cyber attaques déposent le bilan dans les six mois. (Shepherd, 2020). Ci-dessous, une présentation des différents coûts qu'une attaque informatique pourrait engendrer au sein d'une PME (figure 4)

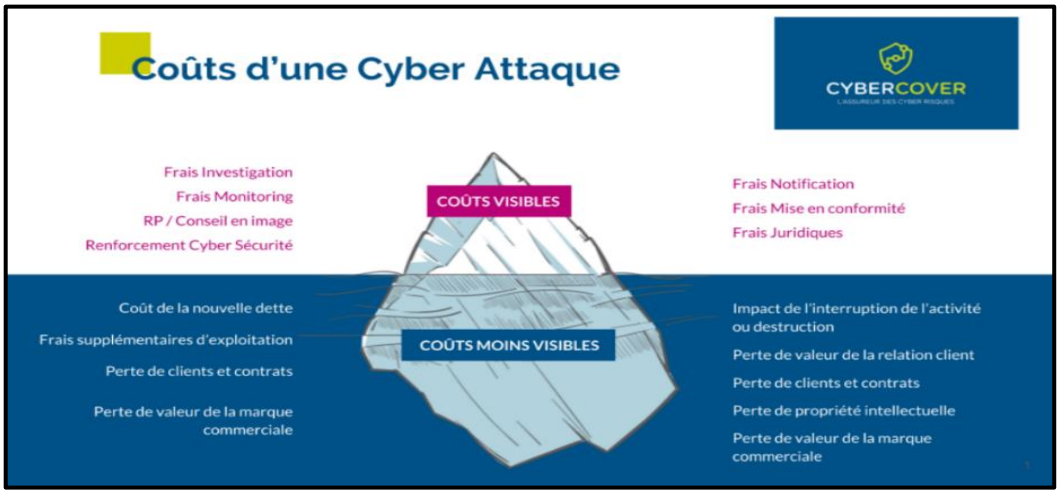


Figure 18 : Coûts d'une cyber attaque au sein des PME/TPE.
Source : (Cybercover)

Face à de telles vulnérabilités, il devient donc de plus en plus essentiel pour les PME exposées d'investir dans leur protection contre les répercussions des attaques informatiques. Il existe aujourd'hui sur le marché plusieurs solutions de défense des PME contre les attaques cybers, parmi lesquelles figure la cyber assurance.

Qu'est-ce que la cyber assurance ? En quoi permet-elle de protéger les PME contre la menace cyber ?

2.4 CYBER ASSURANCE

2.4.1 Définition et état du marché

Diverses mesures ont été mises en place pour aider les entreprises à faire face aux conséquences désastreuses des attaques informatiques. Aujourd'hui, de nombreuses plateformes d'assistance telles que *« cybermalveillance.gouv »* ou encore l'ANSSI sont mises en place pour guider les entreprises victimes de cyber attaques afin de limiter les dégâts. De plus, des experts en cybersécurité proposent leurs services aux entreprises victimes en vue de stopper la menace cyber ou de contre attaquer. Au nombre de ces solutions, figure également la cyber assurance.

« La cyber assurance est une forme d'assurance conçue pour protéger les entreprises contre les dommages causés par les menaces de cyber sécurité. » (Galea, 2021). Les premières offres de cyber assurance avaient été créées en 1990 pour le secteur bancaire. Désormais, avec l'évolution du numérique et la transformation digitale de la société, plusieurs contrats d'assurance cyber sont proposés sur le marché, à différentes catégories d'entreprises de secteurs variés. Aujourd'hui, toute entreprise s'appuyant sur les nouvelles technologies informatiques pour mener ses activités et qui conserve des données électroniques pourrait avoir besoin d'une cyber assurance.

Investir dans une assurance cyber pourrait être très bénéfique pour les entreprises puisque cela leur permettrait d'effectuer une cartographie précise des risques informatiques, afin de les rendre plus aptes à déterminer les actions à mener pour réduire leur exposition au risque. Aussi, en cas d'attaque avérée, seule une assurance cyber risques permet de couvrir les pertes financières et la responsabilité de l'entreprise vis-à-vis de ses clients et partenaires. Ce sont également les seuls contrats d'assurance à prendre en charge les erreurs non intentionnelles de manipulation.

Cependant, malgré l'importance de l'assurance cyber, plusieurs débats sont menés sur la question de la légitimité de l'existence d'un tel contrat d'assurance. Pour beaucoup, le risque cyber n'est pas assurable. En effet, pour qu'un risque soit assurable, il faut qu'il respecte certaines conditions nécessaires au développement d'un marché d'assurance sain et durable qui tiennent à la nature des risques couverts et aux comportements des assurés face à l'éventualité de leur réalisation.

Le chercheur Baruch Berliner explique dans son livre *Limits of insurability of risks*, que les critères d'assurabilité d'un risque peuvent être rangé en trois grandes catégories :

- Des critères actuariels : occurrence aléatoire / indépendance des risques ; perte maximale qui peut être évaluée et couverte, et perte moyenne par événement modérées; exposition au risque suffisante pour établir une base de données statistiques ; aléa moral et sélection adverse limités
- Des critères de marché : la prime d'assurance est jugée abordable par les prospects au regard de la couverture offerte
- Des critères sociétaux et réglementaires : restrictions légales doivent autoriser la couverture *(Le club des juristes , 2018)*

Le risque cyber présente certaines caractéristiques qui le placent à la frontière de l'assurabilité au regard de plusieurs des critères énumérés précédemment.

D'une part, de nombreuses attaques informatiques sont volontairement causées par des individus extérieurs. Les entreprises peuvent en effet, constituer des cibles emblématiques et ainsi être en permanence sujettes à des cyber-attaques. A titre d'exemple, certains cybercriminels privilégient parfois les attaques visant les TPE/PME car, leur système de cyber sécurité étant en général moins développé que celui des grandes structures, elles sont plus accessibles. De la même façon, d'autres cybercriminels préfèrent attaquer les grands groupes car ils estiment qu'ils ont plus de ressources à protéger et seront en mesure de payer des montants élevés en cas de ransomware. Voilà autant de raisons qui font remettre en question le caractère aléatoire du risque cyber.

D'autre part, comme spécifié dans les critères d'assurabilités cités plus haut, afin d'être en mesure de couvrir les pertes associées à un risque donné, il est nécessaire de pouvoir les estimer et les modéliser grâce à l'analyse de séries d'historiques d'événements passés. Dans le cas du risque cyber, il existe un manque de données qui représente un obstacle pour évaluer la fréquence et la sévérité des incidents. Les assureurs fondent ainsi leurs calculs actuariels sur des bases de données assez étroites et parfois tronquées. En réalité, ceci est dû au besoin des entreprises de préserver leur réputation et d'éviter les poursuites judiciaires en cas de violations des données personnelles. Ces dernières préfèrent parfois ne pas dévoiler publiquement les incidents cyber dont elles ont été victimes, ni le montant des pertes subies.

Par ailleurs, l'assurance repose en général sur la mutualisation des risques. Or, dans le cas des cyber-risques, du fait de l'interdépendance des systèmes informatiques, qui augmente les probabilités de propagation des incidents cyber, les assurés sont simultanément vulnérables en cas d'attaque sur l'un des systèmes. Ainsi, un virus informatique, peut se propager en s'autorépliquant dans un programme et infecter presque instantanément des dizaines de milliers d'ordinateurs. De ce fait, tous les assurés d'un même portefeuille peuvent être touchés en cas de cyber attaques. C'est l'exemple de l'attaque de WannaCry décrite plus haut.

Cependant, malgré les controverses autour de l'assurabilité du risque cyber, la cyber assurance constitue aujourd'hui un marché émergent. Selon une étude de l'AMRAE, le volume de primes a augmenté de 49 % entre 2019 et 2020 : il est passé de 87 millions d'euros à 130 millions d'euros. (AMRAE, 2021)

Toutefois, le marché de la cyber assurance en France présente encore de nombreuses faiblesses et plusieurs efforts restent à accomplir. En effet, les couvertures d'assurance

disponibles sont inégalement réparties sur le marché. Le marché est en grande partie porté par les grandes entreprises et peine encore à décoller en ce qui concerne les PME/TPE. D'après la dernière édition de l'enquête LUCY[4] menée par l'AMRAE, malgré leur exposition aux attaques cyber, seuls 0,2% des TPE et PME sont couvertes face au risque cyber en 2021 contre 84 % pour les grandes entreprises. Ces dernières représentent également 82% du volume des primes collectées au titre de la garantie cyber. (AMRAE, 2022)

De même, en raison de l'augmentation considérable du ratio Sinistres/Primes en 2020, (84% en 2019 contre 167% en 2020), les assureurs ont pour la plupart durci les conditions de souscriptions et élevé le niveau des primes. Chez les grandes entreprises par exemple, le taux de prime a été en moyenne multiplié par deux. (AMRAE, 2022). Cette hausse des taux de primes est cumulée à la mise en place de franchises très importantes. (Figure 5)

	Grandes entreprises	ETI	Moyennes entreprises	Petites entreprises	Micro entreprises
Franchise en 2021	3 990 104 €	227 976 €	32 217 €	7 670 €	995 €

Figure 19 : Montants de franchises mises en place par les assureurs.
Source : (AMRAE, 2022)

Ces différents changements entrainent une baisse globale des taux de couverture. Certaines entreprises renoncent même à souscrire à une assurance cyber et se tournent vers diverses solutions qui pourraient être moins couteuses telles que l'auto-assurance. Ci-dessous, un tableau récapitulant les taux de couvertures des entreprises en 2021.

	Effectif total en 2021 selon la typologie de l'Insee	Entreprises assurées en 2019	Entreprises assurées en 2020	Entreprises assurées en 2021	Croissance 2021/2020	Taux de couverture en 2021
Grandes entreprises (plus d'1,5 Md€ de CA)	287	207	251	240	-4,4 %	84 %
Entreprises de taille intermédiaire (50 M€ à 1,5 Md€ de CA)	5 763	307	441	530	+20,2 %	9 %
Petites et moyennes entreprises (10 à 50 M€ de CA)	139 971	311	362	322	-11 %	0,2 %
Petites entreprises (2 à 10 M€ de CA)	3 723 742 (Petites et micro)	616	643	503	-21,8 %	0,2 %
Micro entreprises (moins de 2 M€ de CA)	3 723 742 (Petites et micro)	7 025	7 027	10 433	32,6 %	0,2 %
TOTAL	**3 743 745**	**8 466**	**8 724**	**12 028**	**27,5 %**	**0,3 %**

Figure 20 : Taux de couvertures des entreprises en 2021
Source : (AMRAE, 2022)

Par ailleurs, les montants des garanties souscrites sont bien en dessous des réels besoins des assurés. Par exemple, en moyenne, les grands groupes étaient couverts à hauteur de 31 millions d'euros en 2021 pour un chiffre d'affaires annuel supérieur à 1,5 milliards d'euros. Les ETI quant à elles sont couvertes à hauteur de 6,5 millions d'euros en moyenne en 2021 ; ce qui reste très inférieur à leur réelle exposition au risque. (AMRAE, 2022). Ainsi, il y a sur le marché, un mismatch entre les attentes des entreprises et les capacités fournies par les

[4] LUCY : Lumière sur la Cyber assurance

assureurs. Cela ne permet pas aux entreprises de se protéger convenablement contre les risques informatiques auxquels elles sont confrontées. De ce fait, elles demeurent pour la plupart, vulnérables face à la menace.

Les spécificités du cyber-risque énumérées ci-dessus représentent autant de freins au développement du marché de l'assurance cyber. Néanmoins, l'expérience croissante des assureurs et le recours à la réassurance, notamment, permettront potentiellement une amélioration progressive des offres à destination des entreprises.

Les autorités Françaises et Européennes du contrôle prudentiel ont-elles aussi pris conscience des difficultés existantes sur le marché et ont ainsi suggérés la mise en place de différents mécanismes permettant aux assureurs une meilleur appréhension du risque.

Dans un communiqué de presse, l'ACPR[5] a relevé le manque de structuration du marché de l'assurance cyber, la trop grande variété des offres disponibles ainsi que le manque de données sur les cyber-risques. Elle a ainsi recommandé la création au sein de l'Union européenne, d'un mécanisme commun d'évaluation des offres de cyber-assurance et la standardisation des critères d'analyse des cyber-risques entre les assureurs, afin de créer une nouvelle branche d'assurance dédiée au cyber. (ACPR, 2019)

Le régulateur français a également publié en juin 2021, une liste de vingt-cinq recommandations afin de permettre une meilleure gestion des risques informatiques au sein des compagnies d'assurance et de réassurance. En effet, non seulement le coût des cyber-sinistres augmente, mais la menace pèse également sur de nombreuses entreprises, dont celles du secteur de l'assurance. Les assureurs peuvent eux même être des cibles d'attaques informatiques.

Par ailleurs, au sujet des faiblesses du marché de l'assurance cyber, le régulateur européen, l'EIOPA[6], suggère une amélioration de la lisibilité des contrats de cyber-assurance aussi bien pour les assurés que pour les assureurs. Il recommande également de veiller à faire de l'assurance un outil de prévention des risques cyber, d'accumuler plus de données relatives aux cyber-incidents et de fixer des règles de base et un vocabulaire commun à l'échelle européen sur les offres de cyber-assurance.

2.4.2 Exemples de garanties d'un contrat d'assurance cyber

La construction d'un contrat d'assurance exige au préalable une définition limpide des garanties qui y sont incluses. En assurance cyber, en raison du caractère évolutif du risque, il est en général complexe d'indiquer explicitement ce qui est inclus ou non dans le contrat. Les cyber assureurs présents sur le marché proposent plusieurs garanties pour répondre au mieux aux besoins spécifiques de leurs clients. Ci-dessous, une liste de quelques garanties pouvant figurer sur un contrat d'assurance cyber :

Assistance et gestion de crise : Au titre de cette garantie, l'assureur met à disposition de l'assuré, un expert informatique afin de déterminer la cause et l'étendue de l'attaque. Il

[5] ACPR : Autorité de contrôle prudentiel et de résolution.

[6] EIOPA : European Insurance and Occupational Pensions Authority

détermine également la capacité de l'assuré à éviter ce futur incident. Cette garantie prend également en compte, le remboursement des frais de notifications aux individus s'étant fait voler leurs données personnelles ainsi que les frais de surveillance sur internet ou sur le darkweb des apparitions des données personnelles qui ont pu être volées.

Perte d'exploitation : La garantie perte d'exploitation désigne la couverture des pertes liées à la baisse ou à l'arrêt d'activité économique due à un événement couvert. L'assurance prend en charge les pertes de revenus et dépenses supplémentaires au cours d'une période d'interruption de l'assuré lors d'un cyber attaque. Cette garantie ne fonctionne en général pas si le sinistre dépend d'une responsabilité envers un tiers.

Reconstitution de données : L'assurance prend en charge le remboursement des frais de restauration et/ou remplacement de données et/ou de logiciels volés, infectés, endommagés, cryptés ou encore supprimés.

Enquêtes administratives : L'assurance paie les frais de défense liés à la réclamation auprès de la CNIL lors d'une cyber-attaque.

Cyber Extorsion : Remboursement des rançons payées à la suite de l'endommagement, la destruction, le chiffrage, la modification ou la corruption du système informatique par un cyber-pirate. Il faut noter que cette garantie est généralement appliquée sous certaines conditions. D'une part, l'assurance ne fonctionne que si le tiers effectuant la menace d'extorsion n'est pas salarié ou dirigeant de la société souscriptrice, ou qu'il n'agit pas en collusion avec eux. D'autre part, la société doit être en mesure de prouver que le transfert d'argent s'est produit sous la contrainte. L'assureur doit également pouvoir avertir la police de la menace d'extorsion. Par ailleurs, l'assuré doit maintenir la garantie cyber extorsion confidentielle. Certains assureurs incluent dans leur offre une clause de suspension du contrat d'assurance si la garantie est révélée à un tiers.

La garantie cyber extorsion à pendant longtemps été un sujet de discussion auprès des assureurs/réassureurs et des dirigeants politiques. Pour les dirigeants, payer les rançons des cybers attaques alimenterait la criminalité en ligne et ferait des compagnies d'assurance des cibles emblématiques d'attaques informatiques. Ainsi, la député Valéria Faure-Muntian a suggéré l'inscription dans la loi de « l'interdiction pour les assureurs de garantir, couvrir ou d'indemniser la rançon ». (Faure-Muntian, 2021)

Cependant, l'article 5 du projet de loi orientation et programmation du ministère de l'intérieur, déposé le 16 mars autorise : *« Le versement d'une somme en application d'une clause assurantielle visant à couvrir le paiement d'une rançon par l'assuré dans le cadre d'une extorsion prévue lorsqu'elle est commise au moyen d'une atteinte à un système de traitement automatisé de données du même code, est subordonné à la justification du dépôt d'une plainte de la victime auprès des autorités compétentes au plus tard 48 heures après le paiement de cette rançon »*. (Assemblée Nationale, 2022)

Il existe également plusieurs autres garanties pouvant être incluses dans les contrats d'assurance cyber. Ci-dessous, une répartition des différentes garanties proposées au TPE/PME.

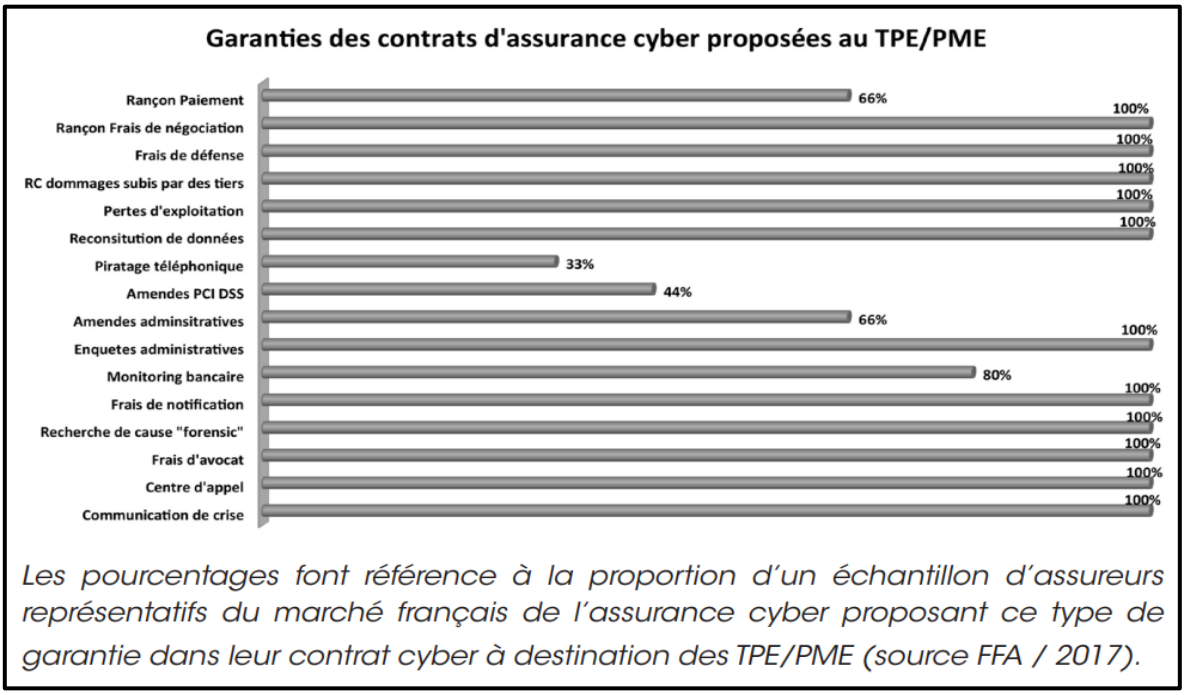
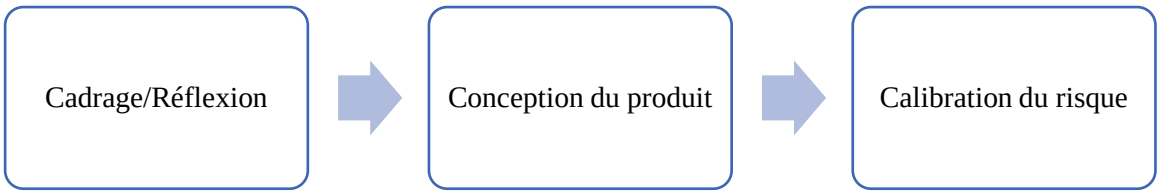


Figure 21 : Répartition des garanties figurant dans les contrats d'assurance cyber à destination des PME

3 CONCEPTION DE L'OFFRE

Un produit d'assurance sert principalement à se couvrir contre un éventuel danger inhérent à une activité ou à une situation et qui pourrait engendrer des coûts très importants. Ainsi, la conception d'un produit d'assurance nécessite minutie et rigueur afin de pouvoir analyser au mieux toutes les caractéristiques du risque à couvrir et de proposer les solutions les plus efficaces pour y faire face. Le processus de création d'une offre d'assurance se fait principalement en 3 grandes étapes :



- **Réflexion** : La phase de réflexion consiste principalement à s'interroger sur la pertinence de l'offre d'assurance proposée. En d'autres termes, il faudrait en premier lieu vérifier l'utilité de l'existence de ce produit, puis, définir les potentielles cibles. Il faudra également au cours de cette étape, réaliser une étude de marché afin d'évaluer les différents aspects des offres similaires disponibles.
- **Conception du produit :** La phase de conception du produit est l'une des plus importante de ce processus. C'est en effet au cours de cette étape, qu'une analyse approfondie du risque à couvrir est réalisée. Il sera nécessaire dans un premier temps de lister toutes les caractéristiques du risque étudié (élément déclencheur, conséquences, probabilité de survenance…), puis de définir les différentes solutions à mettre en place pour le limiter et/ou le couvrir. Au cours de cette phase, il faudra de même, construire une fiche produit claire afin de renseigner toutes garanties et services à inclure dans l'offre d'assurance et décider des différentes exclusions, limites et autres caractéristiques de l'offre.
- **Calibration du risque :** Il s'agit là de la phase finale du processus. Elle consiste à concrétiser toutes les idées soulevées en phase 1 et 2. Elle implique de nombreux de aller-retours entre les équipes marketing et actuariat afin de définir le tarif des différentes garanties à inclure dans l'offre. Cette phase donne également lieu à des ajustements du contenu de l'offre afin d'obtenir un équilibre garanties-prix. A la fin de cette phase, il sera possible de décider ou non de la viabilité du produit d'assurance à proposer.

3.1 SCAN DU MARCHE DE L'ASSURANCE CYBER ET CONSTRUCTION DE L'OFFRE

3.1.1 Construction d'un benchmark

L'objectif de ce mémoire est de proposer un produit d'assurance cyber à destination des PME/TPE de moins de 20 salariés et ayant un chiffre d'affaires inférieur à 10 millions d'euros. Comme stipulé dans les étapes de la conception d'un produit d'assurance présentées plus haut, il est important avant de définir les caractéristiques de l'offre, de réaliser une étude préalable des différents acteurs présents sur le marché de la cyber assurance afin d'analyser les caractéristiques des offres à destination des PME/TPE, de comparer le produit proposé avec ceux existants déjà sur le marché ou encore d'obtenir un avantage concurrentiel. Pour mener à bien cette étude, il convient dans un premier temps de recenser les différentes informations à récolter auprès des différents acteurs du marché étudiés. Ainsi, pour chaque acteur identifié, il faudra collecter les informations ci-dessous :

- **Date de création/ lancement de l'assurance cyber :** Cette information permettra d'identifier le niveau d'expérience de l'acteur étudié dans le domaine de l'assurance cyber et ainsi avoir une idée de la profondeur de l'historique de données disponible sur le marché. En général, les acteurs emblématiques du marché de la cyber assurance (AXA, Hiscox…) seront considérés comme référence.

- **Géographie adressée :** Cela permettra d'identifier s'il y a des garanties qui sont spécifiques ou non au marché de l'assurance français.

- **Taille (PME, ETI, Grandes entreprises…) :** Il est intéressant ici de regarder la taille des entreprises à inclure dans l'étude car cela permettra d'identifier le positionnement de Moonshot par rapport ces derniers.

- **Services de prévention :** La prévention étant essentielle pour limiter les risques d'attaques informatiques, il est important ici d'identifier les services de préventions mis en place par les différents acteurs afin d'évaluer ceux qui semblent permettre une forte réduction du risque numérique. Cela pourra également aider à choisir les services de prévention pertinents pour l'offre d'assurance proposée dans le cadre de ce mémoire.

- **Scope produit :** L'exposition au risque cyber est en partie dépendante de la taille économique (PME, ETI, grandes entreprises…) de l'entreprise. Ainsi, il est primordial dans une offre d'assurance de définir clairement la cible visée. Dans cette étude, l'essentiel des acteurs recensés proposent des produits d'assurance cyber à destination des PME/TPE.

- **Garanties proposées** : L'analyse des garanties proposées par les différents assureurs du marché permettra d'identifier les besoins des PME ainsi que les garanties qu'il serait pertinent d'inclure dans l'offre d'assurance proposée.

Les acteurs considérés dans l'étude réalisée sont pour la plupart des courtiers ou des insurtech. Cela permettra d'identifier le positionnement des concurrents et de construire une offre de produits et services différenciée. Il était nécessaire d'inclure également dans le benchmark réalisé, des concurrents de plus grande taille économique, proposant à la fois des produits à destination des PME/TPE, des ETI mais aussi des grandes entreprises. Ces derniers représentent en grande partie, des leaders mondiaux en matière de cyber assurance et des acteurs présents depuis plusieurs années sur le marché de l'assurance.

3.1.2 Présentation de deux acteurs du Benchmark

3.1.2.1 Stoïk

Créée en 2021, Stoïk se définit comme étant la première cyber-insurtech en Europe qui associe la cyber-assurance à un logiciel de cyber sécurité pour protéger les PME françaises contre les cybers attaques. Il s'agit d'une entreprise française proposant une assurance complète à destination des PME et TPE. (Stoïk, s.d.) Pour accompagner leur offre d'assurance, Stoïk propose une multitude de services de prévention parmi lesquels :

- Des audits réguliers de la surface d'attaque de l'assuré grâce à un scan externe hebdomadaire
- Des tutoriels permettant aux entreprises de remédier aux vulnérabilités les plus fréquentes
- Une sensibilisation à la cybersécurité via des campagnes de simulation de phishing

Les principales garanties proposées par Stoïk sont :

- Conséquences d'une atteinte aux données (Démarches RGPD, Frais des notifications et de suivi, responsabilité civile…)
- Conséquence d'une attaque (remise en état du système informatique, accompagnement en cas de cyber extorsion…)
- Pertes d'exploitation (perte de marge brute d'exploitation, mesures prises pour limiter les dégâts d'une attaque…)
- Services d'urgence (assistance, mise à disposition d'experts techniques et juridiques…) (Stoïk, s.d.)

3.1.2.2 Hiscox

Hiscox est un groupe international d'assurance et un assureur spécialisé proposant un large panel de produits d'assurance à destination des particuliers et des professionnels à travers le monde. En 2019, Hiscox a lancé son produit « cyberclear » qui est une offre complète d'assurance cyber à destination des professionnels. CyberClear a pour objectif principal de protéger les professionnels des coûts liés aux cyber-attaques, menaçant leurs systèmes informatiques, données à caractères personnels et/ou sensibles en leur possession remettant en

cause le bon fonctionnement de leur entreprise. (Hiscox, s.d.). Les principales garanties proposées par Hiscox dans son offre d'assurance sont :

- Assistance (avocat, communication en crise, récupération des données…)
- Enquêtes et sanctions (Frais de défense, amendes et pénalités…)
- Dommages subis par l'entreprise (violation de données, interruption des activités professionnelles…)
- Dommages aux tiers (Transmission de virus, atteinte à la sécurité et à la confidentialité des données personnelles…)

Pour accompagner ces différentes garanties, Hiscox met en place une large gamme de service de prévention tels que :

- La formation des employés et des dirigeants à la cybersecurité grâce à la **Cyberclear Academy**
- Mise à disposition d'experts chargés d'analyser les systèmes informatiques et le niveau de performance des dispositifs de cyber sécurité de l'assuré
- Accompagnement stratégique et opérationnel d'experts en sécurité des systèmes d'information lorsqu'une intrusion est suspectée et/ou détectée.

Bien que les deux acteurs décrits ci-dessus diffèrent par leur taille et leur structure, ils proposent essentiellement les mêmes garanties. Cependant, Hiscox propose un plus grand nombre de garanties et de services de prévention que Stoïk.

L'intégralité du benchmark réalisé est disponible en annexe.

3.2 PRESENTATION DE LA FICHE PRODUIT

La ***fiche produit*** est un document visant à définir les grandes lignes d'un produit d'assurance : les garanties, le périmètre d'intervention, la franchise, le plafond d'indemnisation ou encore les modalités de distribution. Elle est essentielle pour l'actuaire produit dans le processus de tarification.
Après un scan du marché de l'assurance cyber et discussion avec des experts en cybersécurité, cinq garanties essentielles seront inclues dans la fiche produit du contrat d'assurance proposé afin de répondre au mieux aux besoins des PME.

3.2.1 Présentation des garanties retenues

3.2.1.1 Atteinte aux données

Comme son nom l'indique, la présente garantie peut être déclenchée en cas de destruction, perte, altération, divulgation ou d'accès non autorisé aux données personnelles confidentielles de l'assuré ou celles détenues pour un tiers résultant d'une erreur humaine ou d'une cyber-attaque. Ainsi, au titre de cette garantie, sont pris en charge par l'assureur les frais ci-dessous :

Frais de communication et de notification : il s'agit des frais permettant l'identification des personnes dont les données ont fuité, la collecte des informations utiles pour préparer la notification aux personnes concernées et/ou à l'autorité administrative compétente,

l'impression, l'envoi et/ou la publication d'éléments permettant de procéder à cette notification etc.

Frais de monitoring et de surveillance : permettant d'identifier et de contrôler toute utilisation inadéquate des données personnelles détenues par la société assurée dès lors que ces données renferment :

- Tout numéro permettant d'identifier directement ou indirectement un individu et susceptible d'être utilisé, en conjonction avec d'autres informations, pour les besoins de l'ouverture de comptes bancaires ou la souscription d'assurances ; et/ou

- Toutes autres données personnelles pour lesquelles la loi ou la réglementation en vigueur impose la mise en œuvre de cette surveillance.

Frais d'investigation numérique : ce sont les frais d'expertise et autre frais d'investigation servant à :

- Qualifier et identifier l'origine et les circonstances de l'évènement cyber

- Identifier les failles de sécurité existant dans le système informatique de l'assuré ainsi que les données compromises

- Analyser l'ampleur et les conséquences de l'évènement cyber

- Mettre en œuvre les actions permettant de limiter les effets de l'évènement cyber

Frais de décontamination et de restauration du système informatique : il s'agit des différents frais permettant de remettre en état le système informatique. A cet effet, l'assureur prend en charge les frais servants à :

- Nettoyer et décontaminer le système informatique endommagé

- Restaurer le système informatique

Frais de récupération et/ou de reconstitution des données : ils regroupent :

- Les frais de déchiffrement

- Les frais de récupération des donnée grâce à des sauvegardes existantes.

3.2.1.2 Cyber Vol

Cette garantie est déclenchée en cas de virement frauduleux d'une somme d'argent des comptes de l'assuré vers un compte tiers lors d'une attaque cyber et sans collaboration de celle-ci. A ce titre, l'assureur prend en charge le remboursement de la perte correspondant au montant en numéraire de la monnaie scripturale virée ou transférée des comptes.

3.2.1.3 Cyber extorsion

Cette garantie intervient en cas de menace d'extorsion par un cyber-pirate, aux fins d'obtenir le paiement d'une rançon suite à l'endommagement, la destruction, la modification ou la corruption du système informatique de la société assurée. En cas de survenance d'un tel évènement, l'assureur prend en charge le paiement de la rançon ainsi que les frais

d'accompagnement de la société assurée (les honoraires de consultants/négociateurs spécialisés ou de traducteurs/interprètes…)

3.2.1.4 **Responsabilité civile**

La garantie responsabilité civile prend en charge :

Frais de défense : Prise en charge des frais de défense engagés avec l'accord écrit de l'assureur préalablement, dans le cadre de toute enquête ou de toute réclamation à l'encontre de la société assurée à la suite d'un évènement cyber. Ils incluent :

- Les honoraires et frais d'enquête, d'instruction, d'expertise, de comparution, d'avocats dans la limite des tarifs prévue dans le contrat.

- Les frais de procédures judiciaires, administratives, arbitrales et d'exécution de décisions de justice ou de sentences arbitrales.

Transmission de virus et cyber-attaques contre le système informatique : L'assureur garantit les dommages immatériels causés à tous tiers ou préposés pendant la période d'assurance, dès lors qu'une réclamation a été formulée par eux à l'encontre de la société assurée, mettant en cause sa responsabilité, et résultant :

- de la transmission d'un virus depuis le système informatique de l'assuré
- de l'utilisation du système informatique de l'assuré par un cyber-pirate à des fins d'attaque par déni de service dirigée contre ledit tiers ou partenaire commercial
- de toute cyber-attaque dirigée contre le système informatique de l'assuré

Sanctions Pécuniaires : Prise en charge des sanctions pécuniaires prononcées à l'encontre de la société assurée par toute autorité administrative y compris les dommages-intérêts ou tout autre type de pénalité imposée notamment par la commission nationale de l'informatique et des libertés (CNIL) et pouvant être pris en charge par un contrat d'assurance conformément au droit applicable au contrat.

3.2.1.5 **Perte d'exploitation**

La garantie Perte d'exploitation prend en charge :

Perte de marge brute : En cas d'arrêt total ou partiel du système informatique de l'assuré, l'assureur prend en charge sa marge bute, après expiration d'une durée minimale d'interruption dit "délai de carence"

Une **perte de marge brute** est définie comme la différence, pour un exercice comptable donné, entre la valeur nette de vente ou de production et les charges fixes et variables d'exploitation. La perte de marge brute est calculée à partir des éléments suivants :

- L'analyse mensuelle du chiffre d'affaires de la société assurée et de ses charges fixes et variables d'exploitation sur les 12 derniers mois avant la survenance de l'évènement cyber
- Les prévisions de croissance raisonnablement atteignables et prenant en compte les éventuelles modifications de marché.

Frais supplémentaires d'exploitation : L'assureur prend en charge tous les frais supplémentaires d'exploitation que la société assurée doit raisonnablement et nécessairement engager dans le but de reprendre le plus rapidement possible ses activités professionnelles.

Dans le cadre du présent mémoire d'actuariat, seules la garantie **Atteinte aux données** sera tarifée.

3.2.2 Evénements couverts

Dans le cadre de l'offre d'assurance proposée ici, les principaux événements couverts ainsi que des exemples de scénarios associés sont présentés ci-dessous:

Rançongiciels : Un message apparait sur les ordinateurs connectés au système informatique de la société assurée, l'informant du chiffrage de ses données. Pour obtenir la clé de déchiffrement, l'attaquant exige le paiement d'une rançon

Perte de données : Un employé de la société assurée supprime ou altère par erreur des données clients

Hameçonnage/Phishing: Des employés de la société assurée ont été victimes d'emails malveillants visant à les inciter à communiquer leurs données personnelles (ex : mot de passe) entraînant un accès non autorisé à son système informatique

Attaque informatique par déni de service : Le fonctionnement du système informarique de la société assurée est partiellement ou totalement pertubé, paralysé, saturé ou à l'arrêt à la suite d'une attaque malveillante

Logiciels malveillants/ Malware : Lors d'une navigation sur internet de l'un des employé de la société assurée, son système informatique a été infecté par un logiciel informatique malveillant ayant pour effet de détruire ou altérer les données qu'elle détient

Afin d'avoir une meilleure compréhension des garanties inclues dans l'offre, un schéma reliant à chaque garantie du contrat, les événements couverts déclencheurs est présenté ci-dessous :

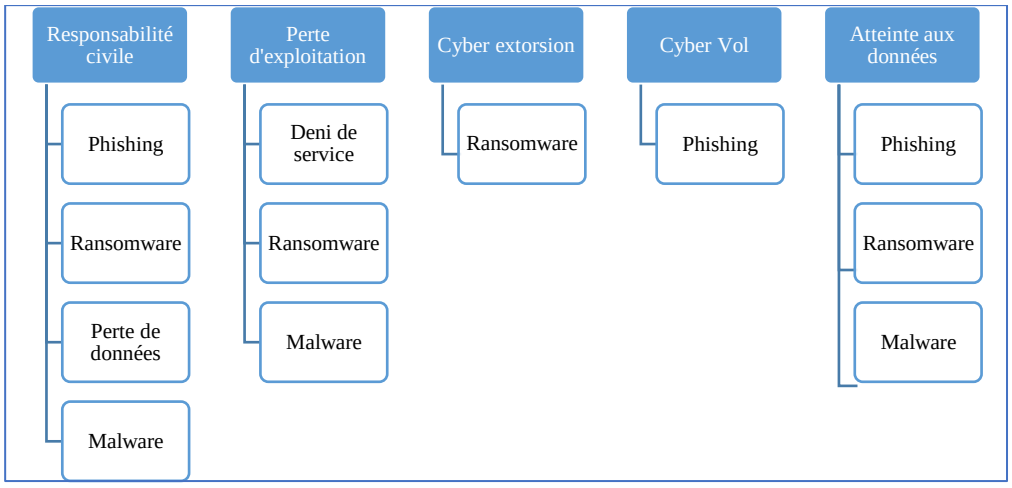


Figure 22 : Association événements couverts et garanties déclenchées

Les principaux secteurs d'activités non couverts par le contrat d'assurance proposé sont :

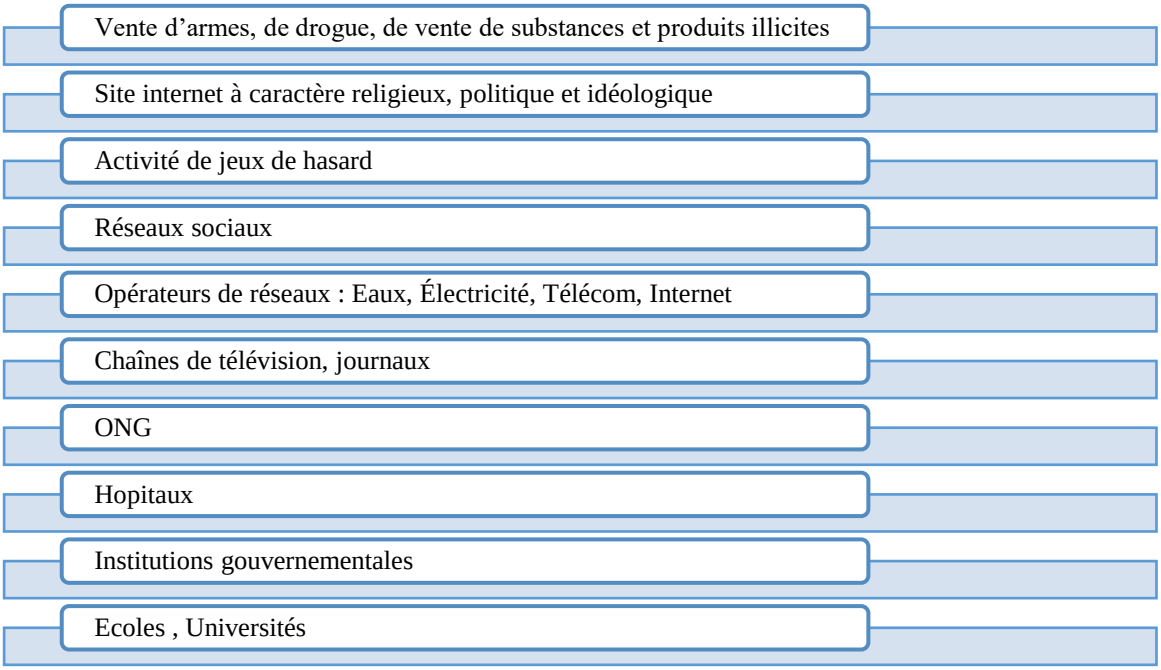


Figure 23 : Quelques secteurs exclus du contrat d'assurance

3.3 PRESENTATION DES SERVICES DE PREVENTION

Comme mentionné en section 2.1, la prévention est un facteur essentiel pour limiter les risques en assurance cyber. Ainsi, dans l'élaboration de ce projet, il est important d'accompagner toutes les garanties proposées de services de prévention adéquats. Les services de prévention à inclure dans l'offre d'assurance proposée seront répartis en trois grandes catégories. Il s'agit, des services axés sur la protection du système informatique de l'assuré, des services axés sur la protection des données clients et enfin des services permettant de limiter le risque d'erreurs humaines. Ci-dessous un récapitulatif de quelques services de prévention disponibles sur le marché :

3.3.1 Système informatique

- **Scan hebdomadaire (monitoring) :** Le monitoring ou la supervision informatique permet de détecter à l'aide d'un scan externe ou interne les failles de sécurité dans les sites web ou les applications web. Il est important d'effectuer ce scan de façon régulière et continue afin de corriger les vulnérabilités détectées avant que les hackers ne s'en servent.
  Cette mesure de sécurité présente cependant quelques limites. En effet, après discussion avec un expert en cyber sécurité, le scan ne permet en général pas de détecter les failles les plus importantes qui pourraient exister sur le site web.

- **Patch Manager** : Il consiste à automatiser les processus de détection, de suivi et de déploiement des mises à jour de sécurité logicielles. En effet, lorsqu'un éditeur comme Windows publie un nouveau patch de sécurité relatif à son produit, ses clients ne sont pas toujours en mesure d'évaluer l'importance de ce dernier ni les risques de son installation. Les solutions de gestion des correctifs proposent alors de stocker localement les correctifs sur un serveur du client, puis d'évaluer l'impact de celui-ci avant éventuellement de le tester puis de l'installer. (JDN, s.d.)
  Le patch manager est une technique de prévention efficace car elle permet de limiter les risques de malware et de s'assurer que les mises à jour logicielles sont faites en temps et en heure. Cependant, ce moyen préventif est assez coûteux et se destine en premier lieu aux grandes entreprises ou aux PME dotées d'un large parc informatique.

- **Antivirus** : Il s'agit d'un programme informatique ayant pour but d'identifier et de détruire un virus informatique. Bien que les antivirus ne garantissent pas une protection absolue contre les virus, il reste important de les installer sur tous les équipements informatiques, car ils contribuent fortement à être protégé des principaux virus connus. Il existe aujourd'hui sur le marché plusieurs antivirus différents, certains étant plus performants que d'autres. Le choix d'un antivirus dépend globalement du budget, des performances de la machine à protéger et des fonctionnalités de protection recherchées. Cependant, le point négatif de ce type de technologie est que son installation pourrait ralentir l'appareil lorsque celui-ci n'est pas très performant.

3.3.2 Données

- **Credential monitor** : Il s'agit d'un mécanisme permettant de tester les accès des collaborateurs de l'entreprise et ainsi de recevoir des alertes en cas de fuites de données sur le darkWeb impliquant leurs informations, accréditations ou mots de passe. Cette technique de prévention permet de limiter les fuites de données et de prévenir les piratages de mots de passe car exigeant que les collaborateurs changent régulièrement leurs mots de passe. Cependant, cette technique de prévention pose quelques problèmes lors de sa mise en place car elle nécessite que les collaborateurs de l'entreprise communiquent leurs différents mots de passe.

- **Authentification multifacteur** : Il s'agit d'un processus de sécurité qui nécessite deux ou plusieurs facteurs de vérification pour prouver l'identité d'un utilisateur. Le plus souvent, cela implique de ne plus se contenter que de la traditionnelle combinaison identifiant et mot de passe pour se connecter à un réseau, une application ou une autre ressource.
  L'utilisation de l'authentification multifacteur permet de limiter les risques de piratages informatiques. En effet, même si un pirate informatique parvient à se procurer l'identifiant et le mot de passe de l'utilisateur, il ne pourra potentiellement pas accéder à son compte, car n'étant pas en possession du second facteur d'authentification.

Après discussion avec un expert en cybersécurité de la société Oppens et le RSSI[7] de Moonshot Insurance, il faut retenir que l'authentification multifacteur est l'une des techniques de préventions les plus efficace contre la perte de données et les piratages. La CNIL recommande d'ailleurs d'activer l'authentification multifacteur chaque fois qu'un service le propose. Cependant, cette mesure de sécurité est assez souvent critiquée car étant assez complexe à mettre en place pour beaucoup d'utilisateurs. En effet, il est nécessaire pour accéder au service, de disposer obligatoirement des deux facteurs d'authentification. Ainsi, en cas de perte, vol ou dysfonctionnement du second facteur, l'accès au compte est temporairement impossible.
De plus, l'authentification multifacteur n'est pas infaillible face à certaines attaques complexes comme l'Hameçonnage.

3.3.3 Comportements humains

- **Formation des équipes** : Comme mentionné en partie 2, la négligence ou le manque de connaissances sont l'une des principales causes de la réussite des cyberattaques. Selon une étude de Verizon, 85 % des violations de la cybersécurité sont causées par une erreur humaine. Ainsi, la formation des employés aux problématiques de sécurité joue un rôle crucial dans le succès de tout programme de cybersécurité.

- **Campagnes de phishing factices :** Il s'agit ici, d'envoyer des mails de phishing fictifs aux employés de l'entreprise afin de d'évaluer ceux qui sont le plus à risque et qui pourraient plus facilement être victime d'une réelle attaque par phishing. Cela permettra également de vérifier la vigilance des employés et de tester l'efficacité des formations à la cybersécurité reçues.

- **Gestionnaire de mot de passe** : Un gestionnaire de mots de passe est une application utilisée pour stocker, générer et gérer les mots de passe d'un utilisateur. Les gestionnaires de mots de passe conservent ces mots de passe dans un format crypté et fournissent un accès sécurisé à toutes les informations de mot de passe à l'aide d'un mot de passe principal. Ces outils sont de véritable facteur de réduction du risque de piratage informatique car ils permettent de générer aléatoirement des mots de passe uniques, complexes et difficilement déchiffrables.

- **Experts en cyber sécurité** : Ce service de prévention consiste à mettre à disposition des entreprises assurées, des experts en cyber sécurité chargé de surveiller leur exposition au risque et d'intervenir pour répondre à leurs éventuels doutes ou questions. Ces experts seront également les premiers à agir en cas de survenance d'un incident afin de limiter les dégâts futurs.
  Cependant, le point négatif ici est que ces services nécessitent un partenariat avec une société de consulting en cyber sécurité, ce qui peut revenir cher autant pour l'assureur que pour les entreprises assurées. De même, il est assez complexe de vérifier qu'un expert est efficace à 100% pour répondre aux problématiques des entreprises.

[7] RSSI : Responsable de la Sécurité des Systèmes d'Informations

Au regard des informations collectées auprès des différents responsables informatiques concernant les limites de chacun des services de préventions présentés ci-dessus, seuls les services de prévention permettant une réduction importante du risque, faciles à mettre en place et peu chers seront retenus. Ainsi, les services de préventions sélectionnés pour le contrat d'assurance proposé dans le cadre du présent mémoire sont :

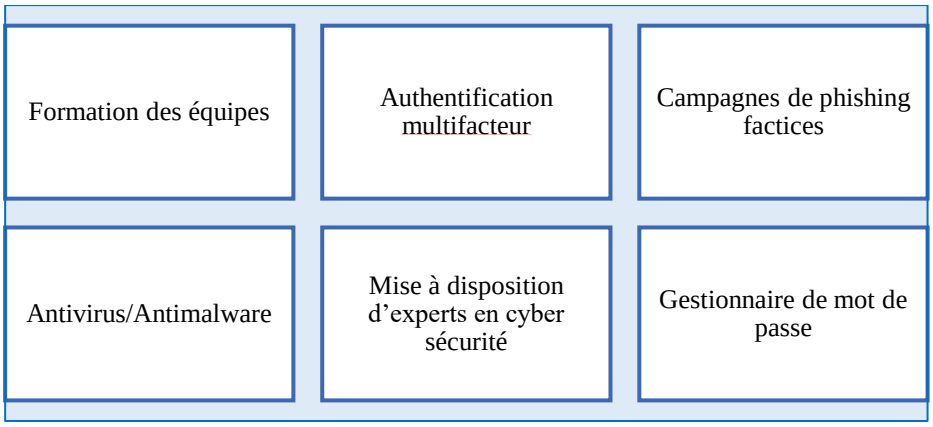


Figure 24 : Services de prévention

Il faut noter que, pour une meilleure gestion du risque et afin de réduire la sinistralité, l'utilisation de **certains services de préventions retenus sera une condition obligatoire à la souscription du contrat** d'assurance proposé. Les autres services feront l'objet de recommandation.

4 TARIFICATION DE LA GARANTIE « ATTEINTE AUX DONNEES »

4.1 TARIFICATION EN ASSURANCE NON-VIE : GENERALITES

La particularité du marché de l'assurance se trouve dans son ***cycle de production inversé***. En effet, habituellement dans un commerce, un produit destiné à la vente est fabriqué avant d'être commercialisé. Ainsi, le vendeur connaît son coût de production, tous les frais engendrés et par conséquent la marge bénéficiaire qu'il réalisera lors de la vente. En assurance en revanche, ce cycle est inversé : l'assureur s'engage à couvrir les pertes financières et/ou matérielles d'un risque dont il ignore la réalisation et le coût. Il devra donc estimer le montant de la couverture qu'il propose afin de déterminer la prime qu'il recevra en compensation. Le mécanisme de permutation du cycle de production se base principalement sur la ***mutualisation*** des risques au sein du portefeuille d'assurance. En effet, un coût insurmontable pour un assuré à lui seul peut être assumé par une collectivité.
Soit X_i le coût d'un sinistre pour un assuré i ($X_i = 0$ si le sinistre ne se produit pas).
Si la compagnie d'assurance possède n assurés dans son portefeuille, et si les X_i , $i = 1, \dots, n$ sont supposés indépendants et identiquement distribués, par la loi des grands nombres :

$$\frac{1}{n} \sum_i^n X_i \xrightarrow[n \to \infty]{} E[X] \text{ , presque surement (p.s)}$$

$E[X]$ étant l'espérance de la charge de sinistre.

Ainsi, si le nombre d'assurés n est suffisamment grand, la charge de sinistre pour l'assureur est approximativement égale à $n\,E[X]$. De ce fait, si chaque assuré paye individuellement une prime $\pi = E[X]$, la compagnie d'assurance devrait pouvoir faire face à ses engagements en cas de survenance d'un sinistre.
La prime d'assurance payée par l'assuré peut se décomposer en trois grandes parties à savoir :

- Le risque couvert, qui représente le montant du potentiel sinistre à assurer
- Les frais (gestion, généraux…) qui sont propres à chaque entreprise et permettent à l'assureur d'amortir ses différentes charges (salaires…)
- La marge de prudence, qui permet à l'assureur de se protéger lui-même contre le risque qu'il couvre.

Le calcul de la prime d'assurance se fait usuellement en deux grandes phases. Les actuaires déterminent en premier une **prime technique** ou **prime pure** puis, une **prime commerciale**.

4.1.1 Prime pure

La prime pure désigne la valeur actuelle probable de la charge totale de sinistre pendant la période de couverture définie dans le contrat d'assurance. Cette charge totale est en générale estimée à partir d'un nombre aléatoire de sinistres.
Soit N le nombre de sinistres et C_i le coût du sinistre i. En supposant que le nombre de sinistre N est indépendant du coût C_i des sinistres, eux même indépendants et identiquement distribués (i.i.d) entre eux et suivant une loi C, alors la prime pure se décompose comme suit :

$$Prime_{pure} = E\left[\sum_{i=1}^{N} C_i\right] = E[N]\ x\ E\ [C]$$

Cependant, l'hypothèse d'indépendance entre la fréquence de sinistre et son coût n'est pas toujours vérifiée en pratique. De même, d'après le théorème centrale limite, la prime pure calculée est insuffisante pour assurer l'équilibre au sein du portefeuille d'assurance à court-terme.

Théorème centrale limite

Soit $X_1, \dots, X_n$ une suite de variables aléatoires indépendantes et de même loi. En notant

$\mu = E[X_i]$ et $\sigma = V[X_i]$ respectivement l'espérance et la variance de X_i , si on pose :

$S_n = \sum_{i=1}^{n} X_i$, alors :

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\sigma^2})$$

$\mathcal{N}(0, \boldsymbol{\sigma^2})$ désignant une loi normale d'espérance nulle et de variance égale à $\boldsymbol{\sigma^2}$

Soit une compagnie d'assurance possédant n assurés et dont les sinistres $S_1, \dots, S_n$ sont i.i.d tels que $E[S_i] < +\infty$. L'assureur encaisse la prime μ.

D'après le théorème centrale limite, pour un nombre d'assurés n assez grand, la probabilité de perte est donnée par la formule :

$$Proba_{perte} = \mathbb{P}\left[\left(n\mu - \sum_{i=1}^{n} S_i\right) < 0\right] = \mathbb{P}\left[\sqrt{n}\left(\frac{S_n}{n} - \mu\right) < 0\right] \approx \frac{1}{2}$$

Ainsi, pour se protéger contre le caractère aléatoire du risque qu'il couvre, l'assureur rajoute une marge de prudence à sa prime pure. Plus le risque sera volatile, plus la marge de prudence sera élevée.

Il existe plusieurs méthodes pour déterminer la marge de prudence :

- L'assureur peut choisir d'inclure un chargement proportionnel à la prime pure, ce qui donnerait une **prime pure sécurisée** égale à :

$$P_{sécu} = Prime_{pure}\ (1 + \boldsymbol{\partial})$$

Où $\boldsymbol{\partial}$ correspond à l'idée que se fait l'assureur de la volatilité du risque à couvrir.

- L'assureur peut également définir le chargement à appliquer en fonction d'une mesure de volatilité de la sinistralité comme l'écart-type ou encore la Value at Risk (VAR).

4.1.2 Prime commerciale

La prime commerciale se définit comme étant la prime pure sécurisée augmentée des différents frais inhérents au contrat. Il existe généralement 3 grandes catégories de frais liés au contrat d'assurance :

- **Frais d'acquisition** : ce sont toutes les dépenses générées par l'acquisition de nouveaux clients. Exemple : frais de marketing, frais de gestion de souscription…

- **Frais d'administration** : ce sont les dépenses issues de la gestion administrative du contrat. Exemple : frais de modification de contrat, encaissements de prime…

- **Frais de gestion de sinistre** : ce sont les dépenses issues de la gestion administrative des sinistres déclarés. Exemple : salaire des gestionnaires de sinistres…

La prime commerciale peut s'écrire :

$$Prime_{commerciale} = P_{sécu} + f_{acquisition} + f_{administration} + f_{gestion-sin} + marge_{assureur}$$

Il faut également noter que, plusieurs critères sont pris en compte dans la construction d'un tarif d'assurance. Ainsi, les futurs assurés sont généralement amenés à répondre à plusieurs questionnaires avant d'accéder à leur devis d'assurance. Dans le cadre de l'assurance cyber par exemple, les assureurs s'appuient sur des caractéristiques telles que :

- Le **secteur d'activité** : de façon générale, plus une entreprise dispose de données sensibles, plus elle est potentiellement à risque et ainsi sa prime d'assurance sera élevée.

- Le **chiffre d'affaires** : plus le chiffre d'affaires d'une entreprise est élevé, plus sa prime est élevée.

- Le **profil de risque de l'entreprise** : une entreprise disposant de service de prévention variés et adéquats est potentiellement moins exposée au risque cyber et ainsi sa prime sera moins élevée.

- Le **nombre d'employés** : Plus le nombre d'employés au sein d'une entreprise est élevé plus le risque humain est élevé et ainsi l'entreprise est plus exposée aux attaques cyber.

Tous ces critères permettent à l'assureur d'avoir une idée de la volatilité du risque qu'il veut couvrir afin de définir au mieux le niveau de la prime. L'assureur prend également en compte dans cette mesure de la volatilité du risque, deux phénomènes souvent rencontrés en assurance qui sont l'aléa moral et l'antisélection. Ces deux phénomènes mettent à mal les hypothèses prises en compte dans le calcul de la prime d'assurance et de ce fait mettent en danger l'équilibre au sein du portefeuille d'assurance.

4.1.3 Antisélection

Selon l'hypothèse d'unicité de la prime d'assurance, tous les assurés d'un portefeuille sont contraints de payer le même montant de prime, quel que soit leur profil de risque. Ainsi, les assurés moins risqués devront payer une prime trop élevée pour leur profil tandis que la mutualisation instaurée par l'assureur permettra aux assurés très à risque de bénéficier d'une prime d'assurance inférieure aux coûts qu'ils engendrent. La mixité du portefeuille permet néanmoins à l'assureur, d'obtenir un équilibre au moment où la tarification est faite : l'excédent de prime payée par les assurés moins risqués sert à financer les pertes crées par les plus à risques.

Cependant, la concurrence existante sur le marché de l'assurance pourrait entrainer un déséquilibre du portefeuille. En effet, si un assureur concurrent effectue une segmentation dans sa méthodologie de tarification, et propose alors des tarifs adaptés à chaque profil de risque, les assurés moins risqués se verront proposer une prime inférieure et seront alors tenté de résilier leur contrat d'assurance avec le premier assureur. A l'inverse, les assurés les plus à risques se verront proposer un tarif plus élevé par la concurrence et seront en conséquence incité à souscrire à la police d'assurance à tarif unique. Il y aura donc une aggravation de la sinistralité au sein du portefeuille d'assurance risquant ainsi de nuire à l'équilibre financier de l'assureur.

Ainsi, pour limiter ce phénomène, les assureurs du marché appliquent presque systématiquement une segmentation de leurs portefeuilles. Etablir le profil de risque d'un assuré peut cependant s'avérer complexe étant donné que certaines informations concernant l'assuré sont légalement inexploitables par l'assureur. De même, une segmentation trop étroite pourrait empêcher la mutualisation des risques.

En assurance cyber, le manque de données et le caractère mouvant du risque pourraient nuire à la construction d'une segmentation efficace. Néanmoins, le secteur d'activité des entreprises, le nombre d'employés ou encore les moyens de prévention contre les incidents cyber mises en place par les entreprises semblent être des caractéristiques pertinentes sur lesquelles l'assureur pourrait baser sa segmentation.

4.1.4 Aléa moral

L'aléa moral désigne le phénomène selon lequel, une personne assurée contre un risque se comporte de manière plus risquée que si elle était totalement exposée à ce risque. La cause principale de l'aléa moral est l'asymétrie d'information : un assuré mal intentionné ou imprudent connait son statut alors que l'assureur lui, l'ignore. Ce phénomène pourrait tout comme l'antisélection nuire à l'équilibre financier de l'assureur car il pourrait entrainer une augmentation de la sinistralité au sein du portefeuille.

De nombreuses solutions sont mises en place par les assureurs pour limiter les risques d'aléa morale. Parmi elles, figure la mise en place des « bonus-malus » en assurance automobile ou encore l'instauration de franchises et de limite d'indemnisation dans les contrats d'assurance de même que les mécanismes de surveillance du portefeuille.

En assurance cyber, un moyen supplémentaire de lutter contre l'aléa moral pourrait être de rendre obligatoire dans les conditions de souscription, l'utilisation de services de préventions efficaces contre les incidents cyber. Cela entrainerait une responsabilisation des entreprises assurées et permettrait ainsi de réduire les conséquences des attaques cyber.

4.2 EXPLORATION DE BASES DE DONNEES

Mesurer l'exposition du risque cyber n'est pas chose aisée. En effet, le caractère complexe et évolutif du risque rend épineux la recherche de bases de données complètes et fiables. Certains assureurs présents sur le marché de l'assurance cyber, disposent de base de données internes. Ces bases peuvent être utiles pour mesurer l'exposition réelle de leur portefeuille à la menace mais pourraient être limitées pour mesurer l'exposition globale au risque. En réalité, une augmentation du nombre de cyber événements dans une base de données privée peut être dû à un changement de comportement des entreprises du portefeuille de l'assureur qui, décident de notifier plus d'incidents et non à une réelle évolution du risque.

Un complément à l'utilisation de bases de données privées est de s'appuyer sur des données publiques. Cependant, il peut également être difficile de mesurer l'exposition au risque cyber en se basant sur ces dernières car n'étant pas toujours fiables, ni bien alimentées. Dans le cadre de ce mémoire, plusieurs bases de données ont été explorées. Moonshot insurance étant nouveau sur le périmètre de l'assurance cyber pour les professionnels, l'assurtech ne possède pas de données internes à disposition.

Ci-dessous une présentation des bases de données publiques explorées :

4.2.1 VCDB (VERIS Community Database)

VERIS (Vocabulary for Event Recording and Sharing) désigne un ensemble de métriques conçues dans le but de fournir un langage commun permettant de décrire les incidents cyber de façon claire, structurée et reproductible. Cet outil permet de répondre efficacement au manque d'informations fiables disponibles en matière de risque cyber et aide les organisations à mieux mesurer et gérer ce risque.

La base de données mise à disposition par ce dispositif est assez riche avec plus de 9000 incidents individuels recensés. Les sources d'informations de ces incidents comprennent, les départements de santé américains (HHS : Health and Human Services), les sites internet des différents procureurs généraux d'Etats américains qui fournissent des informations concernant des incidents qui leur sont rapportés, les rapports des médias ainsi que les communiqués de presse.

La communauté Veris, a pour objectif d'accroitre et de diversifier ses sources d'information afin d'avoir une base qui soit la plus complète possible. Cependant, les données actuelles répertorient en grande partie des incidents survenus dans le domaine médical. Néanmoins, la base de données fournit plusieurs informations utiles sur les caractéristiques des incidents survenus (acteurs, type d'attaque rencontré, perte financière suite à l'incident, délai de réponse etc). De nombreuses caractéristiques relatives aux victimes de l'incident (pays, nombre d'employés, revenu annuel etc) figurent également dans cette base de données.

La base est accessible via le lien suivant : http://veriscommunity.net/index.html/

Dans le cadre de la problématique étudiée dans cette section, cette base de données ne sera pas exploitée. En effet, plusieurs valeurs sont manquantes en ce qui concerne le coût financier de l'incident. L'utilisation de cette variable est donc insuffisante pour mesurer la sévérité des incidents cyber.

Néanmoins, la base Veris sera exploitée dans le cadre de la modélisation du risque de perte d'exploitation suite à une cyber attaque.

4.2.2 Hackmageddon

Il s'agit d'une base entamée en 2011 et construite sous l'initiative d'un ingénieur en sécurité informatique. La base est mise à jour mensuellement et recense de nombreux incidents cybers survenus dans le monde.

Elle est accessible via le lien suivant : https://www.hackmageddon.com/. Cette base ne sera pas exploitée dans la résolution de la problématique de ce mémoire car, étant constituée à la main, elle peut être fortement biaisée.

4.2.3 World's biggest Data Breaches & Hacks

Il s'agit d'un organisme à but non lucratif qui recense des incidents de fuite données survenus dans le monde entier. Plusieurs informations sont fournies sur les caractéristiques de l'incident et de ses victimes (secteur de l'entreprise, type d'attaque, nombre de données perdues, description de l'incident etc).

La base de données mise à disposition est régulièrement mise à jour et ne répertorie que des incidents dont le nombre de données compromises est supérieur à 30 000. Etant sur la construction d'un produit d'assurance à destination des PME/TPE de moins de 20 salariés, cette condition est trop restrictive. De ce fait, cette base de données sera écartée dans les analyses réalisées dans la suite de ce mémoire. La base est accessible via le lien suivant : https://informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

4.2.4 Privacy Rights ClearingHouse (PRC)

La Privacy Rights ClearingHouse est une organisation à but non lucratif fondée en 1992 et ayant à cœur la protection de la vie privée des citoyens américains. Elle répertorie depuis 2005 des incidents de fuites de données affectant des entreprises américaines. Dans cette base de données, plusieurs informations sur les caractéristiques des cyber événements reportés (localisation, date…) sont recensées. Plus de détails sur ces caractéristiques seront fournis dans la suite du mémoire. La base est téléchargeable via le lien suivant : https://www.privacyrights.org/copyright/

Des bases de données disponibles à la vente et fournies par des statisticiens et des experts en cyber sécurité ont également été explorées dans cette quête d'informations. Cependant, il est souvent difficile de savoir comment ces bases ont été construites et de quantifier le biais qu'elles pourraient contenir. De plus, il était compliqué d'obtenir dans les temps, des réponses de la part des entreprises commercialisant ces données.

4.3 MODELISATION DU RISQUE

La base PRC a été retenue pour la suite de ce mémoire car c'est la seule parmi toutes les bases explorées qui correspondait le mieux à la problématique étudiée avec un nombre de lignes pertinent.

4.3.1 Présentation de la base PRC

La base Privacy Rights ClearingHouse (PRC), utilisée dans le cadre du présent mémoire, est une base publique américaine répertoriant des incidents de fuites de données à caractère personnel aux Etats-Unis. Les incidents reportés ici, sont principalement ceux recensés par des agences gouvernementales et ceux médiatisés. Avec l'obligation de notification des vols de données personnelles aux Etats Unis, la base PRC est plus ou moins représentative du territoire américain. Elle est assez dense en termes d'observations avec plus de 9000 incidents recensés entre 2005 et 2019.

La base comporte les variables ci-dessous :

Date Made Public : Date à laquelle l'incident déclaré a été rendu publique

Company : Nom de l'organisation ayant subi la violation

City : Ville où se situe l'organisation ayant subi la violation

Etat : Etat dans lequel se situe l'organisation ayant subi la violation

Type of breach : Type de violations de données subi par l'organisation. Il s'agit d'une variable qualitative dont les modalités sont les suivantes sont résumées dans le tableau suivant :

Type d'attaque	Description
Payment Card Fraud (CARD)	Fraudes impliquant des cartes de paiement (débit ou crédit) sans utilisation de méthode de hacking.
Hacking or Malware (HACK)	Incidents de piratage informatique effectués en se servant de logiciels malveillants
Insider (INSD)	Incidents causés volontairement par des personnes internes à l'entreprise/l'organisation (employé, client, etc…)
Physical Loss (PHYS)	Incidents relatifs à des pertes physiques (non électroniques) de données. Ce sont par exemple, des documents ou papiers égarés, jetés ou volés.
Portable Device (PORT)	Il s'agit d'incidents relatifs à des appareil portables perdus, jetés ou volés. Exemples : ordinateur portable, téléphone mobile, CD, disque dur, clé USB…
Stationary Device (STAT)	Ordinateurs non portables ou serveurs, perdus, jetés, volés, ou dont l'accès s'est fait illégalement.
Unintended Disclosure (DISC)	Incidents de divulgation involontaire d'informations.
Unknown	Type de violation de données inconnu

Tableau 1 : Type d'attaques base PRC

Type of organization : Domaine d'activité de l'organisation ayant subi la violation. Les modalités de cette variable sont listées ci-dessous :

Type d'organisation	Définition
BSF (Businesses-Financial and Insurance Services)	Entreprises de services financiers et d'assurance
BSR (Businesses-Retail/Merchant - Including Online Retail)	Entreprise ayant une activité marchande ou de vente au détail, y compris une activité de e-commerce.
BSO (Businesses-Other)	Entreprises appartenant à d'autres domaines que ceux listés
EDU (Educational Institutions)	Etablissements d'enseignement
GOV (Government & Military)	Services médicaux
MED (Healthcare, Medical Providers & Medical Insurance Services)	Ordinateurs non portables ou serveurs, perdus, jetés, volés, ou dont l'accès s'est fait illégalement.
NGO (Non-Governmental Organization)	Organisation Non gouvernementale

Tableau 2 : Types d'organisations base PRC

Total Records : Nombre de données compromises

Description of incident : Description de l'incident

Information source : Source de l'information

Source URL : Lien vers le site internet d'où provient de l'information

Il faut noter que la base de données contient également des informations sur les coordonnées géographiques (longitude, latitude) du lieu de l'incident.

4.3.2 Retraitement de la base de données

Avant d'exploiter de la base PRC, il était important de la retravailler afin de la rendre cohérente avec la problématique étudiée. Les types d'organisations de la base ont été restreints aux secteurs BSF, BSO et BSR ; les autres secteurs étant exclus du produit d'assurance proposée ici. Pour l'analyse préliminaire, une franchise à 100 données compromises est appliquée aux données.

Par ailleurs, la distribution des tailles des violations semble être séparée en deux avant 1 000 000 de données compromises ((13.8 en log). Cela voudrait dire que les incidents correspondant à des megabreaches (ie nombre de données violées supérieur à 1 000 000), plus rares, ont une distribution différente des autres types d'incidents. Afin d'obtenir un meilleur ajustement des données, les megabreaches ne seront donc pas pris en compte dans l'analyse réalisée dans ce mémoire.

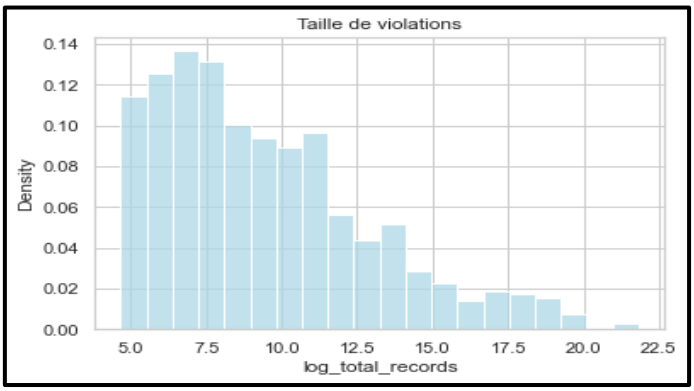


Figure 25 : Distribution log des tailles de violations

4.3.3 Analyses descriptives

Une fois le retraitement de la base effectué, une analyse descriptive des données a été réalisée. L'idée étant de pouvoir analyser au mieux les différentes caractéristiques des incidents cybers reportés dans la base et ainsi avoir une meilleure appréhension du risque étudié.

4.3.3.1 Variables relatives aux victimes

Ci-dessous une répartition des différents types d'organisations de la base de données :

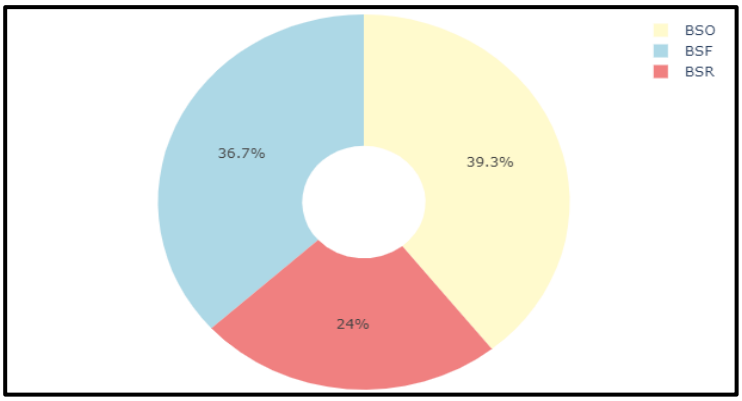


Figure 26 : Répartition des différents types d'organisations

Les différents types d'organisation sont réparties de manière plus ou moins homogène au sein de la base. Il n'y a pas de prédominance excessive d'un type d'organisation par rapport aux autres, même si, une différence d'environ 16% est observée entre les répartitions des entreprises BSO et BSR. Une hétérogénéité trop importante pourrait être source de biais pour l'analyse réalisée, il était donc important d'observer la répartition des différents types d'organisation au sein de la base de données.

La supériorité relative des incidents envers les entreprises du type BSO observée dans la base, pourrait provenir du fait que ce secteur est plus vulnérable face à la menace que les autres car il englobe un spectre plus large d'entreprises que les types BSF et BSR qui eux, sont plus spécifiques.
Afin de mieux comprendre les spécificités des incidents cybers recensés au sein de chacun de ces secteurs d'organisations, il faudrait une analyse détaillée des tailles de violations déclarées.

Secteurs / Paramètres	BSO	BSR	BSF
Moyenne	73 212.28	52 137.55	49 595.64
Ecart-type	171 779.07	135 446	135 399.69
Min	117	111	106
25%	879	325	606
50%	5566	1608	3011
75%	42 250	33 500	25 165
Max	950 000	880 000	950 000

Tableau 3 : Analyse des nombres de données compromises par secteur

La taille des violations est assez homogène entre les différents types d'organisations. Cependant, le nombre moyen de données compromises est plus élevé dans le secteur BSO. Cela pourrait s'expliquer soit par la supériorité des entreprises de ce secteur dans la base de données ou par le fait que les incidents de ce secteur sont de tailles plus importantes que ceux survenus dans les autres secteurs (BSR et BSF).

Une forte dispersion des tailles de violations est également observée dans les trois secteurs, en particulier dans le secteur BSO. Cela traduit une très grande variété des incidents déclarés dans la base, ce qui pourrait conduire à créer, dans la suite, des classes de risques plus

homogènes dans lesquelles les sources d'information seraient séparées, si elles semblent être corrélées avec la gravité du sinistre.

4.3.3.2 Variables relatives à l'incident

A présent, il faudrait s'intéresser aux variables relatives aux incidents. Le graphique suivant représente la répartition des différents types d'attaques dans la base.

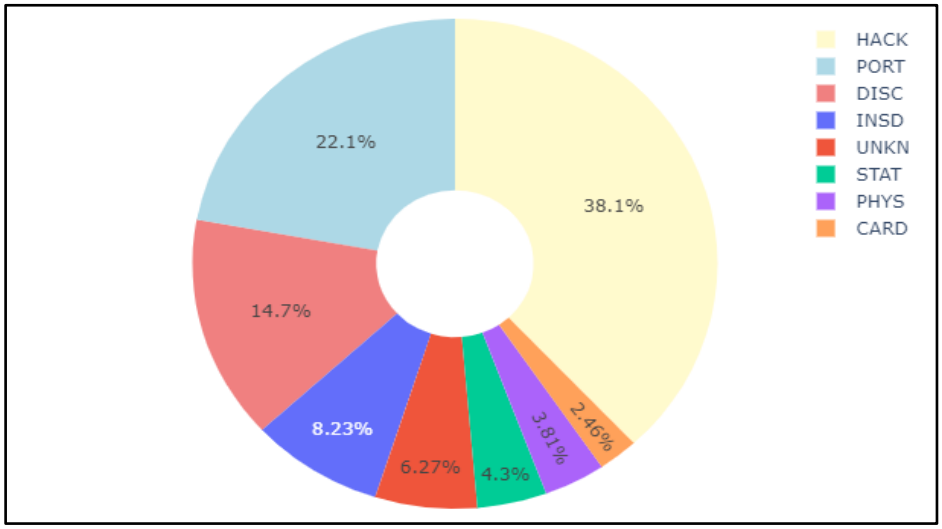


Figure 27 : Répartition des différents types d'attaques

Il apparait une large prédominance d'attaques de type **hacking,** qui à elles seules représentent **38 .1%** des incidents dans la base. Cela pourrait supposer que ce type d'attaque est plus fréquent que les autres.

Cette observation pourrait également provenir du fait que certaines sources d'information spécifient le type d'incidents qu'elles répertorient dans la base**.** Ainsi, une forte présence d'une source recensant des attaques de type **hacking** entrainera la supériorité numérique de ce type d'attaques par rapport aux autres.

Il conviendrait alors d'effectuer une analyse approfondie des différentes sources d'information alimentant la base afin de confirmer ou d'infirmer cette hypothèse et ainsi de quantifier le biais qu'elle engendre.

Les incidents de la base PRC sont rapportés par diverses sources d'informations. Initialement, ces sources d'informations de la base pouvaient être regroupées en 4 grandes catégories :

- **Les agences gouvernementales américaines au niveau fédéral** : Au sein de différents secteurs d'activités, les agences fédérales américaines imposent une notification des différents incidents de fuites de données survenus. Dans le domaine de la santé par exemple, l'HIPPAA (Health Insurance Portability and Accountability Act) exige de notifier tout incident de fuite de données affectant les plus de 500 individus au secrétariat du département américain de la santé et des services humains. Les incidents recensés sont accessibles gratuitement en ligne. Dans la base PRC, les incidents reportés par ces sources d'informations relèvent principalement du secteur médical. (American Medical Association, s.d.)

- **Les agences gouvernementales au sein des Etats américains** : Chaque Etat américain dispose depuis 2018 d'une organisation dédiée aux incidents de fuite de données. Certains Etats décident de rendre publique les incidents recensés. Cependant, ici il n'existe pas encore de règle fixe en ce qui concerne le seuil du nombre de données compromises à notifier obligatoirement.

- **Les Media** : les incidents de fuites de données médiatisés sont aussi reportés dans la base PRC

- **Les organisations à but non lucratif** : la base contient également des incidents recensés par des organismes à but non lucratif tel que ***Databreaches.net***

Après avoir retiré les secteurs non pertinents pour la problématique étudiée, et mis en place la franchise et le plafond sur les tailles des violations, certaines sources d'informations listées plus haut n'apparaissent plus dans la base de données finale. Ci-dessous une représentation de la répartition des différentes sources d'informations restantes dans la base.

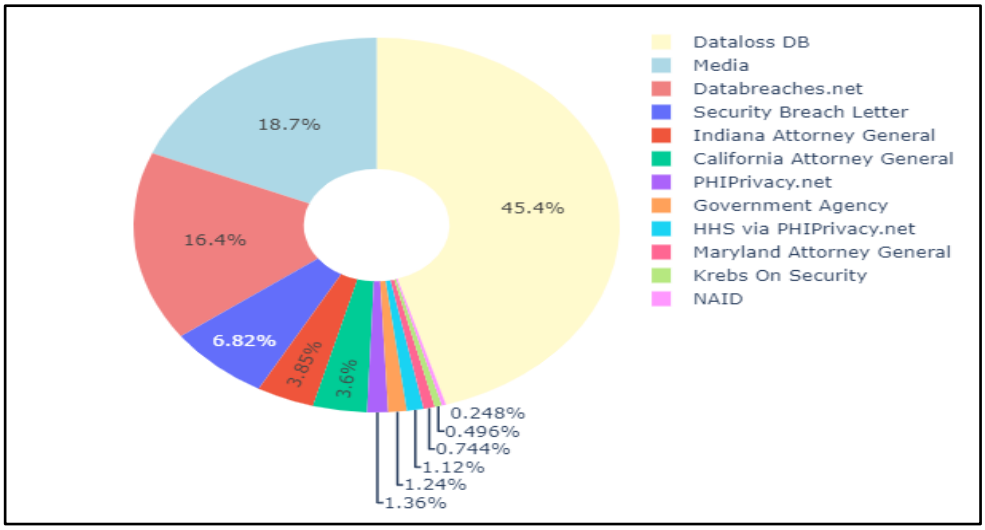


Figure 28 : Répartition des sources d'informations

Il faut noter une majorité d'incidents rapportés par ***Dataloss DB***, qui pourrait se classer parmi les organismes à but non lucratif. En réalité, il s'agit d'un projet de recherche américain fournissant des informations détaillées sur des incidents de fuite de données survenus dans le monde entier. Cependant, il semblerait qu'à ce jour cette organisation ait cessé ses activités.

Pour la suite, afin de ne considérer qu'un nombre limité de sources fiables alimentant la base de données, seules les 4 sources les plus fréquentes seront maintenues, c'est-à-dire *Dataloss DB*, *Media*, *Databreaches.net* et *Security Breach Letter*. Cela permettra également d'analyser un plus grand nombre d'incidents fiables et pertinents, car les sources d'informations retenues sont des organismes à but non lucratif et des médias. Les incidents rapportés sont ainsi assez diversifiés.

4.3.4 Modélisation de la sévérité

4.3.4.1 Sévérité en nombre de données

Dans la détermination de la sévérité d'un incident de violation de données, deux principaux éléments sont à prendre en compte : la taille de la violation (nombre de données compromises) ainsi que la sensibilité de la violation (coût de la donnée perdue). Dans la base étudiée, la variable ***Total Records*** fournit des informations sur la taille des violations.

Afin d'obtenir un tarif final adapté aux caractéristiques de chaque type d'entreprises, il est plus prudent pour la suite, d'évaluer le nombre moyen de données compromises pour chaque secteur pris séparément.

Les graphiques ci-dessous représentent les distributions en densité des tailles des violations pour chaque type d'organisation (BSO, BSR, BSF). Afin de faciliter l'analyse et la lisibilité des graphes, l'étude réalisée dans ce mémoire se portera sur le logarithme des tailles de violations.

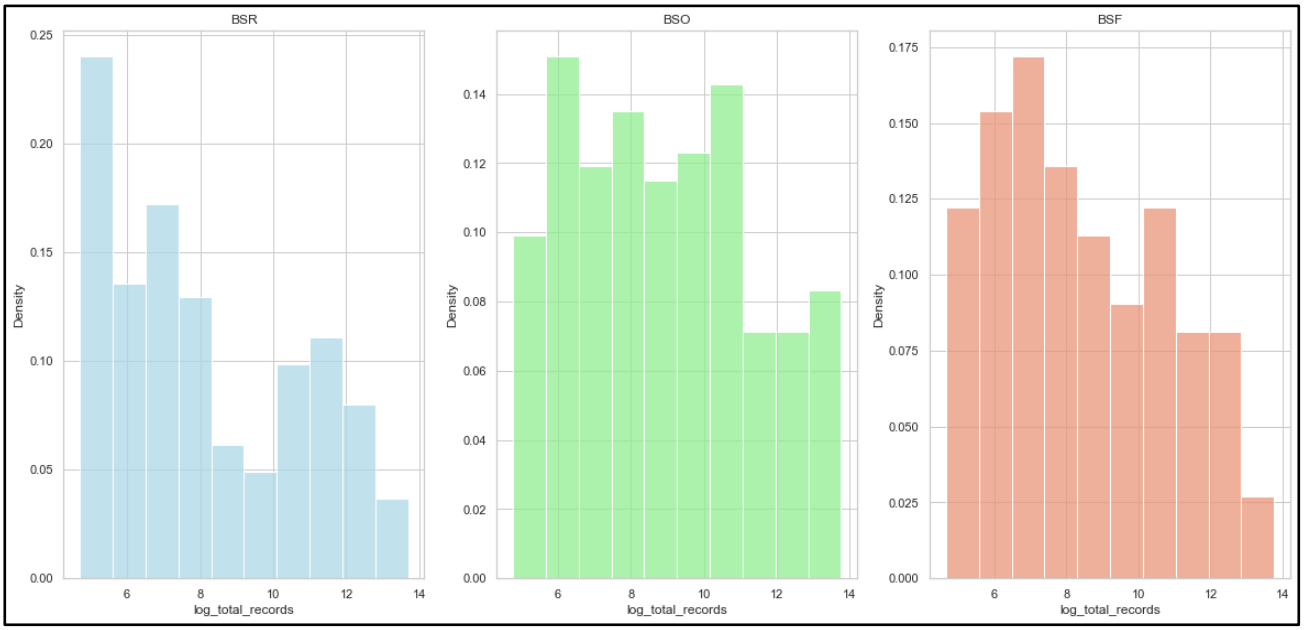


Figure 29 : Distribution des tailles de violations par secteur

A première vue, les distributions des tailles des violations sont assez différentes pour chaque secteur d'entreprise. Ce qui conduit à penser que la répartition des tailles de violations varie en fonction du secteur de l'entreprise. La segmentation des tarifs de l'offre d'assurance en fonction des secteurs d'activités des différentes entreprises semble donc être une bonne idée.
A présent, il est nécessaire de faire une étude approfondie des spécificités des incidents affectant chaque type d'organisation afin de déterminer un coût moyen de sinistre pour chacune d'elles.

4.3.4.1.1 Secteur BSF

Comme notifié plus haut, l'évaluation de la taille des violations pourrait être biaisée si le caractère diversifié des sources d'information alimentant la base de données n'est pas pris en compte. Il serait alors intéressant d'analyser les nombres de données compromises en fonction de la source d'information qui a rapporté l'incident. Le graphique suivant représente la densité des tailles de violations des incidents pour chaque source d'information dans le secteur BSF. L'idée principale étant de vérifier que la distribution des tailles de violations ne dépend pas de la provenance de l'incident.

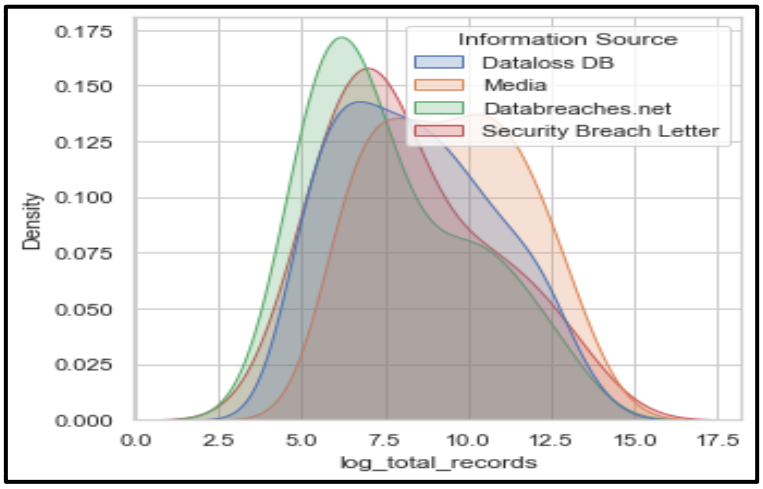


Figure 30 : Densités des logarithmes des tailles de violations BSF par source d'information

Les courbes de densités représentées ont des formes relativement similaires suivant la source d'information, elles se superposent presque. Aucune différence importante n'est relevée entre les répartitions de tailles de violations en fonction des sources d'information pour ce secteur. Pour plus de précision, il faudrait analyser le kurtosis, l'asymétrie ainsi que la moyenne des tailles de violations pour ces différentes sources.

Kurtosis : Le coefficient d'aplatissement ou Kurtosis désigne le moment centré d'ordre 4. Il permet d'étudier la forme pointue ou aplatie d'une distribution. Plus ce coefficient est élevé, plus la distribution des données est pointue en sa moyenne et a des queues de distribution épaisses. La formule de calcul du kurtosis est :

$$K = \frac{E\left[\left(X - E(X)\right)^4\right]}{\sigma^4}$$

σ étant l'écart-type de la distribution.

Skewness : Le coefficient d'asymétrie ou Skewness est le moment centré d'ordre 3. Il permet de mesurer l'asymétrie d'une distribution. Un coefficient de skewness nul indique une distribution symétrique tandis qu'un coefficient positif traduit le fait que la distribution possède une forte queue vers la droite. La formule de calcul du skewness est :

$$S = \frac{E\left[\left(X - E(X)\right)^3\right]}{\sigma^3}$$

σ étant l'écart-type de la distribution.

Ci-dessous les paramètres de la distribution du log de la taille des violations par source d'information.

Sources / Paramètres	Dataloss DB	Media	Databreaches.net	Security Breach Letter
Kurtosis	- 0.89	- 1.08	- 0.65	- 0.48
Skewness	0.31	0.11	0.67	0.62
Moyenne	8.39	9.4	7.69	8.2

Tableau 4 : Paramètres tailles de violations secteur BSF

L'étude des différents paramètres semble confirmer les conclusions tirées plus haut. Les moyennes des logarithmes des tailles de violations se trouvent dans l'intervalle de [7, 9.5] dont l'amplitude est assez faible. Les incidents reportés par les différentes sources d'informations sont alors de tailles assez similaires.

De même, les kurtosis calculés sont négatifs pour les 4 sources d'informations, signifiant que, pour une même variance, les distributions sont relativement « aplaties » (distributions **platikurtiques**). Cependant, le kurtosis de la distribution des tailles des incidents reportés par les media est assez faible par rapport aux autres, ce qui s'observe clairement sur la représentation des densités faites sur la figure 14. Néanmoins, cet écart sera considéré ici comme étant négligeable.

Par ailleurs, les skewness des distributions montrent qu'elles possèdent toutes des fortes queues vers la droite.

Au regard de toutes les informations collectées, tous les incidents reportés par les 4 sources d'informations ci-dessus seront maintenues pour évaluer le nombre moyen de données compromises en cas de survenance d'un incident de violation de données dans le secteur BSF. Une loi de probabilité sera ensuite ajustée sur cette courbe de densité, l'objectif étant de trouver le modèle le plus adapté pour modéliser les tailles de violations.

A l'aide de la fonction ***disfit*** disponible sur python, plusieurs lois de probabilités sont testées sur les données. Certaines de ces lois semblent plus appropriées pour modéliser la taille des violations. Un test d'adéquation de ***Kolmogorov - smirnov*** permettra de trancher parmi celles-ci, laquelle s'ajuste le mieux aux données.

Ci-dessous les résultats obtenus avec la fonction disfit :

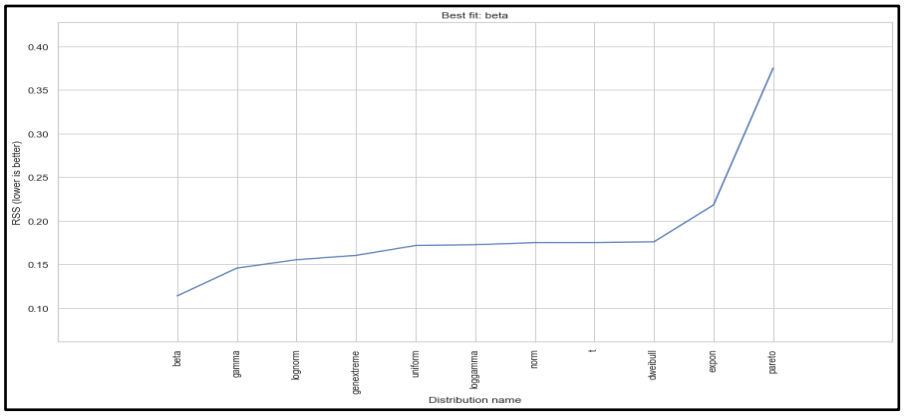


Figure 31 : Résultat fonction disfit sur logarithmes tailles de violations secteur BSF

Le graphique ci-dessus montre que les lois bêta et gamma semblent être les mieux adaptées aux données.

- **Loi Gamma**

Une variable aléatoire X suit une loi Gamma de paramètres α et β strictement positifs et on note
X ~ Γ (α ; β) si elle possède la fonction de densité suivante :

$$f(x) = \frac{x^{\alpha-1}\, exp^{-\frac{x}{\beta}}}{\Gamma(\alpha)\, \beta^{\alpha}} \quad , \quad \forall\, x > 0$$

Γ étant la fonction gamma d'Euler définie comme suit :

$$\Gamma(x) = \int_0^{+\infty} e^{-t}\, t^{x-1}\, dt\, , \forall\, x \in \mathbb{R}$$

- **Loi Bêta**

Une variable aléatoire X suit une loi de bêta de paramètres α et β strictement positifs et on note X ~ B (α ; β), si elle admet pour densité de probabilité la fonction

$$f(t) = \begin{cases} \dfrac{1}{B(\alpha;\beta)}\, t^{\alpha-1}\, (1-t)^{\beta-1} \;, t \in\,]0\,;1[\\ 0\,, \qquad t \notin\,]0;1[\end{cases}$$

Avec :

$$B(\alpha;\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Γ étant la fonction gamma d'Euler.

Afin d'obtenir plus de précision sur l'ajustement des lois et bêta et gamma aux données, analysons la représentation de l'ajustement des lois gamma et bêta sur l'histogramme du logarithme des tailles de violations pour le secteur BSF ainsi que les graphiques Q-Q plot (quantiles théoriques par rapport aux quantiles empiriques) associés.

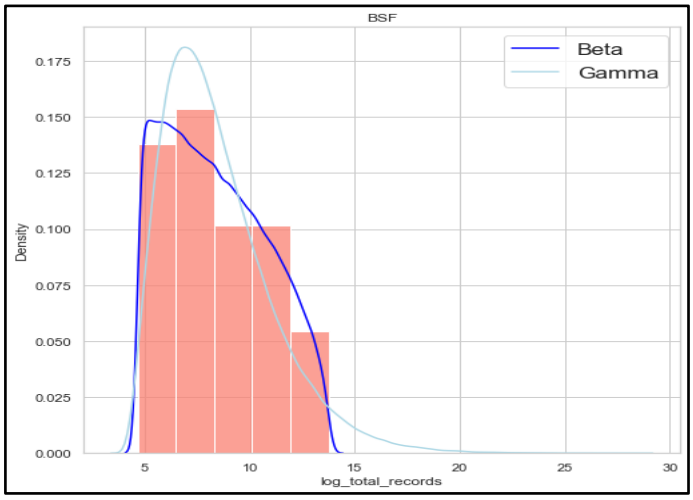


Figure 32 : Ajustement des lois bêta et gamma

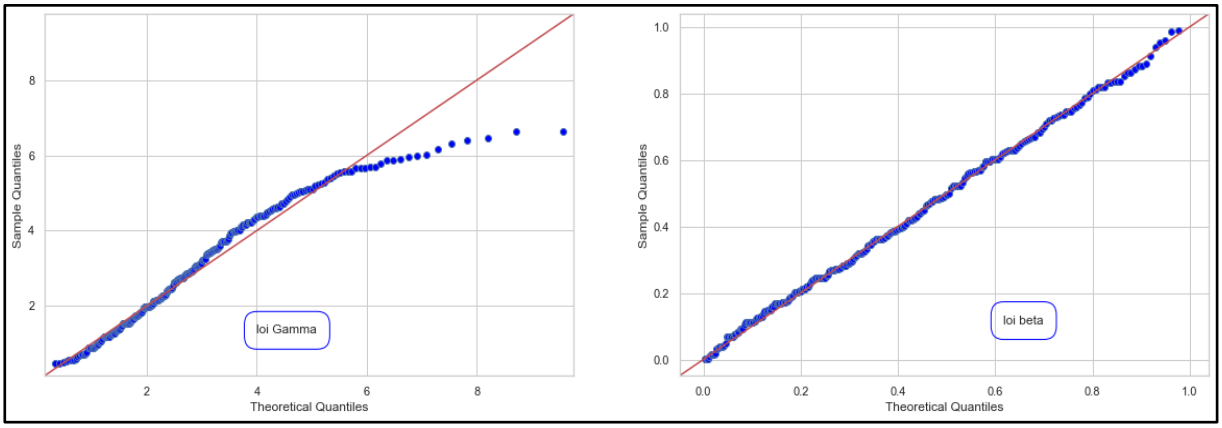


Figure 33 : Q-Q plot lois bêta et gamma sur logarithme des tailles de violations secteur BSF

Une observation des graphiques ci-dessus laisse penser que la répartition du logarithme des tailles de violations suit une loi bêta. Effectuons à présent un test de Kolmogorov-smirnov afin de confirmer ou d'infirmer cette hypothèse.

Portant le nom du mathématicien Andréi Nikoláyevich Kolmogorov, ce test est l'un des tests non-paramétriques les plus courant pour vérifier la qualité de l'ajustement d'une distribution à une loi continue. Le principe du test est assez simple. Il consiste à mesurer l'écart maximum existant entre une fonction de répartition empirique et une fonction de répartition théorique, le test de Kolmogorov-smirnov réalisé étant indépendant de cette loi théorique.

La première étape de réalisation du test consiste à définir deux hypothèses :

H_0 : la distribution suit la loi Lo contre H_1 : la distribution ne suit pas la loi Lo.

Déroulement du test :

Soit n observations ($x_0, x_1, x_2, \ldots x_n$) d'une variable aléatoire X de fonction de répartition F.

Dans un premier temps, les valeurs observées sont ordonnées : $x_1 <= x_2 <= \ldots <= x_n$. Ensuite, on pose : $F(x_1) = 1/n$, $F(x_2) = 2/n$, ..., $F(x_n) = 1$, ce qui définit la fonction de répartition de F en escalier.

On calcule la statistique de test $K = \sup |F(x)-F_0(x)|$, par la formule

$$K = \max_{1 \leq j \leq n} \left(\frac{j}{n} - F_o(x_j), F_o(x_j) - \frac{j-1}{n} \right)$$

F_o étant la fonction de répartition de la distribution théorique considérée ici.

Enfin, la statistique K obtenue est comparée à une valeur critique Dα (n) fournie par les tables de Kolmogorov-Smirnov, α étant le seuil d'acceptation du test. Si K > Dα (n), l'hypothèse Ho est rejetée avec un risque α d'erreur. Dans le cadre de notre étude, le seuil d'acceptation choisi est : α = 5%.

Ci-dessous, les résultats du test de Kolmogorov-smirnov réalisé sur les données :

	Bêta	**Gamma**
K	0.039	0.067
P-value	0.838	0.212

Tableau 5 : Résultats test de Kolmogorov Smirnov loi bêta et gamma secteur BSF

La ***p-value*** désigne la plus petite valeur pour laquelle l'hypothèse H_o est conservée. En général, cette statistique est utilisée pour valider des tests statistiques, en la comparant au seuil d'acceptation α choisit.

- Si p-value > α, on conserve Ho

- Si p-value < α, on rejette Ho

Dans le cadre de ce test, les p-value obtenues sont supérieures au seuil d'acceptation α. Cependant, la p-value de la loi bêta (83.8 %) est largement supérieure à celle de la loi gamma (21.2 %).

De plus, en observant la valeur critique correspondant au test réalisé ici (Dα (n) = 21.2), l'hypothèse Ho est conservée pour la loi bêta et la loi gamma.

Au vu de tous ces résultats, la loi bêta sera retenue pour la modélisation des logarithmes des tailles de violations dans le secteur BSF. Les paramètres de la loi ajustée à la distribution sont résumés dans le tableau ci-dessous :

Paramètres	**a**	**b**	**loc**	**scale**
Valeurs	1.028	1.484	4.662	9.202

Tableau 6 : Paramètres estimés loi bêta secteur BSF

Ensuite, 300 000 réalisations d'une loi de bêta avec les paramètres estimés ci-dessus sont générées, afin d'approcher le nombre moyen de données compromises dans le secteur BSF.

Une fois la simulation effectuée, les valeurs des tailles des violations simulées sont plafonnées à **50 000 données**. Cette manœuvre permet de mettre en place un **plafond de sinistre à 50 000 données violées**. En effet, la problématique actuelle étant la construction d'un produit d'assurance à destination des PME/TPE, laisser le nombre maximum de données compromises à 1 000 000 par exemple, entrainerait un coût moyen de sinistre très élevé et par conséquent une prime d'assurance trop chère pour ce type d'entreprise. De plus, du point de vue de l'assureur, cela représenterait une capacité de couverture trop importante et il serait difficile pour lui de trouver un réassureur prêt à le couvrir, surtout étant donné la complexité du risque cyber. De cet fait, il faudrait mettre en place une limite de sinistre adaptée au périmètre d'étude. La limite à 50 000 données a été choisie de façon arbitraire. Dans la suite, une étude de la sensibilité de la prime obtenue à la variation de ce plafond de sinistre sera effectuée.

Le nombre moyen de données compromises obtenu dans le secteur BSF est égal à **15 424** (taille moyenne réelle observée sur les données brutes : **15 386**).

4.3.4.1.2 Evaluation du coût moyen d'une donnée dans le secteur BSF

Le **Ponemon institute** est un centre de recherche consacré à la protection des données et aux politiques de sécurité de l'information. Il définit également sa mission comme étant de mener des études de haute qualité empiriques sur les questions critiques affectant la gestion et la sécurité d'informations sensibles sur les personnes et les organisations. Ainsi, dans cette logique, il publie chaque année, en partenariat avec **IBM security**, un rapport sur le coût des violations de données intitulé « ***Cost of Data Breach »***. Ce rapport met en lumière différents paramètres des incidents de violations de données tels que les coûts moyens des violations par secteurs, par pays mais aussi divers éléments pouvant influencer ces coûts tels que la taille de l'entreprise ou le type d'attaque subit par cette dernière. Dans l'édition 2019 de ce rapport, figure également, le coût moyen d'une donnée par type d'organisation. Il s'agit d'une information très intéressante pour la problématique actuelle car les coûts estimés dans cette étude prennent en compte plusieurs éléments à savoir :

- **Frais de notification** : Il s'agit des différents coûts permettant d'informer les personnes dont les données ont été compromises lors de la violation. Ce sont par exemple, les frais de communications avec le régulateur ou encore les frais d'engagement d'experts externes.

- **Frais de détection et gestion** : Il s'agit des différents coûts relatifs à la détection et à la réparation de l'incident survenu. Ce sont par exemple, les frais d'assistance et de gestion de crise ou encore les frais d'experts.

- **Coûts post-violation** : Ils regroupent les frais relatifs aux processus mis en place pour aider les personnes affectées par la violation à communiquer avec l'entreprise, ainsi que les coûts associés à la réparation de l'incident. Il s'agit par exemple, des frais d'intervention, des dépenses juridiques ou des paiements des amendes potentielles.

- **Coûts de la perte d'activité** : Ce sont les frais associés à l'interruption potentielle des activités et du système informatique des entreprises victimes de l'incidents. Il s'agit par exemple des coûts liés à la perte de clients, des coûts liés à la perte de revenu ou encore à la perte de réputation.

Ainsi, les coûts estimés dans l'étude se rapprochent fortement de ceux qui sont pris en charge dans la garantie ***atteinte aux données*** proposée dans le produit d'assurance construite dans le cadre du présent mémoire. Les coûts fournis dans le rapport Ponemon 2019 serviront alors de base pour la détermination du coût moyen d'un incident de violation de donnée. Dans la suite, l'évaluation du coût moyen de sinistre dans chaque secteur se fera à l'aide de la formule suivante :

$$\boxed{Coût_{moyen} = Nb_{moy} * unit_{moy}}$$

Avec :

Nb_{moy} : le nombre moyen de données compromises

$Unit_{moy}$: le coût moyen unitaire d'une donnée violée

4.3.4.1.3 Calcul du coût moyen de sinistre dans le secteur BSF

Le rapport « ***cost of data breach 2019 »*** publié par le Ponemon Institute révèle que, en 2019, le coût moyen d'une donnée sensible ou confidentielle compromise dans le secteur finance est de **210$ USD**, tous pays confondus (Ponemon Institute, 2019). Afin d'avoir une estimation plus actuelle de la sévérité des cyber incidents, une actualisation du coût unitaire fourni dans le rapport sera réalisée.

Diverses approches peuvent être utilisées pour actualiser le coût unitaire d'une donnée. Une première idée serait de se baser sur les taux d'inflations annuels entre 2019 et 2022 ou encore de faire une simple translation entre les coûts moyens totaux des incidents de violations de données fournis dans les rapports Ponemon de 2019 et 2022. Pour l'étude réalisée ici, la seconde approche sera privilégiée car elle semble plus pertinente. En effet, le calcul du taux d'inflation annuel se base sur des facteurs tels que l'indice des prix à la consommation ou encore les dépenses moyennes des ménages. Ces éléments paraissent tous assez décorrélés de l'évolution du coût moyen d'une données. De plus, plusieurs éléments, autre que l'inflation, peuvent influer sur la variation du coût moyen unitaire de la donnée. Cela pourrait être par exemple, l'évolution des techniques de sécurité mise en place ou encore l'amélioration des dispositifs de réponse aux incidents cyber.

Actualisation du coût de la donnée

Soient les variables ci-dessous :

- **unit_BSF_2019** : coût moyen unitaire d'une donnée compromise en 2019 dans le secteur de la finance

- **unit_BSF_2022** : coût moyen unitaire d'une donnée compromise en 2022 dans le secteur de la finance. Cette variable représente l'actualisation de la variable précédente.

- **coût_BSF_2019** : coût moyen total d'un incident de violation de donnée en 2019 dans le secteur BSF

- **coût_BSF_2022** : coût moyen total d'un incident de violation de donnée en 2022 dans le secteur BSF

Ci-dessous un tableau récapitulant les valeurs des variables ci-dessous ainsi que leur provenance :

Variables	Valeurs	Sources
unit_BSF_2019	210 $ USD	*Cost of Data breach 2019*
coût_BSF_2019	5.86 M $ USD	*Cost of Data Breach 2019*
coût_BSF_2022	5.97 M $ USD	*Cost of Data Breaches 2022*

Tableau 7: Coûts moyen des incidents de violations de données secteur BSF

Le pourcentage d'évolution du coût moyen total d'un incident de violation entre 2019 et 2022 dans le secteur financier est :

$$t = \frac{\text{coût}_{\text{BSF}_{2022}}}{\text{coût}_{\text{BSF}_{2019}}} - 1 = 1.87\%$$

Ainsi, le coût moyen unitaire d'une donnée dans le secteur BSF en 2022 est approximativement :

$$unit_{BSF_{2022}} = unit_{BSF_{2019}} * (1 + t) = 213.927\ \$\ USD$$

Le coût moyen total de sinistre dans le secteur BSF est donc :

$$Coût_{moyen_{BSF}} = Nb_{moy_{BSF}} * unit_{moy_{BSF}}$$

$$Coût_{moyen_{BSF}} = 15\ 424.17 * 213.927$$

$$\boldsymbol{Coût_{moyen_{BSF}} = 3\ 299\ 646.42\ \$\ USD}$$

Cependant, les coûts estimés prennent en compte la perte d'activité des entreprises. Or, elle fait l'objet d'une garantie à part entière dans l'offre d'assurance proposée dans le cadre de ce mémoire. Il serait alors plus judicieux, afin de ne pas surestimer le montant total de sinistre pour la garantie ***« atteinte aux données »***, d'extraire le coût de la perte d'activité. D'après l'édition 2021 du rapport Ponemon, **les frais liés à la perte d'activité d'une entreprise victime de cyber attaque représentent environ 38 % du coût moyen total de l'incident**. (Ponemon Institute, 2021)

Ce résultat est cohérent avec les estimations faites dans les précédents rapports de l'institut. L'estimation du coût moyen de sinistre pour la garantie *« atteinte aux données »*, sera alors effectuée en prenant en compte ce résultat. Ainsi, en retirant du coût moyen obtenu, la part relative à la perte d'exploitation, on a :

$$Coût_{moyen_{BSF}} = 2\,045\,780.78\,\$\,USD$$

Le coût moyen estimé étant calculé sur la base de données américaines, il pourrait ne pas être adapté à la réalité des entreprises françaises, qui sont la cible de l'offre d'assurance proposée ici. D'après le rapport « ***Cost of Data Breach 2022 »***, le coût moyen total d'un incident de violation de données aux Etats Unis en 2022 représente 2 fois celui d'un incident survenu en France. (Ponemon Institute , 2022)

Cependant, il faut noter que cet écart pourrait provenir du fait que les capacités d'indemnisations des contrats d'assurance cyber proposés sur le marché américain sont plus élevées que celles incluses dans les contrats d'assurance cyber français car, comme relevé dans la section 2.4.1 du mémoire, une sous assurance s'observe sur le marché français.

Cela pourrait également s'expliquer par la forte présence d'entreprises relevant du secteur médical dans les entreprises américaines étudiées dans le rapport de Ponemon. Les montants des incidents affectant les entreprises de ce secteur étant généralement très élevés, cela pourrait entrainer une augmentation du coût moyen global de sinistre observé aux Etats-Unis.

Etant donné le caractère mouvant du risque cyber, le choix a été fait de ne pas tenir compte de cet écart dans l'évaluation du coût moyen de sinistre en France. Ainsi, le coût moyen de sinistre obtenu en France est égal à :

$$\boldsymbol{Coût_{moyen_{BSF}} = 2\,045\,780.78\,€}$$

Le taux de change appliqué étant celui en vigueur au **23/08/2022.**

Le coût moyen obtenu est relativement élevé, ce qui n'est pas surprenant car les données disponibles au sein des entreprises du secteur financier sont assez sensibles (informations bancaires, adresse…).

4.3.4.2 Secteur BSO

A présent, il convient de se pencher sur l'évaluation du coût moyen de sinistre des entreprises du secteur BSO. Tout comme précédemment, la première étape consistera à analyser les densités des tailles de violations pour chaque source d'information.

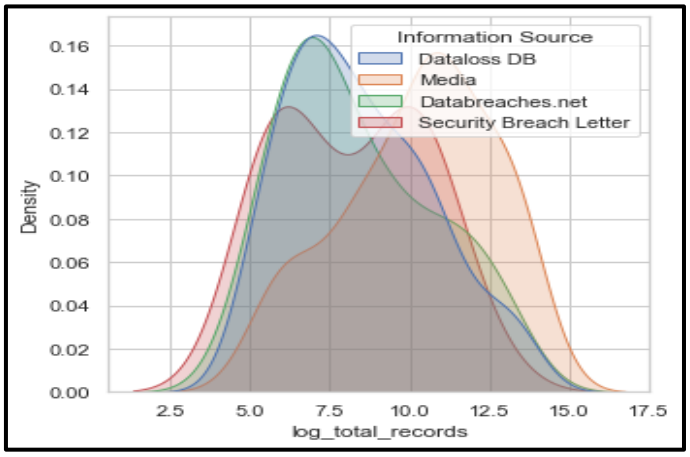


Figure 34: Densités des logarithmes des tailles de violations BSO par source d'information

A première vue, contrairement au constat fait dans le secteur BSF, les courbes de densités représentées ici ont des formes différentes suivant la source d'information. Il faut également noter un décalage à droite de la courbe de densité de la taille des incidents reportés par les **Medias** par rapport à ceux reportés par les autres sources d'informations. Ce résultat n'est pas surprenant car les médias recensent le plus souvent des incidents de grande ampleur ; avec de plus grand nombre de données compromises.

Par ailleurs, la courbe de densité des tailles des violations reportés par Dataloss DB semble plus étalée sur le support que celle des incidents reportés par les autres sources. Cela permet de supposer que Dataloss DB répertorie des incidents de toute taille, permettant ainsi de prendre en compte plus d'éventualités de scénarios qu'avec les autres sources. Pour plus de précision, il faudrait analyser le kurtosis, l'asymétrie ainsi que la moyenne des tailles de violations pour ces différentes sources.

Sources / Paramètres	Dataloss DB	Media	Databreaches .net	Security Breach Letter
Kurtosis	- 0.56	- 0.74	- 0.84	- 1.56
Asymétrie	0.48	- 0.40	0.50	0.07
Moyenne	8.41	10.21	8.36	8.19

Tableau 8 : Paramètres des tailles de violations secteur BSO

Le tableau ci-dessus révèle que les tailles moyennes des violations reportés par les différentes sources d'informations sont assez diversifiées. Les incidents reportés par les médias par exemple, sont de taille supérieure d'environ 22% de plus par rapport à ceux reportés par les autres sources d'informations. De plus, les kurtosis calculés montrent que les distributions des tailles de violations des 4 sources d'informations sont toutes **platikurtiques**.

Cependant, il faut noter que seuls les incidents reportés par les médias ont un skewness négatif, ce qui justifie le décalage à droite observé au niveau de la courbe de densité. Au regard de toutes les informations réunies, afin d'être prudent dans les analyses et d'obtenir des résultats plus réalistes, seuls les incidents reportés par **Dataloss DB** seront pris en compte pour évaluer le nombre moyen de données compromises en cas de survenance d'un incident de violation de données. Ce choix se base sur le fait que, cette source d'information répertorie des incidents de toute taille, ce qui permet de prendre en compte plus d'éventualités de scénarios qu'avec les autres sources.

Une loi de probabilité est ensuite ajustée sur les données finales afin d'identifier le modèle le plus adapté pour modéliser les tailles de violations. Les graphiques suivants montrent l'ajustement des lois gamma et bêta sur l'histogramme du logarithme des tailles de violations pour le secteur BSO, les graphiques Q-Q plot associés ainsi que les résultats du test de Kolmogorov-Smirnov réalisé.

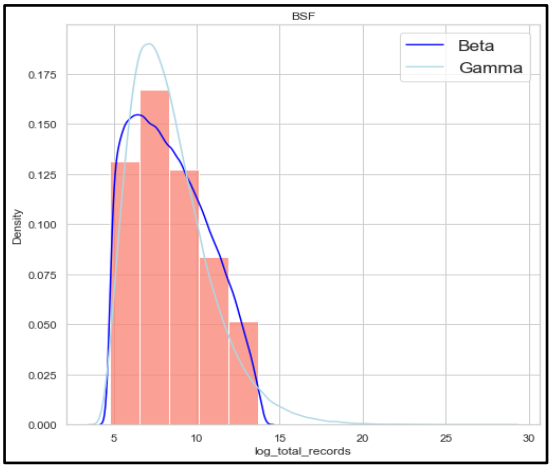


Figure 35 : Ajustement lois bêta et gamma secteur BSO

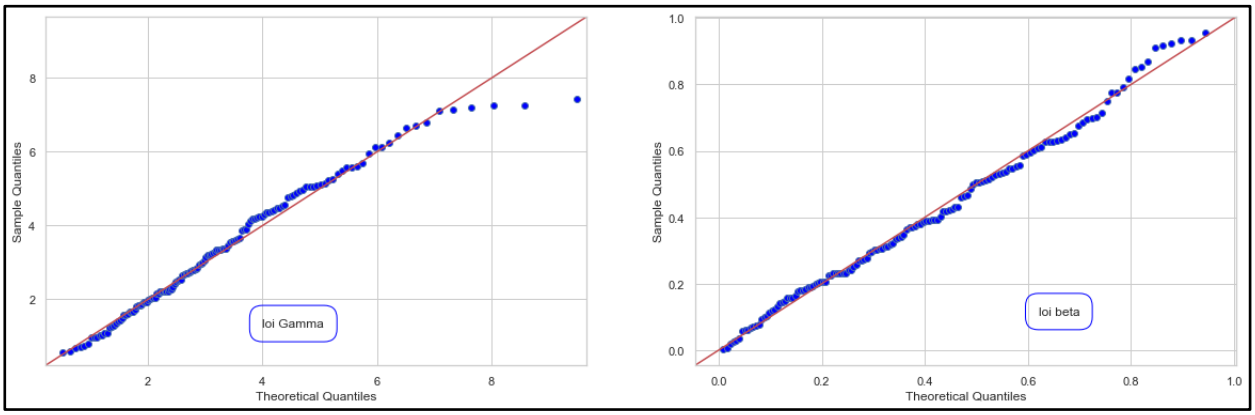


Figure 36: : Q-Q plot lois bêta et gamma sur logarithme des tailles de violations secteur BSO

	Bêta	**Gamma**
K	0.04	0.063
P-value	0.94	0.60

Tableau 9 : Résultats test de Kolmogorov Smirnov lois bêta et gamma secteur BS0

L’analyse des différents graphiques ci-dessus permet de choisir de modéliser la distribution de la taille des violations dans le secteur BSO par une loi bêta. Les paramètres obtenus pour cette loi sont résumés ci-après :

Paramètres	a	b	loc	scale
Valeurs	1.184	1.809	4.749	9.334

Tableau 10 : Paramètres estimés loi bêta secteur BSO

Une fois les paramètres de la loi estimés, tout comme précédemment, 300 000 réalisations de la loi bêta ajustée sont générées et le plafond de 50 000 données compromises est appliqué. Le nombre moyen de données compromises obtenu est sensiblement égal à **14 760 (**taille moyenne réelle observée sur les données brutes : **13 968)**, ce qui est inférieur au résultat obtenu dans le secteur BSF.

4.3.4.2.1 Evaluation du coût moyen de sinistre dans le secteur BSO

Le secteur BSO étant plus large que les autres secteurs, il est nécessaire de prendre en compte plusieurs paramètres dans le calcul du coût moyen unitaire d’une donnée.

La représentation de la répartition des types d’organisations présentes dans l’étude Ponemon de 2019 ainsi que les coût moyens unitaires associés figurent sur le graphique ci-après :

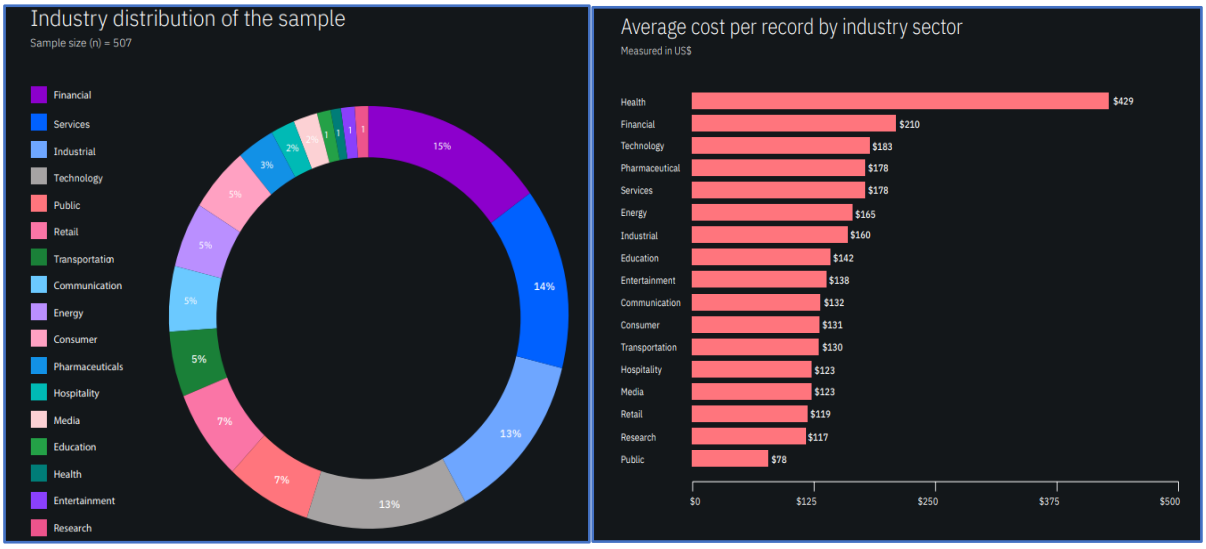


Figure 37 : Répartition des types d’organisations et coût unitaire de donnée dans l’étude Ponemon de 2019
Source : (Ponemon Institute, 2019)

Dans la suite du mémoire, les types d'organisations suivantes : ***Services, technology, research*** et ***communication*** seront considérées comme appartenant au secteur BSO. Ainsi, pour évaluer le coût moyen d'une donnée violée en 2019 dans ce secteur, il faudra effectuer une moyenne pondérée des coûts moyens dans toutes les organisations considérées. Le coût moyen unitaire obtenu est donc :

$$unit_{moy_{2019}} = \frac{183 * 0.13 \ + \ 173 * 0.14 \ + \ 117 * 0.01 \ + \ 132 * 0.05}{0.13 + 0.14 + 0.01 + 0.05}$$

$$\boldsymbol{unit_{moy_{2019}} = 169.03 \text{ \$ USD}}$$

Une fois le coût de la donnée déterminé, il est important tout comme précédemment, de l'actualiser en 2022. Ci-dessous un récapitulatif des résultats obtenus pour ce secteur :

Variables	Valeurs	Sources
unit_BSO_2019	210 $ USD	*Cost of Data breach 2019*
coût_BSO_2019	5.86 M $ USD	*Cost of Data Breach 2019*
coût_BSO_2022	5.97 M $ USD	*Cost of Data Breaches 2022*
Unit_BSO_2022	171.49 $ USD	*Calcul suivant méthode BSF*
Cout_moy_BSO	2,53 M $ USD	*Calcul suivant méthode BSF*

Tableau 11 : Résumé différents coûts secteurs BSO

En retirant les **38% correspondant à la perte d'exploitation**, le coût moyen de sinistre dans le secteur BSO est égal à :

$$\boldsymbol{Coût_{moy_{BSO_{Fr}}} = 1\,575\,248.81 \text{ €}}$$

Le taux de change appliqué étant celui en vigueur au **23/08/2022.**

4.3.4.3 BSR

Enfin, tous comme effectué dans les deux secteurs précédents, il faudrait aussi évaluer le coût moyen de sinistre dans le secteur BSR.
Ci-dessous un graphique représentant les densités des tailles de violations pour chaque source d'information dans le secteur BSR.

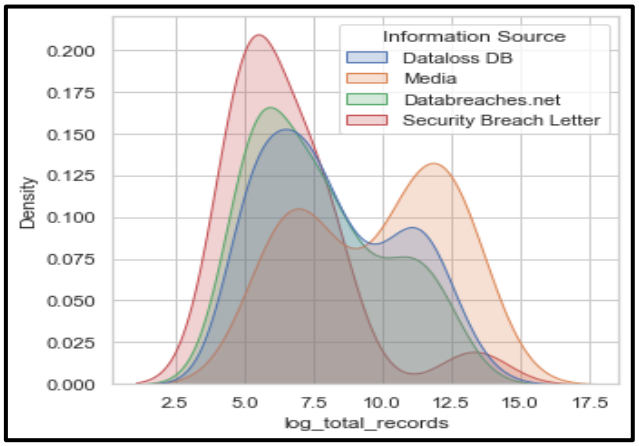


Figure 38 : Densités des logarithmes des tailles de violations BSR par source d'information

Les courbes de densité représentées sont approximativement superposées. L'analyse des différents paramètres (kurtosis, skewness, moyenne) semble confirmer cette observation. Pour la suite, les incidents reportés par toutes les sources d'informations disponibles seront maintenus pour modéliser le nombre moyen de données violées.

Une analyse similaire à celle réalisée dans les secteurs BSF et BSO permet d'obtenir un nombre moyen de données compromises égal à **15 287** dans le secteur BSR. Les différents graphiques d'adéquations ainsi que les résultats du test de Kolmogorov-Smirnov sont disponibles en annexe.
Le calcul du coût moyen de sinistre se fera en se basant sur la même méthode que celle utilisée pour les secteurs précédents.

Ci-dessous un tableau récapitulatif des informations obtenues dans chaque secteur d'activité :

Secteurs	Nombre moyen de données compromises	Coût moyen de sinistre
BSF	15 424	2 045 780.78 €
BSO	14 760	1 575 248.81 €
BSR	15 287	2 010 622.83 €

Tableau 12 : Coûts moyens de sinistres par secteur d'organisation

Les montants de sinistres obtenus dans chaque secteur semblent cohérents avec la réalité du marché, en ce qui concerne la cible visée par l'offre d'assurance proposée dans le cadre du présent mémoire. En effet, d'après le dernier rapport de l'AMRAE, le montant total d'indemnisation versées aux petites entreprises (moins de 20 salariés) victimes d'incident cyber est de 10 M€ (AMRAE, 2022). Il faudrait également noter que ce montant n'inclus pas uniquement les indemnisations au titre de la garantie atteinte aux données.

4.3.4.4 Limites du modèle du coût

Bien que les coûts des incidents cyber estimés soient plus ou moins cohérents avec la réalité d'après le dernier rapport de l'AMRAE, les modèles mis en place comportent tout de même certaines limites.

D'une part, il faut noter que le coût marginal d'une donnée n'a pas été pris en compte dans les calculs. En effet, il est tout à fait possible que le nombre de données violées ait un impact non négligeable sur le coût unitaire de la donnée. Les frais de notifications aux personnes concernées par exemple, pourraient ne pas être facturés par donnée mais plutôt pour un ensemble de données violées. Ainsi plus le nombre de données personnelles compromises est élevé, plus le montant de ces frais sera faible et par conséquent le coût unitaire de la donnée le sera aussi. En se basant sur une simple moyenne sur les coûts de données compromises, les modèles de coûts construits ne prennent pas en compte cet effet.

D'autre part, les calculs effectués sont principalement basés sur les montants d'incidents disponibles dans les rapports fournis par l'institut Ponemon. Or, ces derniers ont été évalué pour des incidents dont les nombres de données violées sont compris entre 2000 et 100 000. Ainsi, étant donnée la mise en place du plafond de données compromises à 50 000 qui a été effectué sur les données PRC, il est possible que les coûts de sinistres obtenus aient été surestimés. Cependant, aucun écart aberrant n'a été observés entre les coûts de sinistres estimés et les valeurs disponibles dans des rapports tels que LUCY.

Par ailleurs, une autre limite importante à relever ici est l'absence de la prise en compte de la nature des attaques dans l'évaluation du coût des incidents cyber. En effet, il est probable que la gestion de certains types d'attaque soient plus coûteux que d'autres. D'ailleurs, le rapport Ponemon 2022 illustre bien ce phénomène en fournissant le montant des incidents cyber par type d'attaques. Il y apparait notamment que les attaques de type phishing sont plus coûteuses que celles causées par des erreurs humaines. Ainsi, le fait que cet aspect n'ait pas été pris en compte dans les modèles construits pourrait entrainer une sous-estimation dans l'évaluation du risque considéré dans la mesure où si les attaques les plus coûteuses sont également les plus fréquentes, cela ne sera pas pris en compte.

4.3.5 Modélisation de la fréquence

Certaines raisons laissent penser que la base PRC n'est pas adéquate pour modéliser rigoureusement la fréquence de sinistre. En effet, le mode d'alimentation de la base ne permet pas d'obtenir une estimation stable du niveau d'exposition des entreprises au risque. La diversité des sources d'information alimentant la base ne permet pas de faire la distinction entre l'évolution de la sinistralité causée par une réelle augmentation du risque et celle venant de l'instabilité de de ces sources d'informations. Les figures ci-dessous montrent la répartition des différents types d'attaques selon les sources d'information ayant rapportés les incidents.

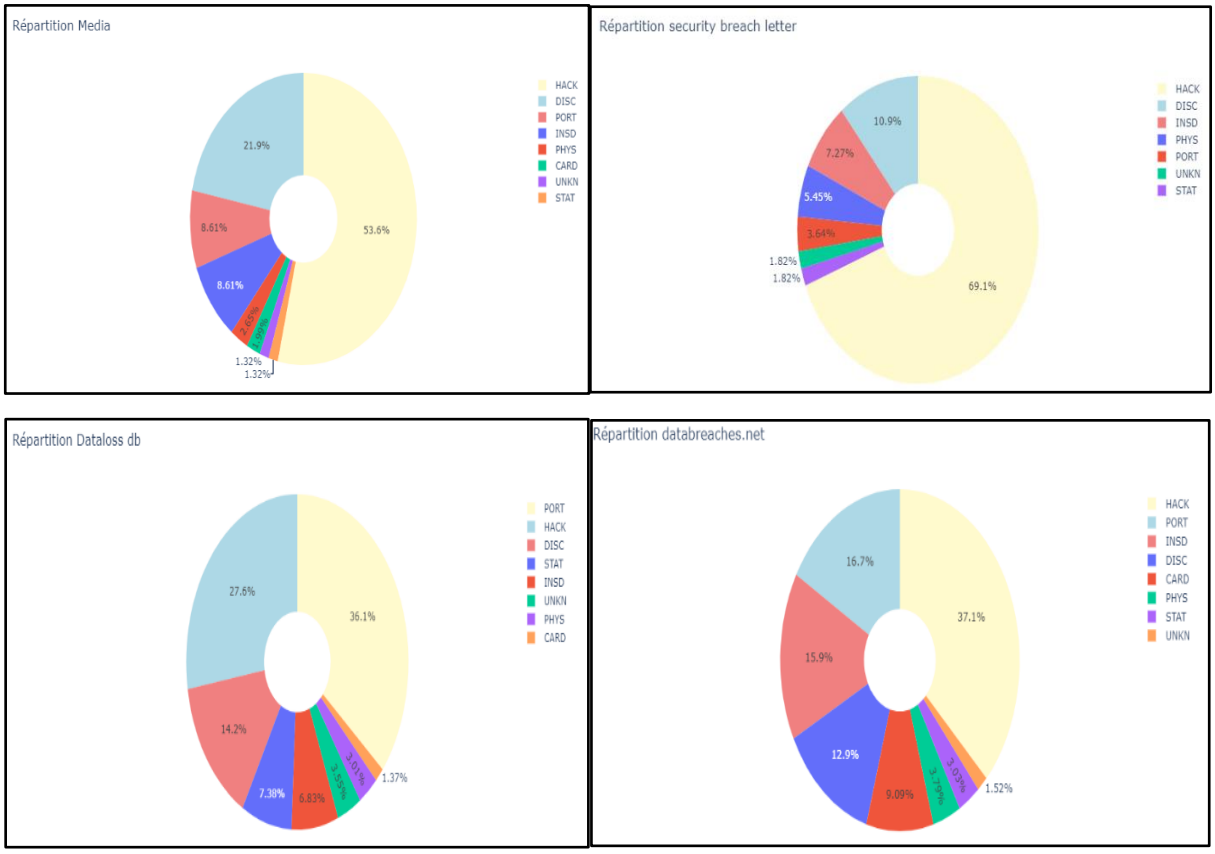


Figure 39 : Répartition des types d'attaques rapportés par source d'informations

Ces différents graphiques révèlent que les différentes sources d'informations semblent spécifier les types d'attaques qu'elles répertorient dans la base. Tandis que les médias et les agences gouvernementales (security breach letter) reportent une grande majorité d'attaques de type **hacking**, les organismes à but non lucratifs tels que Dataloss DB et Databreaches.net, eux enregistrent une majorité d'attaques de type **Port**. Cela a également été vérifié dans l'analyse de la sévérité où, il avait constaté que le nombre moyen de données compromises étaient plus important dans le cas des incidents reportés par les médias.

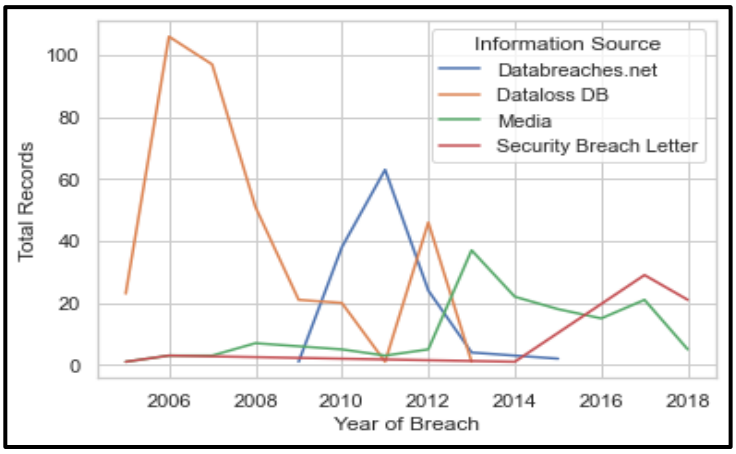


Figure 40 : Evolution des sources d'informations par années

De plus, l'observation du dernier graphique révèle que, les différentes sources qui alimentent la base PRC n'ont pas la même stabilité au cours du temps. Par exemple, les organismes à but non lucratifs (Dataloss DB et Databreaches.net) semblent avoir cesser d'enrichir la base à partir de 2013 tandis que les Medias l'ont massivement alimenté entre 2012 et 2014.

Néanmoins, la base PRC reste un bon indicateur pour mesurer l'exposition au risque de cyber attaques. Elle a d'ailleurs servi de base à la réalisation de nombreux travaux scientifiques sur l'étude de ce risque tels que ceux de Farkas sur la modélisation des sinistres cyber à l'aide d'une régression de Pareto généralisée (Sébastien, Olivier, & Maud, 2020) ou encore ceux menés en 2017 par Eling et Loperfido sur la tarification et la mesure du risque cyber. (Martin & Nicola, 2017)

En outre, étant donné que la problématique actuelle est de fournir une approximation de la prime pure d'une garantie *« atteinte aux données »*, l'utilisation de la base pourrait permettre d'obtenir une estimation plus ou moins réaliste de la réelle exposition du risque. Cette estimation pourra être améliorée au cours de l'évolution du portefeuille d'assurance.

Dans la suite du mémoire, l'estimation de la fréquence du risque se basera principalement sur deux hypothèses. La première hypothèse est celle selon laquelle, tous les incidents de violations de données enregistrés aux Etats Unis ont été notifiés et enregistrés dans la base PRC. Cette hypothèse peut paraître a priori irréaliste, mais elle reste pertinente étant donné que la notification des incidents de violations de données à caractère personnels est obligatoire sur presque tout le territoire américain, et cela depuis plusieurs années.

La seconde hypothèse consiste à considérer qu'au sein d'un même secteur, toutes les entreprises ont la même probabilité d'être victime de cyber attaques. Cette dernière hypothèse pourrait ne pas se vérifier dans la réalité car il se pourrait que certaines entreprises soient plus à risque que d'autres. Par exemple, les petites entreprises sont souvent considérées comme des cibles plus faciles et pourraient ainsi plus souvent victimes d'incidents cyber que les plus grandes. La base PRC ne disposant par d'informations sur la structure des entreprises (nombre d'employés, chiffre d'affaires…), il n'est pas possible ici d'obtenir une segmentation plus fine.

Différents rapports relatifs au nombre d'entreprises existantes sur le territoire américains sont disponibles sur le site du bureau de recensement des Etats-Unis [8]. Afin de rapprocher les données présentes dans ces rapports et ceux de la base PRC, il a fallu retirer toutes les entreprises constitué d'une seule personne *(Sole Proprietorships)*. Ci-dessous une répartition du nombre d'entreprises en 2019 dans chacun des secteurs retenus.

	BSR	**BSF**	**BSO**	**Total**
Nombre entreprises	459 852	164 464	1 009 950	1 634 266

Tableau 13: Nombre d'entreprises par secteur aux USA en 2019

Le précédent tableau montre que plusieurs entreprises américaines n'apparaissent pas dans la base PRC. Dans la logique de l'étude réalisée dans ce mémoire, cela voudrait dire que très peu d'entreprises sont victimes de cyber attaques aux Etats Unis. Cependant, il est plus probable que cette observation vienne du fait que plusieurs incidents cybers survenus aux Etats Unis n'ont pas été enregistrés dans la base.

Ce phénomène pourrait également s'expliquer par le fait que les incidents reportés dans la base PRC sont relatifs à certains types d'entreprises. Par exemple, il est plus envisageable que les sources d'informations telles que les médias rapportent principalement dans la base des incidents affectant de grandes organisations comme Yahoo ou encore Facebook plutôt que ceux dont sont victimes les PME. Cependant, le manque d'informations concernant les caractéristiques des entreprises dans la base PRC ne permet malheureusement pas de vérifier cette information.

Ci-dessous un résumé du nombre d'attaques subies par les compagnies :

Nombre d'incidents	**Nombre d'entreprises**
0	1 633 608
1	623
2	26
3	7
4	2

Tableau 14 : Répartition nombre de compagnies par nombre d'incidents

Le tableau ci-dessus montre un déséquilibre important entre les différentes classes constituées. Ainsi, afin de pallier à ce phénomène, les données de la base PRC seront considérées comme étant 0-tronquées dans l'application des modèles de détermination de la fréquence de sinistre.

Des données sont dites 0-tronquées, si les observations ayant pour variable réponse 0 ont été retirées de l'ensemble des données disponible. Ce type de phénomène peut par exemple s'observer lors du comptage du nombre de jours passés à l'hôpital par des patients atteints d'une certaine maladie. Dans ce cas précis, il n'est pas possible de connaitre le nombre de patients atteints de cette même maladie et qui n'ont pas eu besoin d'aller à l'hôpital.

[8] https://www.census.gov/

C'est exactement ce même problème qui est rencontré avec l'exploitation de la base PRC pur modéliser la fréquence des incidents cyber.

Pour estimer le nombre de sinistre, un modèle linéaire généralisé (GLM) prenant pour variables explicatives les secteurs d'entreprises (BSR, BSO, BSF) sera ajustée aux données. Cela permettra de prendre en compte les spécificités de chaque type d'organisation et ainsi d'obtenir une fréquence plus adaptée à chaque secteur.

Trois types de GLM seront testés sur les données : un GLM basé sur une loi de poisson 0-tronquée, un sur une loi géométrique 0-tronquée et un dernier sur une loi binomiale négative 0-tronquée.

4.3.5.1 Modèles linéaires généralisés

Les modèles linéaires généralisés sont des méthodes paramétriques utilisées pour étudier la liaison existante entre une variable réponse Y à laquelle est associée une loi de probabilité et un ensemble de variables explicatives X qui peuvent être qualitatives ou quantitatives.

Notons :

- $Y = (Y_1, \dots, Yn)'$ vecteur colonne représentant les variables réponses,
- $X_i = \left(X_{i,1}, \dots, X_{i,p}\right)'$ vecteur colonne représentant les variables explicatives pour la variable réponse Y_i
- X la matrice de taille n x p dont les lignes sont les lignes X_i'
- $\beta = (\beta_1, \dots, \beta_p)'$ vecteur colonne représentant les p paramètres du modèle
- ϵ le terme d'erreur, c'est-à-dire $\epsilon = Y - E[Y|X]$
- $\mu\,(X) = E[Y|X]$

Un modèle est dit linéaire généralisé s'il vérifie les hypothèses suivantes :

- $Y \mid X = x \sim \mathbb{P}_{\theta,\varphi}$ appartient à une famille exponentielle ;

- $g\big(\mu(X)\big) = g(E[Y|X]) = X\beta$, g étant une fonction bijective appelé fonction lien.

La formule du modèle linéaire généralisé est :

$$g(E[Y|X]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x'\beta$$

Les coefficients β du modèle sont estimés à l'aide de la méthode du maximum de vraisemblance. Cette dernière consiste à maximiser la probabilité d'obtenir les réalisations observées i.i.d ($y_1, \dots, y_n$) sur un échantillon de taille n.

L'utilisation des GLM pour la modélisation est assez bénéfique car elle permet de maintenir la simplicité des modèles linéaires tout en autorisant une forme plus générale. De même, les coefficients obtenus en utilisant les GLM sont estimés par la maximation d'une vraisemblance provenant d'une famille exponentielle de lois, ce qui permet une estimation plus précise.

4.3.5.2 Poisson zéro-tronquée

La loi de poisson apparaît souvent comme premier choix lors de la modélisation d'événements rares, d'où son utilisation ici. Un avantage du processus de poisson est qu'il dépourvu de mémoire : la survenance d'un incidents cyber dans la seconde suivante est indépendante du passé, bien que n'étant pas toujours réaliste.

Soit X une variable aléatoire suivant une loi de poisson zéro-tronquée de paramètre λ. Elle possède la fonction de masse ci-dessous :

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!\,(1 - e^{-\lambda})}\ , x = 1{,}2\ \ldots$$

On note X ~ P (λ)

L'espérance d'une loi de poisson tronquée est donnée par la formule :

$$\mu = \frac{\lambda}{(1 - e^{-\lambda})^{2}}$$

4.3.5.3 Binomiale négative zéro-tronquée

Tout comme la loi de poisson, la loi binomiale négative est souvent utilisée dans les processus de comptages. Elle constitue une bonne alternative au processus de poisson lorsque le modèle utilisé présente une variabilité plus grande que la moyenne des données (sur dispersion).

Soit X une variable aléatoire suivant une loi binomiale négative zéro-tronquée. Elle possède la fonction de masse ci-dessous :

$$P(X = x) = \frac{\Gamma(x + r)\,p^{r}\,(1 - p)^{x}}{\Gamma(x)\,x!\,(1 - p^{r})}$$

Avec p la probabilité de succès comprise entre [0,1], r le nombre de succès et x le nombre d'échecs enregistrés avant l'obtention du r-ième succès. Dans la modélisation de la loi binomiale négative tronquée, le paramètre r ne sera pas forcé à être un nombre entier.

On note X ~ Negbin (r,p)

L'espérance d'une loi binomiale négative tronquée est donnée par la formule :

$$\mu = \frac{r(1 - p)}{p(1 - p^{r})}$$

4.3.5.4 Géométrique zéro-tronquée

La loi géométrique classique modélise le premier temps de succès dans une suite d'expérience de Bernoulli de paramètre p ϵ] 0,1[.

Si X est une variable aléatoire suivant une loi géométrique zéro-tronquée, elle possède la fonction de masse ci-dessous :

$$P(X = x) = p(1-p)^{x-1}\,, x = 1{,}2 \ldots, \text{p} \in]\,0{,}1[\text{ et P}(1) = 1 \; quand \; p = 1$$

On note X ~ G (p)

Son espérance est donnée par la formule suivante :

$$\mu = \frac{1}{p}$$

4.3.5.5 Application des modèles et résultats

Les GLM testés sur les données peuvent s'écrire :

$$g(E[N|X]) = X\beta\,, avec \begin{cases} N \sim P(\lambda) \; et \; g(x) = \log(x) \\ ou \\ N \sim Negbin(r,p) \; et \; g(x) = \log(x) \\ ou \\ N \sim G(p) \; et \; g(x) = \log(x) \end{cases}$$

Une fois les paramètres de ces trois modèles estimés, il est essentiel de vérifier que ces modèles sont valides et reflètent bien la réalité des données. Plusieurs critères permettent d'évaluer la qualité d'ajustement d'un modèle statistique. Dans le cadre de ce mémoire, les critères d'évaluation qui seront pris en compte sont la déviance et l'AIC.

AIC

Le critère d'information d'Akaike ou AIC a été proposé en 1973 par Hirotugu Akaike et se base sur le maximum de vraisemblance tout en pénalisant les modèles comportant un nombre élevé de paramètres. En effet, ces derniers sont critiqués car ils généralisent mal et sur-apprennent les données.
La formule de l'AIC est :

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

k étant le nombre de paramètres du modèle et L le maximum de la fonction de vraisemblance du modèle.
L'objectif de ce critère est la construction d'un modèle statistique parcimonieux reposant sur un équilibre entre simplicité (utilisation d'un nombre faible de paramètres) et qualité d'ajustement (utilisation du nombre de paramètres nécessaires). Parmi plusieurs modèles, le plus adéquat est celui qui a l'AIC le plus faible.

Déviance

Une façon classique d'évaluer la qualité d'ajustement d'un GLM est de comparer le modèle M ajusté à un modèle plus général ayant autant de paramètres que de variables réponse : le modèle saturé $\widetilde{M}$. Il s'agit d'un GLM ayant la même distribution et la même fonction de lien que le modèle M. En notant $\tilde{l}$ et $\hat{l}$ les vraisemblances respectives des modèles $\widetilde{M}$ et M, la déviance est définie comme :

$$\Delta = 2 \log (\tilde{l} - \hat{l})$$

Ainsi une déviance trop élevée suggère que le modèle estimé décrit mal les données par rapport au modèle saturé. Or d'après le théorème de Wilks, si les hypothèses du GLM sont vérifiées, alors :

$$\Delta \xrightarrow{\mathcal{L}} \chi^2_{n-p}$$

P étant le nombre de paramètres à estimer.

Comparer les modèles $\widetilde{M}$ et M revient alors à comparer la déviance à une loi du χ^2 au bon degré de liberté. Plus la déviance est grande, plus la qualité de l'ajustement est mauvaise. Ainsi, de deux modèles, celui qui a le meilleur ajustement est celui qui à la déviance la plus faible.

Ci-dessous les déviances et AIC des trois GLM testé sur les données :

Modèles	Poisson 0-tronquée	Negbin 0-tronquée	Géom 0-tronquée
AIC	384.07	375.78	375.97
Déviance	378.07	367.78	369.97

Tableau 15 : AIC et Déviance des modèles testés

Au regard des critères énumérés précédemment (AIC et Déviance), le modèle GLM basé sur loi binomiale négative 0-tronquée semble être plus adapté pour modéliser la fréquence des incidents cyber. Il sera donc retenu pour la suite de ce mémoire.

La fréquence annuelle de sinistres a été estimée en considérant que les entreprises de la base **ont été exposées pendant 6 ans**, bien que la base PRC recense des incidents survenus entre 2005 et 2019 (14 ans). En réalité, la figure 24 montre qu'aucune source d'information n'a été stable pendant plus de 6 ans. Seul les médias ont enrichi la base de données plus ou moins régulièrement entre 2012 et 2018.

Les résultats obtenus avec ce modèle ainsi que la valeur des paramètres estimés sont résumés dans le tableau suivant :

Type d'organisation	Coefficients estimés	Intervalle de confiance (95 %)	Espérance sinistre
BSF	- 0.51	[- 0.76 , - 0.25]	0.8 %
BSO	- 0.47	[- 0.72 , - 0.22]	0.9 %
BSR	- 0.58	[- 0.85 , - 0.31]	0.5 %

Tableau 16 : Résultats du GLM loi binomiale négative 0-tronquée

Le tableau révèle que les entreprises du secteurs BSO ont plus de chances d'être victimes d'attaques cyber que les autres entreprises. Cela pourrait se justifier par la diversité d'activités exercées par les entreprises de ce secteur.

Il faut également noter que, il n'a pas été jugé pertinent de réadapter les niveaux de fréquence obtenu au marché français, bien que les estimations aient été effectuées à l'aide de données américaines. En effet, du fait de la mondialisation importante existante dans le secteur informatique, il est tout à faire probable que les niveaux de digitalisation des entreprises françaises et américaines soient similaires. Cette décision permettrait également de rester prudent dans l'estimation de la mesure du risque et de prendre en compte le caractère nouveau et évolutif du risque cyber.

De même, bien que la base PRC semble ne contenir que des incidents dont sont victimes des entreprises de grandes tailles, les fréquences estimées ne seront pas réajustées à la cible de l'offre d'assurance proposée dans le cadre du présent mémoire (entreprises de moins de 20 salariés). Ce choix est avant tout basé sur le fait que, le risque cyber est complexe, protéiforme et changeant. Ainsi, aucun moyen ne permet de justifier avec certitude qu'une plus grande entreprise est plus susceptible d'être victime d'attaques cyber qu'une plus petite. Toutes les tailles d'entreprises ont des raisons d'être attrayantes pour les hackers. Aussi, une des principales causes des incidents cyber sont les erreurs humaines : que ce soit au sein d'une grande ou d'une petite entreprise, un employé pourrait commettre des erreurs.

Par ailleurs, les fréquences de sinistres estimées dans le tableau 17 ne prennent pas en compte le type d'événement survenu. Il serait judicieux afin d'améliorer la précision des estimations faites de vérifier l'impact des différents types d'attaques sur la fréquence de sinistres. Il se pourrait que dans la réalité, certaines attaques soient plus fréquentes que d'autres.

Ainsi, dans la suite du mémoire, il sera considéré que, lors de la survenance d'un incident cyber, sa nature est déterminée par une variable multinomiale aléatoire dont les paramètres dépendent du type d'organisation. Si T désigne le type d'organisation d'une entreprise victime d'incident cyber et A le type d'attaque subie, alors :

$$P(A = a \,|T = t) = \frac{e^{\beta_{t,0} + \beta_{t,a}}}{\sum_{a'} e^{\beta_{t,0} + \beta_{t,a'}}}$$

Avec $\beta_{t,0}$ qui correspond à un type d'organisation de référence. La référence choisie ici est le secteur BSR.

Le modèle multinomial réalisé conduit à l'estimation d'un grand nombre de paramètres avec très peu de données pour les calibrer. Face à cette contrainte, Farkas recommande dans son article, l'utilisation de la méthode de réduction de dimension de LASSO pour réduire le nombre de paramètres à estimer (Sébastien, Olivier, & Maud, 2020). Cette technique consiste globalement à pénaliser la log-vraisemblance du modèle en appliquant une pénalité de dimension 1 aux coefficients $\beta_{t,a}$ estimés.

Ci-dessous un tableau recensant les résultats de l'estimation des coefficients $\beta_{t,a}$ de la loi multinomiale testées sur les données :

Type d'attaques	CARD	DISC	HACK	INSD	PHYS	PORT	STAT
Intercept	-1.02	0 .67	1.58	0.13	-0.88	1.11	-0.49
BSO	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BSF	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Tableau 17 : Coefficients estimés loi mutinomiale

Le type d'organisation pris comme référence (BSR) semble avoir plus d'impact sur les fréquences de chaque type d'attaque que les autres.

Afin de mieux comprendre l'importance des coefficients, le calcul des probabilités de chaque type d'attaque est utile. Ci-dessous les résultats obtenus :

Type d'attaques	CARD	DISC	HACK	INSD	PHYS	PORT	STAT
Intercept	0.003	0.098	0.61	0.033	0.004	0.241	0.01
BSO	0.029	0.157	0.392	0.091	0.033	0.246	0.049
BSF	0.029	0.157	0.392	0.091	0.033	0.246	0.049

Tableau 18 : Fréquences estimées des types d'attaques

Les probabilités calculées sont assez faibles pour tous les types d'attaques. Cependant, tout comme c'était le cas dans la base de données initiales, les attaques de type Hacking semblent être les plus fréquentes. Par exemple, si une entreprise du secteur BSR est victime d'une attaque cyber, il y a 61% de chance qu'elle soit de type Hacking. Il sera important de retenir ces informations pour le suivi de l'évolution du portefeuille d'assurance qui sera construit dans le cas où une segmentation sera réalisée sur le coût moyen des sinistres engendrés par chaque type d'attaques.

Cependant, l'estimation faite ne prends pas en compte les attaques cyber qui existent mais ne sont pas répertoriées dans la base PRC.

4.3.5.6 Limites du modèle de fréquence

Tout comme les modèles de coûts construits, les modèles de fréquences présentent eux aussi quelques limites.

Le premier point à retenir est que les fréquences ont été évaluées sur la base de données relatives à une population assez générale. Or, en pratique l'actuaire en tarification évalue l'exposition de son risque par rapport à la population qui sera assurée. Le fait de ne pas prendre en compte cette réalité dans la construction des modèles empêche d'adapter le tarif final obtenu aux spécificités des entreprises ciblées par l'offre d'assurance. Toutefois, le risque cyber étant atypique par rapport aux risques traditionnellement étudiés, cela pourrait ne pas représenter une véritable limite.

De plus, du fait de l'instabilité des sources d'informations alimentant la base et de l'inconsistance des données, il n'a pas été possible d'identifier avec exactitude, la durée d'exposition des entreprises de la base. Ainsi, la fréquence obtenue est assez approximative et devra être réajustée avec l'évolution du portefeuille d'assurance.

Par ailleurs, les phénomènes d'accumulation qui pourraient entrainer une augmentation brusque de la fréquence de sinistre et nuire à l'équilibre financier du portefeuille n'ont pas été pris en compte dans les modèles.

4.3.6 Impact de la prévention

La mise en place de mécanismes de cybersécurité efficaces au sein des entreprises est bénéfique pour prévenir le risque cyber et en atténuer les conséquences. Ainsi, dans l'offre d'assurance proposée dans le cadre de ce mémoire, certains services de prévention seront obligatoires à la souscription. Ces services de préventions seront fournis aux entreprises assurées grâce à la collaboration de Moonshot avec un prestataire en cyber sécurité. La mise en place de ces services de prévention pourrait avoir un impact important sur le coût des potentiels incidents cyber dont seront victimes les entreprises qui souscriront à l'offre. Les **services de préventions obligatoires** retenus sont :

- Formation des équipes à la cyber sécurité
- Authentification multifacteur
- Mise à disposition d'experts en cyber sécurité

L'édition 2022 du rapport Ponemon permet d'obtenir des informations relatives à l'impact de ces différents services de préventions sur le coût moyen d'un incident cyber. Ci-dessous un résumé de l'impact financier de ces différents moyens de prévention :

Services	Impact sur le coût moyen
Formation des équipes	- 247 758 $ US
Authentification multifacteur	- 186 765 $ US
Experts en cybersécurité	- 136 244 $ US

Tableau 19 : Impact des services de prévention sur le coût moyen d'un incident cyber
Source : (Ponemon Institute , 2022)

Il est également mentionné dans le rapport que la souscription d'une entreprise à une offre d'assurance pourrait réduire le coût d'un incidents cyber de **240 488 $ US**. Cela pourrait être dû à la mise en place par les assureurs de mécanismes de suivi du risque en temps réel et de sensibilisation des entreprises assurées.

Au total, l'inclusion de services de prévention pourrait réduire le coût moyen de sinistre de **811 255 $ US**.

4.3.7 Modélisation de la prime pure de la garantie « Atteinte aux données »

Cette section sera consacrée à la construction de la grille tarifaire de l'offre d'assurance proposée. Les primes pures de chaque secteur d'organisation seront estimées selon la méthode de fréquence coût présentée en section 3.

Rappelons que les caractéristiques de l'offre comprennent la mise en place d'une franchise à 100 données compromises et un plafond de sinistre à 50 000 données.
Ci-dessous un résumé des coûts moyens de sinistres après mise en place de l'impact de la prévention :

Secteurs	Coût moyen de sinistre
BSF	1 234 525.78 €
BSO	763 993.81 €
BSR	1 199 367.83 €

Tableau 20 : Coûts moyens de sinistre après impact de la prévention

Les résultats obtenus pour l'estimation des primes pures annuelles d'assurance par secteur sont résumés dans le tableau ci-dessous :

Secteurs	Fréquence	Franchises	Limite	Prime pure
BSF	0.8 %	13 263 €	6 603 000 €	9 770.10 €
BSO	0.9 %	10 480 €	5 225 000 €	6 781.62 €
BSR	0.5 %	13 152 €	6 576 000 €	5 931.08 €

Tableau 21 : Primes pures annuelles par secteurs d'activités

Une fois les primes évaluées, il est important de les comparer aux tarifs des contrats d'assurance cyber existants sur le marché afin d'évaluer le positionnement de l'offre d'assurance proposée. La société Stoïk fournit dans un rapport, un résumé des tarifs de son offre d'assurance par secteur d'activité des entreprises :

Secteur d'activité	Prix Stoïk / an
E-commerce	2 061 €
Restaurant, bar et discothèque	1 383 €
Industrie manufacturière	1 685 €
Création de sites web	1 685 €
Services et fonds d'investissement	2 061 €

Tableau 22 : Tarifs annuels contrats d'assurance cyber Stoïk. Source : (Stoïk, 2022)

Les tarifs ci-dessus sont basés sur les critères ci-dessous :

1. Chiffre d'affaires compris entre 30 000€ et 35 000€
2. 30 employés
3. Profil de risque optimal
4. Sans les options RC transmission de virus et cyber-fraude (compter +10% avec les options)
5. Franchise : 1000 €
6. Plafond de garantie : 100 000€ (Stoïk, 2022)

L'observation faite ici est que les tarifs fournis par Stoïk sont basés sur des niveaux de franchises ainsi que de plafonds d'indemnisation différents de ceux de l'offre d'assurance construite. De ce fait, il faut réadapter la prime d'assurance calculée en se basant sur les mêmes critères de tarification que ceux utilisés par Stoïk. En fixant une franchise à 1000 € (équivalent à environ 10 données compromises) et un plafond d'indemnisation à 100 000 € (équivalent à 800 données compromises environ), la prime commerciale obtenue pour le secteur BSF est de 233.52 €/an environ.[9] Cette prime est inférieure à celle de l'offre de Stoïk qui propose une prime annuelle de 2 061€ pour une société qui pourrait être classée dans le secteur BSF.

L'écart entre les montants de primes provient probablement du fait que, les primes estimées dans le présent mémoire ne prennent en compte que la garantie *« atteinte aux données »*. Ainsi, la comparaison entres les tarifs est assez difficile à réaliser. Toutefois, elle permet d'obtenir un ordre de grandeur des niveaux de primes fixés sur le marché.

4.3.8 Etude de sensibilité des tarifs obtenus

Afin d'évaluer la soutenabilité et la viabilité des tarifs obtenus, il serait judicieux d'effectuer une étude de la sensibilité de ceux-ci à différents facteurs tels que le niveau de fréquence considéré. Au sein de l'entreprise, cette étude sera effectuée pour les tarifs obtenus dans les trois secteurs d'activités retenus (BSF, BSO, BSR). Cependant, dans le cadre de ce mémoire, seule la sensibilité de la prime dans le secteur BSF sera analysée.

L'étude de sensibilité réalisée est basée sur la variation de la fréquence de sinistre car il s'agit de du paramètre qui présente la plus grande incertitude.

Avant de faire varier le niveau de fréquence, il serait pertinent dans un premier temps de vérifier si les fréquences estimées dans chaque secteur d'entreprise sont cohérentes avec celles observées sur le marché. Cela permettrait à la fois d'observer si le niveau d'exposition au risque a été sous-estimé ou surestimé mais aussi d'avoir une idée des différents niveaux de fréquences à tester dans l'étude de sensibilité.

Coalition, l'un des principaux fournisseurs d'assurance cyber aux Etats Unis met à disposition son rapport annuel de sinistre déclarés au sein de son portefeuille d'assurance. Il en ressort que les fréquences moyennes de sinistre enregistrés en cas de violations de données dans les secteurs BSF et BSO étaient respectivement de 1.058 % et 0.78 % en 2021. Ces résultats sont proches des estimations faites avec le modèle GLM utilisé. Ainsi, les niveaux de fréquences prédits semblent être réalistes. Aucune information n'est donnée concernant le secteur BSR.

Afin d'évaluer l'impact de la fréquence sur le niveau de prime, trois niveaux de fréquences seront considérés. Dans un premier temps un chargement léger ($\pm$ 0.1 %) sera appliqué, ensuite un changement moyen ($\pm$ 0.5 %) et enfin un changement plus important (+ 1 %). Dans cette partie, c'est l'effet fréquence qui sera analysé : les différentes variations de tarifs seront obtenues à coût moyen inchangé.

[9] Ce résultat a été obtenu grâce une étude de la structure du tarif que sera présentée en annexe

Ci-dessous les résultats obtenus :

Variation	**- 0.5 %**	**- 0.1 %**	**+ 0 %**	**+ 0.1 %**	**+ 0.5 %**	**+ 1 %**
Fréquence	0.3 %	0.7 %	0.8 %	0.9 %	1.3 %	1.8 %
Prime pure	3 663.79 €	8 548.83 €	9 770.10 €	10 991.36 €	15 876.41 €	21 982.72 €
S/P	27.38 %	63.87 %	73 %	82.12 %	118.62%	164.25 %

Tableau 23 : Analyse de la sensibilité du tarif

Les hypothèses de calculs du ratio Sinistre/Prime (S/P) seront précisées en annexe.

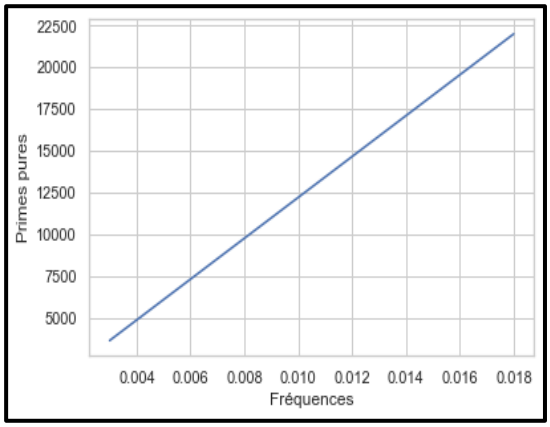


Figure 41: Evolution de la prime pure par rapport à la fréquence de sinistres

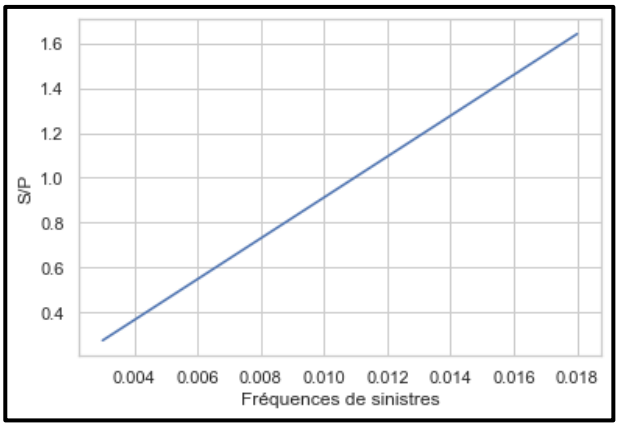


Figure 42 : Evolution du S/P par rapport à la fréquence de sinistres

Il ressort de toutes ces informations, que la prime pure évolue assez rapidement par rapport à l'évolution de la fréquence de sinistre. Une légère variation de la fréquence de sinistre entraine une variation considérable du niveau de prime pure ainsi qu'une forte dégradation du S/P. Cela montre que le tarif obtenu est assez sensible aux variations de la fréquence. Ainsi, une variation positive de la fréquence estimée ne serait-ce que d'un demi-point de pourcentage rendrait le risque insoutenable pour l'assureur (118% de S/P).

Il serait alors judicieux pour l'assureur de limiter au maximum son niveau de sinistralité. Ainsi, un pilotage renforcé devra être mis en place en vue d'assurer une analyse fine de l'évolution de la sinistralité au sein du portefeuille. De même, il est nécessaire au vue des résultats de l'analyse de sensibilité de mettre en place un programme de réassurance.

D'autres mécanismes de mitigation du risque étudié pourraient également être mis en place. Au nombre de ces mécanismes figurent : la mise en place dans le contrat d'assurance de clauses de révision infra-annuelles du tarif ou encore de clauses permettant à l'assureur de rompre le contrat en cas de sinistralité fortement dégradée.

4.3.9 Evolution du risque

Une étape importante du travail de l'actuaire en charge de la tarification d'un produit d'assurance est d'analyser l'évolution du risque qu'il étudie. Cette étape consiste principalement à définir les critères d'évolution du tarif construit et à améliorer les modèles

de tarification utilisés. Dans le cas du risque cyber, il n'est toujours aisé de définir des règles strictes concernant les critères d'évolution du tarif. Il est néanmoins possible de définir certains indicateurs sur lesquels seront basés l'amélioration des modèles construits.

Concernant les coûts des incidents évalués, l'évolution du portefeuille d'assurance permettra de disposer d'un historique de sinistres et ainsi de réaliser une segmentation plus fine des tarifs selon les caractéristiques des entreprises (nombre d'employés, chiffres d'affaires etc…).

Ainsi, chaque entreprise aura un tarif d'assurance plus adapté à son profil de risque, permettant ainsi de limiter le phénomène d'antisélection. Le fait de disposer d'un historique de sinistre permettra également d'analyser l'impact réel du de nombre de données violées sur le coût marginal de la donnée et ainsi, d'améliorer les estimations faites dans les modèles de coûts construits.

Enfin, grâce aux données récupérées, il sera possible d'identifier les potentielles spécificités du marché de l'assurance cyber français par rapport à l'américain et ainsi de réadapter si besoin les primes d'assurances calculées aux réalités des entreprises françaises.

5 ESTIMATION DE LA FREQUENCE DE SINISTRE D'UNE GARANTIE DE PERTE D'EXPLOITATION

Pour rappel, la garantie perte d'exploitation couvre les pertes liées à la baisse ou à l'arrêt de l'activité économique d'une entreprise à la suite d'un événement donné. Dans le cadre de l'offre d'assurance proposée dans le présent mémoire, seules les pertes d'activité engendrées par des cyber attaques seront couvertes.

Pour cette garantie, la fréquence de sinistre sera approchée par le pourcentage d'entreprises qui subissent une interruption d'activité suite à une attaque cyber. Plusieurs éléments peuvent faciliter la perte d'activité d'une entreprise après une cyber attaque.

En effet, cela pourrait dépendre soit de la nature de l'attaque, soit des mécanismes de prévention mis en place par l'entreprise attaquée ou encore des raisons ayant motivés l'incident. Par exemple, des attaques informatiques ayant pour motif principal l'enrichissement de l'attaquant (hacker) seraient potentiellement plus susceptibles d'entrainer une perte d'exploitation. C'est le cas des attaques par ransomware au cours desquels les hackers bloquent le système informatique d'une entreprise afin de la contraindre à leur verser une rançon pour pouvoir reprendre ses activités.

A l'inverse, les attaques de type cyber espionnage consistent tout simplement à infiltrer clandestinement les systèmes informatiques d'une organisation afin de s'emparer de ses données sans pour autant chercher à l'empêcher de poursuivre ses activités.

Par ailleurs, une entreprise ayant mis en place un mécanisme de cyber sécurité efficace serait probablement plus susceptible de limiter les dégâts causés par une attaque informatique et ainsi éviter une interruption de ses activités.

Cependant, étant donné la volatilité du risque cyber, il est difficile d'établir un lien direct entre les différents facteurs énumérés précédemment (type d'attaque, motif de l'attaque…) et la survenance ou non d'une perte d'activité. Avec l'évolution permanente des méthodes de cyber attaques, ces dernières engendrent des conséquences de plus en plus graves et diversifiées. De ce fait, il serait plus prudent de considérer que tous les types d'attaques ont la même probabilité de causer une perte d'exploitation.

Aussi, il est complexe de définir avec précision la motivation des attaquants. A titre d'exemple, bien qu'une attaque de type ransomware puisse être causée par un hacker qui ne cherche qu'à s'enrichir, elle peut également provenir d'un concurrent qui cherche à nuire à l'organisation attaquée. Ainsi, dans la suite du mémoire, la motivation des attaquants ne sera pas prise en compte dans la modélisation du risque de perte d'exploitation à la suite d'une attaque cyber.

5.1 PRESENTATION DE L'APPROCHE BAYESIENNE

Estimer le pourcentage d'entreprises victimes de perte d'exploitation après une attaque informatique s'est avéré être une tâche complexe. Cela est en partie dû à la non-disponibilité de base de données publiques fiables sur le risque cyber et à la non-déclaration des incidents par certaines entreprises victimes de cyber attaques. De plus, il est parfois difficile de différencier les pertes d'exploitation dues à des incidents cyber et celles qui pourraient être provoquées par un dysfonctionnement des systèmes informatiques des entreprises. Cela peut s'observer notamment dans le cas des attaques informatiques silencieuses (cyber espionnage, malware…).

La modélisation de la fréquence de sinistre de la garantie perte d'exploitation se fera à l'aide de méthodes bayésiennes. Ces dernières permettent globalement de déduire la probabilité d'un événement en se basant sur des probabilités d'autres événements déjà estimées : on cherche à exprimer ce que l'on sait sur les inconnus du problème sachant ce qui est connu (ou assumé comme tel). Ces méthodes reposent sur la notion de probabilité personnelle et leur cohérence est garantie par des règles mathématiques du calcul des probabilités.

A l'inverse des approches statistiques fréquentistes usuelles, basées uniquement sur des données déjà observées dans le passé, les méthodes dites bayésiennes utilisent le théorème de Bayes pour combiner des opinions et des faits de toute provenance avec des observations en lien avec l'étude considérée.

Ainsi, ces techniques résolvent certaines difficultés rencontrées par les méthodes classiques. En particulier, elles permettent de pallier à la faible quantité de données disponibles dans le cas de l'étude de nouveaux risques tels que le risque cyber. Elles servent également de moyen pour traiter des données d'observations que l'on juge faiblement informatives et peuvent être appliquées à des problèmes dont la structure est trop complexe pour que les méthodes classiques puissent s'utiliser.

L'approche bayésienne peut être décrite comme une généralisation de l'approche fréquentiste. En effet, les paramètres ne sont plus des valeurs fixes inconnues, mais des variables aléatoires dont il faut déterminer la distribution. La distribution donnée au moment de la modélisation est dite ***a priori***. L'incorporation de l'information apportée par les données observées se fait par le calcul d'une distribution dite ***a posteriori*** à l'aide de la formule du théorème de Bayes.

5.1.1 Rappel sur les probabilités conditionnelles et formule de Bayes

Soient A et B deux événements aléatoires, tels que $\mathbb{P}(A)$ et $\mathbb{P}(B)$ sont non nuls. La probabilité de B conditionnellement à la réalisation de A ainsi que la probabilité de A conditionnellement à la réalisation de B sont données par les formules suivantes :

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \quad (1) \qquad\qquad \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (2)$$

$\mathbb{P}(A \cap B)$ étant la probabilité que les événements A et B se réalisent simultanément

De la formule 2, se déduit $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\ x\ \mathbb{P}(B)$

En remplaçant l'expression de $\mathbb{P}(A \cap B)$ dans la formule 1, on peut en déduire **la formule de Bayes** liant les probabilités conditionnelles $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$ suivante :

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\ x\ \mathbb{P}(B)}{\mathbb{P}(A)}}$$

5.1.2 Principe de l'approche bayésienne

Soit un modèle paramétrique statistique d'observations peu informatives, issu d'une variable aléatoire X (*v.a*) distribuée selon *f(x|θ)*, θ étant un paramètre inconnu appartenant à un espace Θ de dimension finie, et *x* l'ensemble des observations de X. On note $x = (x_1, \ldots., x_i, \ldots x_n)$; en supposant que l'on dispose d'un échantillon de taille n. Conformément à la statistique inférentielle, les x_i sont considérées comme étant des réalisations de variables aléatoires, notées X_i.

L'objectif de l'analyse statistique bayésienne est de modéliser le paramètre θ à travers une structure probabiliste. Pour ce faire, elle vise à se servir efficacement de l'information apportée par X sur le paramètre θ, pour ensuite construire des procédures d'inférence sur ce dernier. Ainsi, le modèle statistique paramétrique bayésien consiste en la donnée de la loi des observations x et d'une loi ***a priori***.

5.1.2.1 Loi des observations

La loi des observations est la loi conditionnelle de X sachant θ. Sa densité est notée ***f(x|θ)***. Si X est discrète, ***f(x|θ)*** représente ***Pr(X = x|θ)***. On fait l'hypothèse ici que les v.a X_i sont indépendantes conditionnellement à θ. Alors :

$$\boxed{f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)}$$

La loi des observations est égale à la vraisemblance notée **L(θ,x)**.

5.1.2.2 Loi a priori

L'information *a priori* sur le paramètre θ, désigne toute information disponible sur ce dernier autre que celles apportées par les observations. La connaissance *a priori* sur θ est entachée d'incertitude et s'exprime à travers une loi de probabilité nommée loi *a priori* dont la densité est notée ***π(θ)***. L'information *a priori* peut être issue de l'opinion d'experts ou bien construite sur la base de l'analyse d'autres données similaires ou encore sur l'avis du statisticien réalisant l'étude. Il existe deux types de loi à priori : les **lois informatives** et les **lois non informatives**.

Loi a priori informative et non informative

Une distribution *a priori* est dite informative si elle est non dominée par la vraisemblance et a un impact sur la distribution a posteriori. Les lois à priori informatives doivent être définies avec précaution dans la pratique et privilégient certaines valeurs de la quantité d'intérêt. Ces préférences sont généralement basées sur des études antérieures.

Cependant, très souvent, les informations a priori sur le modèle sont assez imprécises ou peu fiables rendant ainsi le choix de la distribution a priori très difficile. De même, dans certains cas, afin de construire un modèle le plus objectif possible, le statisticien préfère limiter au maximum les hypothèses subjectives et baser l'inférence sur le modèle d'échantillonnage seul.

Ainsi, en l'absence d'information a priori, des lois a priori non informatives sont utilisées afin de rester dans un cadre bayésien, bien que ne disposant pas d'information a priori. Différentes approches peuvent être utilisées pour le choix de l'a priori non informatif :

- Lorsqu'il n'y a pas d'a priori, il est naturel de proposer une loi uniforme sur θ car elle permet d'obtenir une probabilité égale aux intervalles de longueur l donnée : c'est le principe de la raison insuffisante (principe de Laplace). Cette approche semble être assez fiable.

- Figure également parmi les lois non informatives, la loi a priori impropre de Haldane $\pi(\theta) = [\theta(1-\theta)]^{-1} 1_{]0,1[}(\theta)$

- Jeffreys propose aussi une approche en 1960 basée sur l'information de Fisher $I(\theta)$, telle que la distribution apriori dans le cas uni dimensionnel est donnée par : $\pi^*(\theta) \propto I^{\frac{1}{2}}(\theta)$

Avec : $I(\theta) = E_\theta \left[\left(\frac{\partial \log f(x|\theta)}{\partial(\theta)} \right)^2 \right]$

5.1.2.3 Loi a posteriori

Il s'agit de la loi conditionnelle de θ sachant x. Sa densité est notée $\pi(\theta|x)$ et elle fournit l'information disponible sur θ après observation de X. Elle représente la conciliation de l'information a priori donnée par la fonction $\pi(\theta)$ avec celle de l'information tirée de l'échantillon. En appliquant la formule de Bayes, l'expression de la loi a posteriori est :

$$\pi(\theta|x) = \frac{\pi(\theta) * \pi(x|\theta)}{\int_\theta \pi(\theta) * \pi(x|\theta)\, d\theta} = \frac{\pi(\theta) * \pi(x|\theta)}{h(x)}$$

$h(x) = \int_\theta \pi(\theta) * \pi(x|\theta)\, d\theta$ est la loi marginale de x également appelée **constante d'intégration**. Etant donné que h(x) est indépendante de θ, la loi a posteriori peut s'écrire comme suit :

$$\pi(\theta|x) \propto \pi(\theta)\ x\ \pi(x|\theta)$$

Pour résumer toute ses informations, le mécanisme de l'inférence bayésienne comporte les étapes ci-dessous :

- Le choix d'un modèle relatif aux données disponibles et appartenant à une famille de lois de probabilité

- La détermination d'une distribution à priori du paramètre à estimer à l'aide de connaissances disponibles sur le sujet étudié (avis d'experts, essais antérieurs…)

- L'application du théorème de Bayes pour obtenir une distribution a posteriori

- L'estimation des paramètres à l'aide de la distribution a posteriori obtenue

Ce mécanisme peut être représenté à l'aide du schéma ci-dessous :

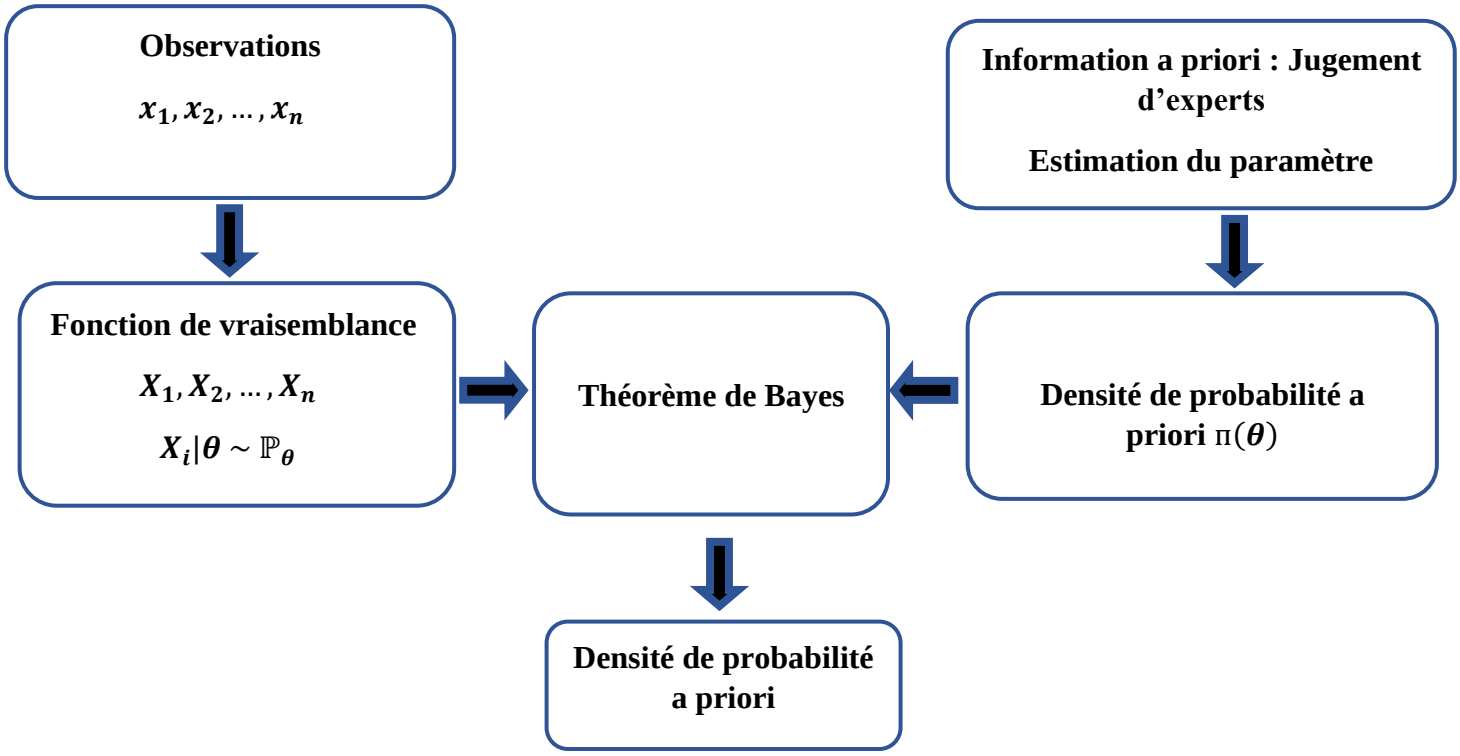


Figure 43 : Principe de l'approche Bayésienne

5.1.2.4 Opinion d'experts

La difficulté de l'application de l'inférence bayésienne réside principalement dans la détermination de la loi à priori. Elle est déterminée après un travail conjoint entre un expert sur le sujet d'étude et le statisticien chargé de la construction du modèle.

Généralement, un expert, c'est quelqu'un qui possède des connaissances approfondies dans un domaine particulier grâce à l'expérience. (CNRS) L'avis des experts est assez souvent utilisé en actuariat lors de la modélisation de nouveaux risques tels que le cyberrisque. Cela permet de donner une rigueur supplémentaire au modèle dès lors que très peu de données sont

disponibles ou que l'on doute de la fiabilité de celles-ci. D'ailleurs, l'ajout de l'opinion d'experts dans la construction de modèle mathématique a été étudiée par plusieurs mathématiciens. A titre d'exemple, dans son livre « *Experts in Uncertainty: Opinion and Subjective* », Cooke donne un aperçu historique du sujet tout en examinant comment l'opinion d'expert est utilisée aujourd'hui, comment l'incertitude d'un expert est ou devrait être représentée, comment la qualité et l'utilité de l'opinion d'expert peuvent être évaluées, et comment les opinions de plusieurs experts pourraient être combinées. (Cooke, 1991)

5.2 APPLICATION DE LA METHODE

5.2.1 Description des données

Les données d'observations utilisées pour la construction du modèle proviennent de la base VERIS présentée dans la partie 1 du mémoire. Le choix de cette base réside dans le fait qu'elle recense des milliers d'incidents cyber survenus dans le monde entier avec des variables qui permettent une description complète et détaillée de ces incidents. Elle a d'ailleurs fait l'objet de nombreuses études sur l'impact des cyber incidents. A titre d'exemple, Walker-Roberts et ses collègues se sont appuyés sur les informations de la base VERIS pour analyser l'étendue des risques engendrés par des incidents de cybersécurité. (Walker-Roberts, Hammoudeh, Aldabbas, & al., 2020).

La base étudiée regroupe 9134 incidents survenus entre 1971 et 2021. Elle comporte 2546 variables décrivant le **suivi des différents incidents** (numéro de l'incident, description de l'incident…), les **caractéristiques des entreprises victimes** (localisation, nombre d'employées, revenus…), la **description des incidents** (acteurs, conséquences…) ainsi que les différentes **conséquences des attaques sur les victimes**.

Afin d'évaluer la fréquence de perte d'exploitation survenant après une attaque cyber, il serait plus intéressant de se concentrer sur les variables décrivant les conséquences des attaques. En effet, bien que la portée et l'étendue réelles des conséquences de ces attaques peuvent être difficiles à mesurer, Veris s'appuie sur trois perspectives d'impacts afin de fournir une compréhension et une mesure des conséquences associées aux différents incidents. Ils cherchent ainsi à :

- catégoriser les différentes pertes subies
- estimer leur ampleur
- saisir une évaluation qualitative de l'effet global sur l'organisation.

La dernière perspective mise en place par Veris sera utilisée pour distinguer les incidents ayant causés une perte d'exploitation. Dans cette perspective, une évaluation des incidents selon leurs impacts sur l'organisation affectée est réalisée. Ainsi, les différents niveaux d'impacts recensés dans la base sont :

- **Insignificant** : Regroupant les attaques dont les impacts ont été absorbés par les activités quotidiennes des entreprises victimes

- **Distracting** : Regroupant les attaques ayant entrainé des pertes financières limitées mais qui ont contraint les entreprises victimes à passer du temps à trouver des solutions pour stopper l'incident plutôt que de poursuivre leurs activités normales.

- **Painful** : Regroupant les attaques ayant entrainé des pertes financières limitées mais qui ont contraint les entreprises victimes à passer du temps à trouver des solutions pour stopper l'incident plutôt que de poursuivre leurs activités normales.

- **Damaging** : Incidents ayant eu un impact réel et sérieux sur le résultat net des entreprises victimes ainsi que sur leur aptitude à gérer des revenus à long terme.

- **Catastrophic** : Incidents ayant contraints les entreprises victimes à stopper définitivement leurs activités.

- **Unknown** : Incidents dont les impacts sur les organisations victimes sont inconnus.

Dans le cas de la problématique étudiée dans cette deuxième partie du mémoire, seules les incidents classés dans les catégories **Catastrophic** et **Damaging** seront retenus comme ayant entrainé une perte d'exploitation. Ces dernières catégories semblent en effet, être les seules qui, à travers leurs définitions laissent penser à la survenance d'une interruption totale d'activité au sein des organisations victimes d'incidents cyber.

Sur les 9134 incidents reportés dans la base, seuls 9053 ont été confirmés. Ces derniers serviront de données d'observations pour la construction des modèles bayésiens. Ci-dessous une répartition du nombre d'incidents recensés dans la base par année :

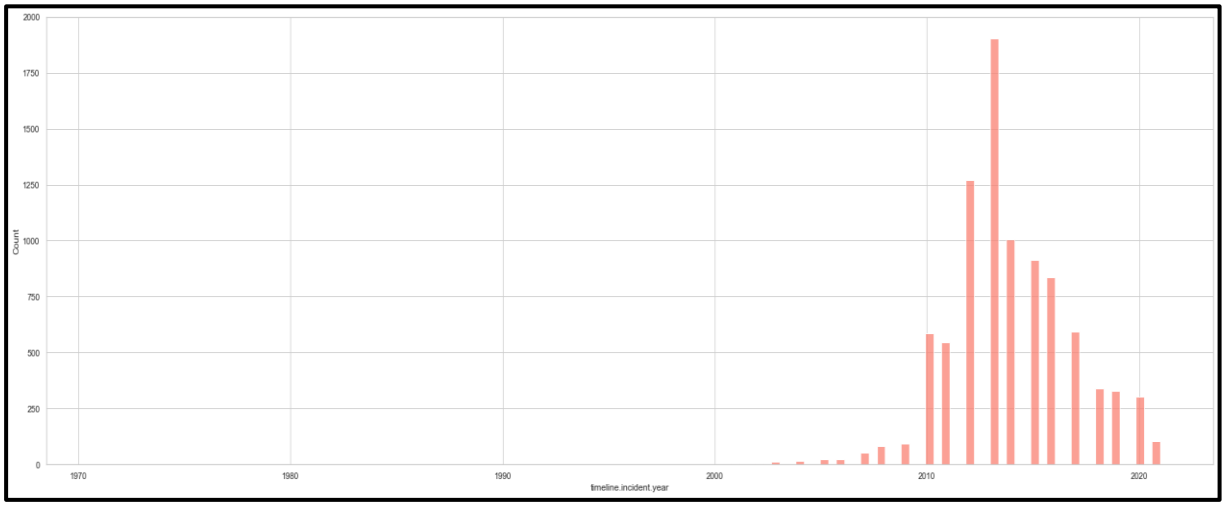


Figure 44 : Nombre d'incidents enregistrés par année dans la base VERIS

Ce graphique montre que l'essentiel des incidents recensés dans la base Veris se sont produits entre 2005 et 2021.

Afin de rester cohérent avec la modélisation faite dans le cas de la garantie ***atteinte aux données***, la fréquence de sinistre ici, ne sera estimée que pour les secteurs BSF, BSR, BSO. En effet, la base Veris dispose d'une colonne ***victim.industry.name*** qui permet d'associer à chaque entreprise victime de cyber attaque, un secteur d'activité. Ainsi, les secteurs d'activités retenus sont :

BSF : Finance

BSR : Retail, Trade

BSO : Utilities, Other Services, Accomodation, Entertainment, Information

Il serait à présent judicieux, d'analyser les impacts des incidents sur les entreprises appartenant aux secteurs énumérés.

Ci-dessous une répartition des incidents selon les niveaux d'impacts qu'ils génèrent :

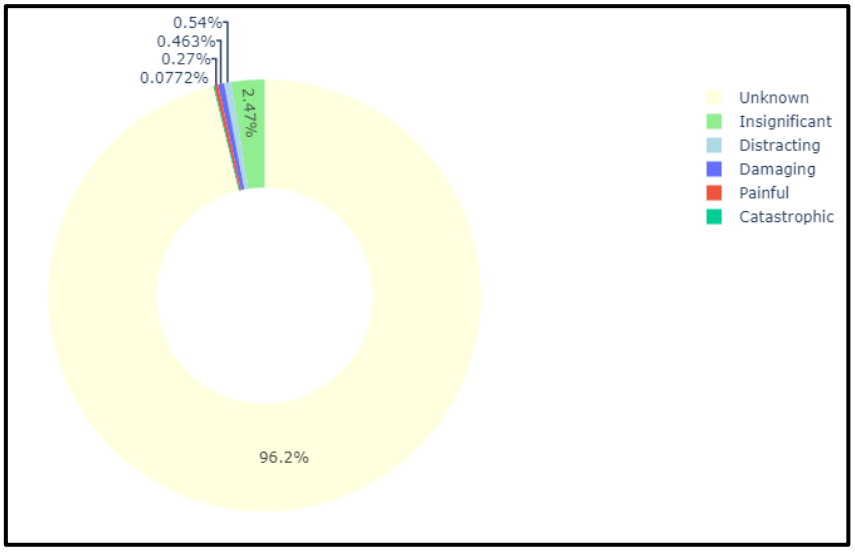


Figure 45 : Répartition des incidents selon leur impacts

Il ressort de cette représentation qu'il manque de nombreuses informations sur les conséquences engendrées par les cybers attaques sur les entreprises. Cela se justifie probablement par le fait que la base VERIS est remplie à la main. De ce fait, la personne qui recense un incident dans cette base pourrait ne pas avoir toutes les informations relatives à cet incident. C'est d'ailleurs pourquoi l'utilisation d'une approche Bayésienne pour modéliser la fréquence de sinistre est pertinente car cela permettra de limiter le biais que pourrait créer ce manque de données.

Ainsi, deux modèles bayésiens seront utilisés dans la suite du mémoire. Cela permettra de comparer deux résultats différents et d'analyser la pertinence des résultats obtenus avec chaque méthode.

5.2.2 Approche 1 : Utilisation d'une loi a priori informative

Soit θ la probabilité qu'une entreprise victime de cyber attaque subisse une perte d'exploitation à l'issue de celle-ci. Comme mentionné plus haut, l'estimation du paramètre θ se fera à l'aide des données d'observations de la base Veris. Soient des variables aléatoires X_i indépendantes, identiquement distribuées (i.i.d) et désignant les entreprises de la base. Les réalisations de ces variables aléatoires X_i seront notées x_i ; $x_i \in \{0,1\}$ étant l'état de l'entreprise i après une attaque cyber où $x_i = 1$ si l'entreprise est victime d'une interruption d'activité et $x_i = 0$ sinon. Ainsi : **$X_i \sim$ Bernoulli (θ)**

5.2.2.1 Loi a priori

Dans ce premier modèle, une loi a priori informative sera utilisée pour la construction du modèle. Lorsqu'il existe peu d'informations a priori, une option pour le choix de la loi a priori est d'utiliser une **loi a priori conjuguée**. Il s'agit d'une modélisation qui permet une manipulation explicite des lois a posteriori.

Définition : La loi des observations étant supposée connue, soit F une famille de loi de probabilité sur Θ. En supposant que la loi a priori appartienne à F, si la loi a posteriori appartient encore à F, on dit que **la loi a priori est conjuguée**.

Généralement, les lois a priori sont associées à un type particulier de loi d'échantillonnage : ces lois constituent des familles exponentielles. Une **famille exponentielle** à s-paramètres désigne toute famille de lois de distribution P_θ dont la densité à la forme suivante :

$$f(x|\theta) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x)$$

$\eta_i(.)$ et $B(.)$ étant des fonctions du paramètre θ et $T_i(.)$ des statistiques.

La loi binomiale par exemple appartient à la famille exponentielle.

Soit X ~ Binomiale (n, θ)

$$P(X = x \mid \theta) = C_n^x \, \theta^x (1-\theta)^{n-x}$$

$$P(X = x \mid \theta) = C_n^x \exp\{x \log \theta + (n - x)\log(1-\theta)\}$$

$$P(X = x \mid \theta) = C_n^x \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + n \log(1-\theta)\right\}$$

Avec : s = 1, $\eta_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $T_1(x) = x$, $B(\theta) = n \log\left(\frac{\theta}{1-\theta}\right)$ et $h(x) = C_n^x$ ■

Ci-dessous un tableau regroupant des exemples de lois a priori conjuguées naturelles pour quelques familles exponentielles :

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $N(\theta, \sigma^2)$	Normale $N(\mu, r^2)$	Normale $N\left(\left(\frac{\sigma^2\mu+r^2x}{\sigma^2+r^2}\right), \left(\frac{\sigma^2 r}{\sigma^2+r^2}\right)\right)$
Poisson $P(\theta)$	Gamma $G(\alpha, \beta)$	Gamma $G(\alpha + x, \beta + 1)$
Exponentielle $Exp(\lambda)$	Gamma $G(\alpha, \beta)$	Gamma $G(\alpha + k, \beta + t)$
Binomiale $B(n,\theta)$	Bêta $Be(\alpha, \beta)$	Bêta $Be(\alpha + x, \beta + n - x)$
Binomiale négative $Neg(m,\theta)$	Bêta $Be(\alpha, \beta)$	Bêta $Be(\alpha + m, \beta + x)$

Tableau 24 : Lois a priori conjuguées

Dans le cadre de cette étude, $X_i|\theta \sim$ Bernoulli(θ). La conjuguée naturelle de la loi de Bernoulli étant la loi bêta, on a : $\theta \sim Be(\alpha, \beta)$. Alors la densité de la loi a priori ici est :

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}\ 1_{[0,1]}(\theta)$$

où $\boldsymbol{B(\alpha,\beta)} = \frac{\boldsymbol{\Gamma(\alpha)\Gamma(\beta)}}{\boldsymbol{\Gamma(\alpha+\beta)}}$

Γ étant la fonction gamma d'Euler définie comme suit :

$$\Gamma(x) = \int_0^{+\infty} e^{-t}\ t^{x-1}\ dt\ ,\ \forall\ x \in \mathbb{R}$$

5.2.2.2 Calcul de la loi a posteriori

Etant donné le modèle bayésien défini plus haut, on a :

$$f(x|\theta) = \prod_{i=1}^{n} \mathbb{P}(X = x_i|\theta) = \theta^{s}(1-\theta)^{n-s}$$

n étant le nombre d'entreprises de la base Veris et $s = \sum_{i=1}^{n} x_i$.

D'après la formule de Bayes :

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}$$

Par calcul, on obtient $\int_{\theta} f(x|\theta)\pi(\theta)d\theta = \frac{B(a,b)}{B(\alpha,\beta)}$. (La démonstration sera effectuée en annexe.)
Avec $a = \alpha + s$ et $b = \beta + n - s$; α et β étant les paramètres de la **loi a priori**.

Par conséquent, la densité de la loi a posteriori s'écrit :

$$\pi(\theta|x) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} 1_{[0,1]}(\theta)$$

Alors, $\boldsymbol{\theta|x \sim Beta\ (a,b)}$ soit $\boldsymbol{\theta|x \sim Beta\ (\alpha + \sum_{i=1}^{n} x_i\ ,\ \beta + n - \sum_{i=1}^{n} x_i)}$, α et β étant les paramètres de la loi a priori.

L'estimateur de Bayes, encore appelé moyenne a posteriori est donné par la formule suivante :

$$\widehat{\theta_B} = E[\theta|x] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n}$$

Afin de déterminer cet estimateur il faudrait avant tout inclure l'opinion de l'expert considéré.

5.2.2.3 Elicitation de l'avis d'expert

Les experts auxquels qui pourraient être contactés dans le cadre de la problématique actuelle seront probablement des professionnels de la cyber sécurité. L'opinion de ces experts est requise dans le but d'obtenir des informations statistiques (moyenne, variance…) sur le paramètre d'intérêt θ. N'ayant pas pu obtenir des rendez-vous avec des experts en cyber sécurité capables de fournir des informations sur le paramètre θ , l'expertise sera ici basée sur les résultats obtenus dans le *« baromètre de la cybersécurité des entreprises »* réalisé par Opinionway pour le CESIN[10]. De cette étude sera déduite une estimation de l'espérance de θ ainsi qu'un intervalle de confiance à 95% pour cette valeur.

En effet, du 24 novembre au 21 décembre 2021, Opinionway réalise un sondage sur un échantillon de 282 entreprises membres du CESIN. Parmi les répondants au sondage figurent 8% de TPE/PME, 38% d'ETI et 54% de Grandes entreprises. D'après ce sondage, **54%** des entreprises interrogées ont subi au moins une attaque en 2021. Le rapport de Opinionway recense les conséquences subies par les entreprises victimes à la suite de ces cybers attaques. Ci-dessous une répartition de ces différentes conséquences :

[10] Club des Experts de la Sécurité de l'Information et du Numérique

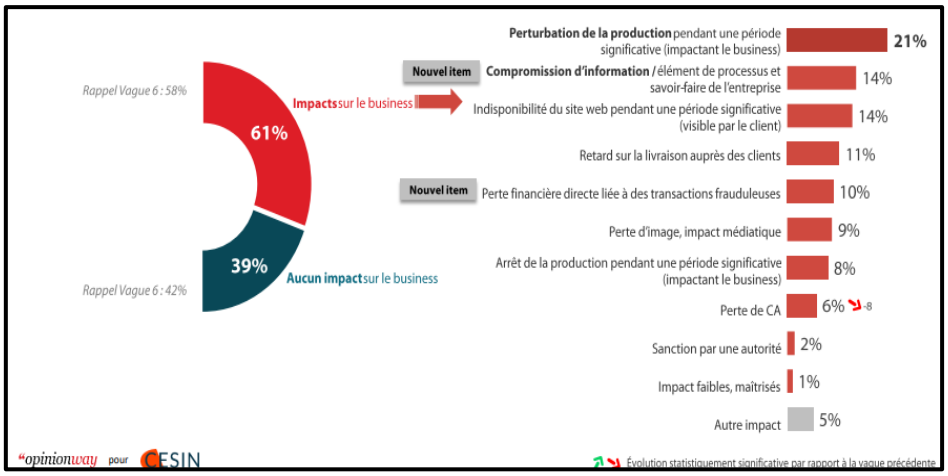


Figure 46 : Répartition des impacts des cyber-attaques sur les entreprises victimes
Source : (Opinionway pour CESIN, 2022)

Parmi les impacts répertoriés ci-dessus, les conséquences qui seront considérées comme de la perte d'exploitation sont :

- Arrêt de la production pendant une période significative (impactant le business) : 8%
- Indisponibilité du site web pendant une période significative (visible par le client): 14%

Ainsi, d'après ce rapport, **22%** des entreprises victimes de cyber attaque subissent une interruption d'activité à la suite de celle-ci. Alors, l'estimation de l'espérance de θ par les experts est : $\theta_{\mathbf{expert}} = \mathbf{22\%}$

<u>Construction d'un intervalle de confiance à 95% pour θ_{expert}</u>

D'après l'étude, 152 entreprises sont victimes d'une cyber attaque en 2021. Soient $X_1,\ldots X_{152}$ des réalisations de variables aléatoires i.i.d désignant l'état des entreprises interrogées après une cyber attaque. $X_i = 1$ si l'entreprise i subit une interruption d'activité due à une attaque cyber et $X_i = 0$ sinon.

Ainsi, si θ est la probabilité qu'une entreprise subisse une interruption d'activité à la suite d'une attaque, alors la moyenne empirique de θ est :

$$\hat{\theta} = \bar{X} = \frac{1}{152}\sum_{i=1}^{152} X_i$$

Des informations précédentes, il ressort que $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\mathbf{expert}} = \mathbf{22\%}$.

La variance empirique de θ est donnée par la formule ci-dessous :

$$\widehat{\sigma^2} = \frac{1}{152}\sum_{i=1}^{152}(X_i - \bar{X})^2$$

$$\widehat{\boldsymbol{\sigma^2}} = \frac{1}{152}\left[33 * (1 - 0{,}22)^2 + 119 * (0 - 0{,}22)^2\right]$$

$$\widehat{\boldsymbol{\sigma^2}} = \mathbf{0,17}$$

D'après le théorème de la limite centrale, $\frac{\bar{X}-\boldsymbol{m}}{\frac{\sigma}{\sqrt{n}}} \sim \boldsymbol{N(0,1)}$ N étant une loi normale, m la moyenne de l'échantillon et $\boldsymbol{\sigma}$ sa variance.

Ce théorème permet de conduire un intervalle de confiance à 100(1- α) % pour m à partir de l'égalité suivante :

$$1-\alpha = \mathbb{P}\left[-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right]$$

$$1-\alpha = \mathbb{P}\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

$z_{\frac{\alpha}{2}}$ étant le quantile d'ordre $\alpha/2$ d'une loi normale centrée réduite.

Application numérique

Pour α = 5%, on obtient un intervalle de confiance à 95% pour m = θexpert.

$$IC_{95\%} = \left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

$$\boldsymbol{IC_{95\%} = [0.15\,;0.28]}$$

Etant donné qu'un expert ne peut être sur à 100% des statistiques qu'il fournit sur le paramètre θ , la construction de l'intervalle de confiance précédent permet de prendre en compte l'incertitude de l'expert autour de la valeur de θ

Estimation des paramètres de la loi a posteriori

D'après les informations recueillies auprès de l'expert :

$$\begin{cases} E[\theta] = \dfrac{\alpha}{\alpha+\beta} = 0.22 \\ \mathbb{P}(a_o \leq \theta \leq b_o) = \displaystyle\int_{a_o}^{b_o} \pi(\theta|\alpha,\beta)\, d\theta = F_{\alpha,\beta}(b_o) - F_{\alpha,\beta}(a_o) = 0.95 \end{cases}$$

$a_o \; et \; b_o$ étant les bornes de l'intervalle de confiance construit plus haut ($a_o = 0.15 \; et \; b_o = 0.28$).

$F_{\alpha,\beta}(.)$ est la fonction de distribution jointe de la loi a priori Beta (α, β)

Une résolution numérique du système d'équations précédent permet d'obtenir les résultats suivants :

$$\boxed{\alpha = 33 \text{ et } \beta = 119}$$

α et β étant les paramètres de la loi a priori.

Les données d'observations de la base Veris montrent que 14 entreprises sont victimes de perte d'exploitation sur les 99 entreprises ayant subi des attaques cyber. Ainsi, en combinant ces informations avec celle de l'expert, les paramètres de la loi a posteriori sont :

$$\text{a} = \alpha + \sum_{i=1}^{99} x_i = 33 + 14 = 47 \quad \text{et} \quad \text{b} = \beta + n - \sum_{i=1}^{99} x_i = 119 + 99 - 14 = 204$$

Ainsi la loi a posteriori est la loi **Beta (47,204)**.

L'estimateur de Bayes est donc :

$$\widehat{\theta_B} = E[\theta|x] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n} = \frac{47}{204 + 47} = 0.1873$$

Ainsi, une estimation de la fréquence de sinistre pour la garantie perte d'exploitation est de **18.73%**.

Il serait intéressant pour l'étude réalisée d'analyser le poids de l'a priori dans la réponse bayésienne. Cela permettra de comprendre plus explicitement comment l'information a priori et l'information contenue dans les observations se combinent l'une à l'autre pour obtenir la réponse bayésienne. Une façon assez fréquente de reparamétrer la loi a posteriori obtenue de la façon suivante :

$$E[\theta|x] = \frac{\lambda}{\lambda + n} E[\theta] + \frac{n}{\lambda + n} \bar{x}$$

Soit :

$$E[\theta|x] = k\, E[\theta] + (1 - k)\, \bar{x}$$

Avec $\boldsymbol{\lambda} = \boldsymbol{\alpha} + \boldsymbol{\beta}$ et n la taille de l'échantillon

L'estimation bayésienne de θ peut donc être interprétée comme une moyenne pondérée de $\bar{x}$ et de la moyenne a priori $E[\theta]$. Le poids λ de $E[\theta]$ dans cette expression s'interprète comme la précision de l'a priori. Il faut remarquer également que, lorsque $\boldsymbol{\lambda} = \boldsymbol{n}$, l'expression ci-dessus devient :

$$E[\theta|x] = \frac{1}{2} E[\theta] + \frac{1}{2} \bar{x}$$

Ainsi, l'estimation bayésienne de θ se trouve exactement au milieu de l'intervalle $[E[\theta], \bar{x}]$.

Dans le cas où $\boldsymbol{\lambda} > \boldsymbol{n}$, l'estimation est plus proche de $E[\theta]$ que de $\bar{x}$ et inversement dans le cas contraire. L'influence de l'a priori sur $E[\theta|x]$ peut ainsi être examinée en s'intéressant aux situations limites : $\boldsymbol{\lambda} \to \mathbf{0}$ et $\boldsymbol{\lambda} \to \infty$, n et $E[\theta]$ étant fixés.

- Lorsque $\boldsymbol{\lambda} \to \mathbf{0}$, le poids de l'a priori est nul dans la réponse Bayésienne et $E[\theta|x] \to \bar{x}$.

- Lorsque $\boldsymbol{\lambda} \to \infty$, c'est le poids des données d'observations qui est nul et $E[\theta|x] \to E[\theta]$.

Ci-dessous un résumé de ces différentes éventualités observées :

	$\boldsymbol{\lambda \to 0}$	$\boldsymbol{\lambda \to \infty}$
$\mathbf{Var(\theta)}$	$E[\theta](1 - E[\theta])$ Variance maximale	0 Variance minimale
Loi a priori	Loi de Haldane	Loi concentrée en $E[\theta]$
Interpretation	Situation non informative	Situation extrêmement informative
Estimation Bayésienne	$\bar{x}$	$E[\theta]$

Tableau 25 : observations des cas limites de la réponse bayésienne
Source : (DUPUIS, 2007)

De plus, plus la taille de l'échantillon est grande ($n \to +\infty$), $E[\theta]$ et λ étant fixé, plus le poids de l'a priori est négligeable et la réponse bayésienne ne dépends plus que des observations.

5.2.2.4 Etude de la sensibilité de la réponse bayésienne à la loi a priori

Le choix de l'a loi a priori est principalement basé sur des décisions subjectives et cela que ce soit en ce qui concerne la distribution de probabilité choisie (ici la loi bêta) ou encore en ce qui concerne les valeurs fournies par l'expert interrogé. Il sera donc utile pour cette étude, d'analyser l'impact de ces différentes décisions sur les résultats obtenus. Afin de réaliser cette étude de sensibilité, des perturbations seront faites sur l'estimation de θ fournie par l'expert. Pour cela, différentes valeurs de θ seront testées et les résultats obtenus seront analysés.

Ci-dessous les différents estimateurs de Bayes obtenus :

$\boldsymbol{\theta_{expert}}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\widehat{\theta_B}}$
5 %	7.5	135	8.9 %
15 %	22.64	127	14.74 %
20 %	20.51	120	17.50 %
22 %	**33**	**119**	**18.73 %**
25 %	37.67	113	20.69 %
30 %	45.56	105	23.87 %
50 %	76	76	35.86 %

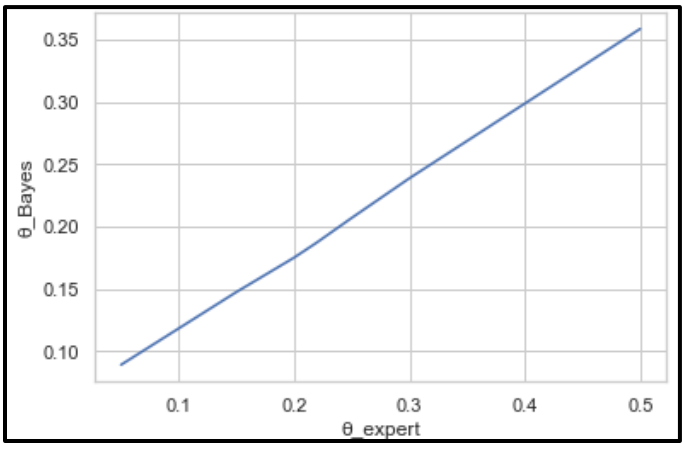


Tableau 26 : Analyse de sensibilité de la réponse bayésienne

Dans le tableau, α et β désignent les paramètres de la loi a priori correspondante et $\widehat{\theta_B}$ l'estimateur de Bayes obtenu.

De ces informations, il ressort que l'estimateur de Bayes évolue dans le même sens que l'estimation fournie par l'expert. Plus la valeur de l'estimation de l'expert est élevée, plus

l'estimation Bayésienne l'est : la réponse bayésienne est instable face à la variation de l'opinion de l'expert. Ainsi, l'information fournie par l'expert a ici un impact non négligeable dans la réponse bayésienne. De ce fait, si l'information a priori s'avère être erronée, cela pourrait fortement biaiser les résultats obtenus. Il faudrait ainsi faire attention à ne pas inclure dans le modèle, un a priori sur confiant.

Néanmoins, l'étude de sensibilité révèle également que les variations des estimations de bayes évoluent moins fortement par rapport aux variations de l'estimation de l'expert. Une variation de la valeur fournie par l'expert de 7% n'entraine qu'une évolution de l'estimateur de Bayes de 5 %.

Afin réduire l'impact de l'a priori dans la réponse bayésienne, une seconde approche basée sur une loi a priori non informative sera explorée dans la suite.

5.2.3 Approche 2 : Utilisation d'une loi a priori non informative

Cette approche se base sur l'utilisation d'une loi a priori non informative. Cette loi a priori est ensuite mise à jour avec les données d'observations de la base Veris dans un premier temps, ce qui permet d'obtenir une première distribution a posteriori. Dans un second temps, la loi a posteriori obtenue sera elle aussi mise à jour avec des informations fournie par un expert. Cela permettra ainsi d'obtenir une deuxième loi a posteriori.

5.2.3.1 1ère étape : utilisation d'une loi uniforme

Ne disposant pas d'informations précises sur le paramètre d'intérêt θ, une loi a priori non-informative sera utilisée. La loi non informative retenue ici est la loi uniforme sur [0,1]. Ce choix se justifie par le fait que lorsqu'il n'y a pas d'a priori, la loi uniforme apparait comme un choix naturel d'après le principe de la raison suffisante de Laplace. Ainsi, $\theta \sim U(0,1)$.

Il faut rappeler que d'après les données d'observations de la base Veris, 14 entreprises sont victimes de perte d'exploitation sur les 99 entreprises ayant subi des attaques cyber. Alors, la variable aléatoire $X \sim Bin(99, \theta)$, avec 14 sinistres *(« succès »)* observés.

D'après la formule de Bayes :

$$\boxed{\pi(\theta|X) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}}$$

$\int_\theta f(x|\theta)\pi(\theta)d\theta$ est d'après la section précédente, égale à une constante. Ce qui permet d'écrire :

$$\pi(\theta|X) \propto f(x|\theta)\pi(\theta)$$

$$\pi(\theta|X) \propto \theta^{14}(1-\theta)^{99-14}$$

$$\pi(\theta|X) \propto \theta^{14}(1-\theta)^{85}$$

Il s'agit là de l'expression d'une loi proportionnelle à a loi Beta (15,86)

Ainsi, $\boldsymbol{\theta|X}$ **~ Beta (15,86)** . Ce qui représente la **première distribution a posteriori** de ce modèle. L'estimateur de Bayes obtenu est donc :

$$\widehat{\theta_B} = E[\theta|X] = \frac{15}{15+86} = 0{,}15$$

Le résultat obtenu est assez proche de celui observée avec les données de la base Veris (14/99= 0.14). De plus, l'estimateur obtenu peut se réécrire comme suit :

$$\widehat{\theta_B} = \frac{14}{99}\,C + \frac{1}{2}(1-C)$$

$C = \frac{99}{101} = 0.98$ étant le facteur de crédibilité. Ce facteur montre qu'un poids important est accordé aux données d'observations de la base Veris. Cela peut s'expliquer par le fait que la moyenne de la loi uniforme a priori est non informative par rapport à la quantité de données disponible.

5.2.3.2 2ième étape : Incorporation de l'opinion d'experts

A cette étape, la loi a posteriori obtenue précédemment devient la loi a priori. Cette loi est mise à jour avec l'information apportée par l'expert. Pour cette méthode également on imaginera que les experts qui seront généralement des professionnels de la cyber sécurité disposent de connaissances statistiques (moyenne, variance…) sur le paramètre d'intérêt θ. L'expert peut soit être sûr à 100% de son estimation de la valeur de θ ou alors définir une distribution de probabilité pour θ qui prendra en compte son incertitude sur la valeur du paramètre estimé.

Afin de prendre en compte l'incertitude contenue dans l'estimation de l'expert, il serait intéressant de construire une distribution bêta à partir de la valeur de $\theta \in [0,1]$ fournie par ce dernier. Ainsi, on pourrait demander à l'expert, une estimation de la moyenne et la variance de la loi bêta obtenue, ce qui permettrait d'obtenir les paramètres $\alpha \; et \; \beta$ de cette loi grâce à la méthode des moments.

Cependant, il faut reconnaitre qu'il pourrait ne pas être aisé pour l'expert contacté de fournir avec précision la moyenne et la variance de la loi de bêta. Une autre alternative à cela serait de poser la question suivante à l'expert :

« Sur 100 entreprises ayant subi des attaques cyber, indiquez un intervalle de confiance à 95% pour le nombre d'entreprise parmi elles qui pourraient être victimes d'une interruption d'activité »

L'intervalle de confiance à 95% ainsi fourni par l'expert permettra d'obtenir les 2,5ème et 97,5ème quantiles de la loi bêta et ainsi de calculer les paramètres $\alpha\ et\ \beta$ de cette loi.

En supposant que $[\boldsymbol{a_o}, \boldsymbol{b_o}]$ est l'intervalle de confiance à 95% obtenu en divisant par 100 les bornes de l'intervalle fourni par l'expert. a_o, b_o sont compris dans l'intervalle [0,1] et sont des probabilités. Les paramètres $\alpha\ et\ \beta$ de la loi de Bêta peuvent être obtenus à l'aide de méthodes numériques. Ainsi, la fonction **Minimize** du package **Optimize** sur python sera utilisée pour minimiser numériquement l'expression de la somme des erreurs au carré ci-dessous par rapport à $\alpha\ et\ \beta$:

$$f(\alpha\ ,\beta) = \left(\int_0^{a_o} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{1-\beta}dx - 0.025\right)^2 + \left(\int_0^{b_o} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{1-\beta}dx - 0.975\right)^2$$

La fonction **Minimize** permet de trouver le minimum global de $f(\alpha\ ,\beta)$ ainsi que les valeurs de $\alpha\ et\ \beta$ pour lesquelles ce minimum est atteint, pour un intervalle spécifique $[a_o, b_o]$. Une fois la loi de θ trouvée, cette information sera incluse au modèle afin d'obtenir une seconde distribution a posteriori.

Elicitation de l'avis d'expert

De façon générale, l'information d'expert est obtenue en réalisant des sondages ou des interviews auprès de plusieurs experts. En raison du manque de temps et de la complexité du risque étudié, les intervalles de confiance fournis par les experts seront simulés aléatoirement. Ainsi, pour ce faire, deux nombres aléatoires seront générés : b_o sera généré à partir d'une distribution uniforme sur [0,1] et a_o à partir d'une distribution uniforme sur $[0, b_o]$. 100 simulations seront réalisées et les résultats seront examinés dans chaque cas. Cela correspondrait à un sondage auprès de **100 experts en cybersécurité et ainsi 100 intervalles de confiance à 95% sur la valeur de $\boldsymbol{\theta}$ seront probablement obtenus.**

A présent rappelons les résultats obtenus à la première étape :

$$\boldsymbol{\theta|X} \sim \textbf{Beta (15,86)}$$

Il s'agit là de la loi a priori de cette deuxième étape. Ainsi, après incorporation de l'avis d'expert, on a :

$$\pi(\theta|\mathrm{X}, \mathrm{Expert}) \propto f\ (Expert|\theta, \mathrm{X})\pi(\theta|\mathrm{X})$$

$$\pi(\theta|\mathrm{X}, \mathrm{Expert}) \propto \theta^{\alpha}\ (1-\ \theta)^{\beta}\ \theta^{14}(1-\ \theta)^{85}$$

$$\pi(\theta|\mathrm{X}, \mathrm{Expert}) \propto \theta^{\alpha+14}\ (1-\ \theta)^{\beta+85}$$

De façon analogue à l'approche utilisée à la première étape, on trouve :

$\boldsymbol{\theta|\mathrm{X}, \mathrm{Expert} \sim \mathrm{Beta}\ (\alpha+15, \beta+86)}$: ce qui représente la distribution a posteriori de cette 2ième étape et du modèle bayésien final.

Ainsi, l'estimateur de Bayes est donné par la formule suivante :

$$\widehat{\theta_{\mathrm{expert}}} = E(\theta|\mathrm{X}, \mathrm{Expert}) = \frac{\alpha+15}{\alpha+15+\ \beta+86} = \frac{\alpha+15}{\alpha+\ \beta+101}$$

Cet estimateur peut être réécrit comme suit :

$$\widehat{\theta_{expert}} = \frac{\alpha}{\alpha + \beta}\ C + \frac{15}{101}\ (1 - C)$$

Avec $C = \frac{\alpha+\beta}{\alpha+\beta+101}$ le facteur de crédibilité, qui montre que, plus les valeurs de $\alpha\ et\ \beta$ sont élevées, plus le poids accordé à l'information apporté par l'expert est important.

Simulations d'intervalles de confiance

Etant donné que la fonction $f(\alpha, \beta)$ dépends également des bornes $a_o\ et\ b_o$ de l'intervalle de confiance fourni par les différents experts simulés, l'expression de la fonction change d'un expert à l'autre. Ainsi, pour certaines fonctions f obtenues, la fonction **Minimize** utilisée n'est pas en mesure d'atteindre le minimum. Dans ces situations, il n'est alors pas possible de d'estimer les paramètres $\alpha\ et\ \beta$ de la loi a priori. Néanmoins, cela ne constitue pas une véritable limite pour l'estimation du paramètre d'intérêt car cette situation est similaire à celle qui aurait été observée dans la réalité. En effet, si 100 experts avaient réellement été contactés, il est fort probable que certains parmi eux ne répondent pas au sondage, ce qui conduirait également à un cas de figure où les paramètres de la loi a priori ne peuvent être connus.

Ainsi, dans le cadre des simulations effectuées ici, la fonction **Minimize** a permis d'obtenir le minimum de 62 fonctions $f(\alpha, \beta)$, ce qui corresponds à 62 réponses d'experts.

Ci-dessous les résultats des 5 premières simulations effectuées :

a_o	b_o	α	β	$\widehat{\theta_{expert}}$
0.21	0.95	3.08	1.81	0.17
0.25	0.59	12.96	17.93	0.21
0.0022	0.76	0.70	2.21	0.15
0.28	0.92	4.39	2.47	0.18
0.10	0.85	2.37	2.78	0.16

Tableau 27 : Résultats des simultations effectuées

Les deux premières colonnes du tableau ci-dessus répertorient les bornes des intervalles de confiances fournies par les experts interrogés. Les colonnes 3 et 4 quant à elles, rassemblent les paramètres $\alpha\ et\ \beta$ de la loi beta du paramètre θ. Ces paramètres ont été obtenus avec la minimisation de la fonction $f(\alpha, \beta)$ explicitée plus haut. Enfin, la dernière colonne donne l'estimateur de Bayes du paramètre θ obtenue.

De ce tableau, il ressort que les intervalles de confiance fournis par les experts sont d'amplitudes variables et différents les uns des autres. Cela est assez crédible avec la réalité car étant donné la complexité du risque cyber, il serait difficile de trouver des experts dont les

avis convergent. Il faut également remarquer que plus l'intervalle $[a_o, b_o]$ est restreint, plus les valeurs de $\alpha \; et \; \beta$ sont élevées. Cela s'explique par le fait que la densité de probabilité de la variable θ fournie par l'expert, est plus pointue autour d'une certaine valeur de θ lorsque la différence entre $a_o \; et \; b_o$ est petite et dans ce cas, il est caractéristique pour la distribution Bêta d'avoir des valeurs de paramètres plus grandes pour α et β, signifiant ainsi qu'un poids important est accordé à l'information fournie par l'expert.

Une moyenne sur les 62 valeurs de $\widehat{\theta_{\text{expert}}}$ obtenues sera utilisée comme estimation pour la fréquence de sinistre finale. Ainsi la fréquence de sinistre estimée obtenue avec cette deuxième méthode est de **18.28 %**. Ce résultat est assez proche de celui obtenu avec la méthode 1 (18,73%). Afin de rester prudent, il serait judicieux de choisir la fréquence la plus élevée des deux.

Ainsi, la fréquence de sinistre pour une garantie de perte d'exploitation est d'environ **18.73%**. La fréquence de sinistre estimée semble assez cohérente avec la réalité car elle se rapproche fortement de celle obtenue dans le rapport publié par Opinionway (**22%**).

5.2.4 Limites des modèles utilisés

Bien que les approches bayésiennes soient adaptées pour résoudre de nombreuses problématiques, elles restent tout de même critiquables. En effet, comme mentionné plus haut, les probabilités a priori sont choisies de manière subjective. L'information préalable prise en compte dans les modèles varie alors d'un analyste à l'autre. Ainsi, il est complexe d'obtenir des résultats convergents pour la modélisation d'un même risque, ce qui constitue un des principaux inconvénients liés à l'utilisation de la statistique bayésienne.

De plus, dans le cas où très peu de données d'observations sont disponibles, un poids important est accordé aux informations issues de l'avis d'expert et l'a priori considéré est sur confiant. Cela pourrait fortement biaiser les résultats de la réponse bayésienne dans le cas où l'expertise ne serait pas fiable ou erronée.

Ainsi, face à cette contrainte, l'idéal aurait été de comparer les résultats extraits du rapport de Opinionway avec ceux d'autres études ou avec des estimations faites par experts en cybersécurité. Malheureusement très peu d'études évaluant la probabilité d'une interruption d'activité causée par une cyber attaque sont disponibles.

Néanmoins, une étude de sensibilité pourrait permettre d'analyser l'impact de ces décisions subjectives sur les résultats des estimations et de quantifier le biais que pourrait engendrer un avis d'expert incorrect.

Par ailleurs, dans le cas de la problématique étudiée dans cette section, l'utilisation de la base Veris pourrait fausser l'estimation obtenue. En effet, comme il l'a été constaté lors de la construction des modèles, la base met à disposition très peu d'informations sur les caractéristiques des incidents répertoriés. A titre d'exemple, pour 96% des incidents de la base il est impossible de savoir si l'entreprise attaquée a été victime ou non d'une perte d'activité.

Aussi, la base étant remplie à la main, elle pourrait contenir des informations erronées. Certains approvisionneurs de cette dernière pourraient par exemple ne pas bien comprendre son mode de remplissage.

Toutefois, les résultats obtenus en utilisant la loi a priori informative et ceux fournis par la loi a priori étant très proches, on pourrait penser que le biais contenu dans les données a priori est faible.

CONCLUSION

L'objectif principal de ce mémoire était de présenter une méthode de construction d'une offre d'assurance cyber à destination des PME/TPE. Pour y parvenir, une étude approfondie du risque cyber, de son ampleur et de ses spécificités a été réalisée en amont. De cette étude, il a été déduit que le risque numérique est évolutif, systémique et protéiforme, rendant ainsi complexe sa modélisation pour les assureurs.

L'étude de risque réalisée a également permis de définir les conséquences des attaques informatiques sur les PME/TPE afin d'inclure dans l'offre d'assurance des garanties qui leurs permettraient d'y faire face efficacement. Ainsi, les garanties d'assurance qui ont été retenues sont : *atteinte aux données, perte d'exploitation, cyber vol, responsabilité civile et cyber extorsion.*

Par ailleurs, l'étude a révélé une absence de base de données publiques complètes et fiable en ce qui concerne le risque cyber. Dans le mémoire, une méthode de modélisation de ce risque malgré l'existence de telles contraintes a été présentée à travers la tarification de la garantie *« Atteinte aux données »* incluse dans l'offre d'assurance construite.

La tarification de cette garantie a été effectuée à l'aide de l'application de la méthode de fréquence x coût sur une base de données publique américaine (PRC). Dans l'élaboration du processus de tarification, les limites de cette base (incomplétude, manque de données sur les caractéristiques des victimes…) ont été partiellement surmontées grâce à l'utilisation de méthodes mathématiques adaptées à des données incomplètes telles que les lois 0-tronquées. Cela a permis d'aboutir à des montants de primes pures par secteur d'activités qui ne semblent pas aberrants par rapport à ceux des offres disponibles sur le marché français.

Les modèles de tarification construits pourraient également être élargis à une utilisation sur des bases de données privées et servir de bons indicateurs pour la quantification du risque cyber.

Afin d'introduire la mise en place d'un processus de tarification de la garantie *« perte d'exploitation »*, une étude de la fréquence d'interruption d'activité survenant à la suite d'une attaque cyber a été réalisée. Cette étude a permis de remédier en partie, au manque de sonnées existant à travers l'utilisation de méthodes d'inférences bayésiennes. Ces dernières ont permis d'obtenir des résultats cohérents avec ceux d'études réalisées sur le risque de perte d'exploitation.

Cependant, quelques limites ont été relevées sur la pertinence des résultats obtenus ainsi que le choix des hypothèses utilisées. Il s'agit notamment de l'absence de la prise en compte du coût marginal d'une donnée dans l'évaluation du coût moyen des incidents cyber dans le cas de la garantie *« atteinte aux données »*. De même, les méthodes bayésiennes utilisées sont basées sur des décisions subjectives qui pourraient contraindre la fiabilité des estimations obtenues.

Toutefois, l'évolution du portefeuille d'assurance permettra de disposer d'un historique de sinistres et ainsi d'améliorer les modèles de tarifications construits afin d'obtenir des tarifs plus adéquats pour chaque catégorie d'assuré.

En somme, les différentes problématiques abordées dans ce mémoire ont permis de mettre lumière les difficultés auxquelles pourraient faire face les actuaires telles que le manque de données et la modélisation de risques complexes. Ce fut également l'occasion de comprendre les bases de la construction d'un produit d'assurance ainsi que d'analyser les différents obstacles inhérents à l'étude des nouveaux risques issus de la digitalisation de la société.

La réalisation de ce mémoire marque le début d'un projet à long terme au sein de Moonshot Insurance. A présent, il faudrait compléter les travaux effectués pour une étude approfondie de la rentabilité du produit d'assurance proposé afin d'évaluer les risques encourus par l'assureur dès son lancement. Il faudrait également s'intéresser aux modes de distributions de l'offre afin de décider si cette dernière sera incluse dans un contrat d'assurance *multirisques professionnels* ou si elle fera l'objet d'un contrat à part entière.

BIBLIOGRAPHIE

(2021, 11 12). *l'internaute*.

ACPR. (2019). *La distribution des garanties contre les risques cyber par les assureurs.*

American Medical Association. (s.d.). *HIPAA Breach Notification Rule*. Récupéré sur https://www.ama-assn.org/practice-management/hipaa/hipaa-breach-notification-rule#:~:text=HIPAA's%20Breach%20Notification%20Rule%20requires,and%20security%20of%20the%20PHI.

AMRAE. (2021). *LUCY.*

AMRAE. (2022). *Lumière sur la cyber assurance.*

ANSSI. (2020). *'ANSSI ET LE BSI ALERTENT SUR LE NIVEAU DE LA MENACE CYBER EN FRANCE ET EN ALLEMAGNE DANS LE CONTEXTE DE LA CRISE SANITAIRE.*

Assemblée Nationale. (2022). Projet de loi d'orientation et de programmation du ministère de l'intérieur.

Berliner, B. (1982). *Limits of insurability of risks.*

CEPro. (2021, Février 24).

Christian, B., Martin, E., & Jan, H. ,. (2015). *INSURABILITY OF CYBER RISK : AN EMPIRICAL ANALYSIS.*

Cision PRWeb. (2020, Février 19). New Study Reveals One In Three SMBs Use Free Consumer Cybersecurity And One In Five Use No Endpoint Security At All. San Francisco.

CNRS. (s.d.). *centre National de Ressources Textuelles et Lexicales.* Récupéré sur https://www.cnrtl.fr/definition/expert

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective.*

CPME. (2019). 16 chiffres clés sur la Cybersécurité des entreprises (-50 salariés).

CRO Forum. (2014). *The cyber risk challenge and the role of insurance.*

Cyber Data-Risk. (2020). Récupéré sur https://databreachinsurancequote.com/cyber-insurance/cyber-insurance-data-breach-insurance-premiums/

Cybercover. (s.d.). CYBERATTAQUES & PME : QUELLES CONSÉQUENCES FINANCIÈRES ?

CyberCover. (s.d.). CYBERATTAQUES & PME : QUELLES CONSÉQUENCES FINANCIÈRES ?

Cybermalveillance. (2021). *Rapport d'activité 2021*.

Cybermalveillance.gouv. (2019, Octobre 9). *Attaque DDoS, que faire ?* Récupéré sur Cybermalveillance.gouv: https://www.cybermalveillance.gouv.fr/tous-nos-contenus/fiches-reflexes/attaque-en-deni-de-service-ddos

Cybermalveillance.gouv. (2020, Janvier 10). *Que faire en cas de phishing ou hameçonnage ?* Récupéré sur cybermalveillance.gouv: https://www.cybermalveillance.gouv.fr/tous-nos-contenus/fiches-reflexes/hameconnage-phishing

Cybermalveillance.gouv. (2022, Mars 2). *Qu'est-ce qu'un ransomware ou rançongiciel ?* Récupéré sur Cybermalveillance.gouv: https://www.cybermalveillance.gouv.fr/tous-nos-contenus/actualites/ransomware-rancongiciel-definition

DUPUIS, J. (2007). *Statistique bay´esienne et algorithmes MCMC.* Récupéré sur https://www.math.univ-toulouse.fr/~dupuis/bayesmim08.pdf

Faure-Muntian, V. (2021). *Rapport La cyber Assurance.*

Fireeye. (s.d.).

France Assureurs. (2022). *Cartographie prospective 2022 des risques de la profession de l'assurance et de la réassurance.*

Galea. (2021). Le risque cyber .

Hiscox. (s.d.). Récupéré sur Site web de l'entreprises: https://www.hiscox.fr/pourquoi-hiscox/a-propos-d-hiscox

INSEE. (2019, Novembre 04). Équipements et usages dans les établissements scolaires. France.

INSEE. (2019). *L'économie et la société à l'ère du numérique.*

JDN. (s.d.). Récupéré sur http://www.journaldunet.com/solutions/0501/050126_gestion_correctifs.shtml#:~:text=La%20gestion%20des%20correctifs%20ou,%C3%A0%20jours%20de%20s%C3%A9curit%C3%A9%20logicielles.

Journal officiel de l'union européenne. (2016). DIRECTIVE (UE) 2016/97 DU PARLEMENT EUROPÉEN ET DU CONSEIL. *Journal officiel de l'union européenne*.

LAROUSSE. (2021). *Dictionnaire.*

Le club des juristes . (2018). *Assurer le risque cyber* . Paris.

LESAFFRE, C. (2021). Cybersécurité : les PME n'investissent pas assez pour se protéger. *Europe 1*.

Martin, E., & Nicola, L. (2017). *Data breaches: Goodness of fit, pricing, and risk measurement.*

McKinsey & Company. (2020). *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever.* New York: New York office.

MEURANT, S., & CARDON, R. (2021). *Rapport d'information.*

Netwrix. (2022, Janvier 13). Les 10 types de cyberattaques les plus courants.

Opinionway pour CESIN. (2022). *Baromètre de la cyber-sécurité des entreprises.* Paris.

Ponemon Institute . (2022). *Cost of Data Breach.*

Ponemon Institute. (2019). *Cost of Databreaches.*

Ponemon Institute. (2021). *Cost of Databreaches.*

Saragaglia, B. (2017). Les 10 plus gros braquages de l'histoire du Web . *CAPITAL*.

Sébastien Farkas, Olivier Lopez, & Maud Thomas. (2020). *Cyber claim analysis through Generalized Pareto.*

Shepherd, M. (2020, Decembre 16). 30 Surprising Small Business Cyber Security Statistics . *Fundera*.

Stoïk. (s.d.). Récupéré sur Site de l'entreprise: https://www.stoik.io/a-propos

Stoïk. (2022). *Stoïk.* Récupéré sur https://www.stoik.io/assurance-cyber/prix

Verizon. (2021). *Data Breach Investigation report.*

Walker-Roberts, S., Hammoudeh, M., Aldabbas, O., & al. (2020). Threats on the horizon: understanding security threats in the era of cyber-physical systems. *The Journal of Supercomputing 76*, 2643–2664. Récupéré sur https://doi.org/10.1007/s11227-019-03028-9

XERFI, l'économie réelle. (2021, Mai 10). Cyber assurance : un marché en progression de 20% par an d'ici 2023.

ANNEXE A : BENCHMARK

		Stoïk	Coalition	Cyber Sérénité	Zeguro	Hiscox	Allianz
Informations	Statut	Insurtech	Insurtech	Courtier	Courtier	Assureur	Assureur
	Date de création/lancement assurance	2021	2017	NA	2016	2019	2020
	Nationalité	France	USA	France	USA	UK	Allemagne
	Taille (effectif)	15	320		11	2500+	5000+
	Géographie adressée	France	USA	France	USA	Monde	Monde
Scope produit	PME/TPE	Oui	Oui	Oui	Oui	Oui	Oui
	Grandes entreprises	Non	Oui	Oui	Non	Non	Non
Services de prévention	Scan hebdomadaire du réseau (Monitoring)	Oui	Oui	Non	Oui	Non	Non
	Mise à disposition d'experts en cyber sécurité	Oui	Oui	Oui	Oui	Oui	Oui
	Protection du réseau	Oui	Oui	Oui	Oui	Non	Oui
	Formation à la cyber sécurité	Non	Oui	Non	Oui	Oui	Oui
	Diagnostic en ligne	Non	Non	Oui (en option)	Oui	Non	Non
Garanties	Atteinte aux données (prise en charge des différents frais)	Oui	Oui	Oui	Oui	Oui	Oui
	Perte d'exploitation (continuité d'activité)	Oui	Oui	Oui	Oui	Oui	Oui
	Couverture des amendes (Sanctions pécuniaires)	Non	Non	Non	Oui	Oui	Oui
	Cyber extorsion	Oui	Oui	Oui (en option)	Oui	Oui	Oui
	Cyber Fraude	Oui (en option)	Oui	Non	Oui	Oui	Non (cyber Vol)
	Gestion de crise (Prise en charge des frais d'assistance technique)	Oui	Oui	Oui	Oui	Oui	Oui
	Rc "transmission de virus"	Oui (en option)	Non	Oui	Oui	Oui	Oui
	usurpation d'identité de l'entreprise	Non	Non	Non	Oui	Non	Non
	e reputation (du dirigeant?)	Non	Oui	Non	Non	Oui	Non
	Prise en charge des frais liés aux démarches RGPD (notification, etc)	Oui	Oui	Oui	Oui	Oui	Non
	Prise en charge des frais de communication	Oui	Oui	Oui	Oui	Oui	Oui
	Prise en charge des frais de poursuite judiciaires (frais de défense etc, inclus dans la RC)	Oui (en option)	Oui	Oui	Oui	Oui	Oui

ANNEXE B : SEVERITE DE LA GARANTIE ATTEINTE AUX DONNEES : SECTEUR BSR

Cette annexe reporte les éléments relatifs à la modélisation de la sévérité de la garantie atteinte aux données dans le secteur BSR.

Sévérité en nombre de données

Sources / Paramètres	Dataloss DB	Media	Databreaches .net	Security Breach Letter
Kurtosis	-1.08	-1.54	- 0.90	5.94
Asymétrie	0.40	- 0.19	0.59	1.97
Moyenne	8.12	9.74	7.68	6.50

Figure 47 : Paramètres des tailles de violations secteurs BSR

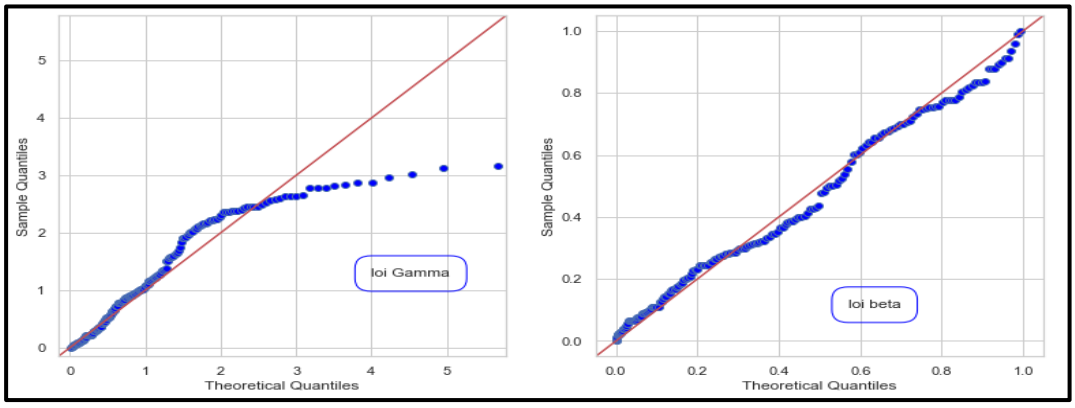


Figure 48 : Q-Q plot lois bêta et gamma sur logarithme des tailles de violations secteur BSR

Evaluation du coût moyen:

Variables	Valeurs
unit_BSR_2019	119 $ USD
coût_BSR_2019	1.84 M $ USD
coût_BSR_2022	3.28 M $ USD
Unit_BSR_2022	212.13 $ USD
Cout_moy_BSR	2.010 M $ USD

Tableau 28: Coûts moyens relatifs au secteurs BSR

ANNEXE C : DEMONSTRATION MATHEMATIQUE

Dans la section 5.2 du mémoire, il avait été admis le résultat ci-dessous :

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta \ = \frac{B(a,b)}{B(\alpha,\beta)}$$

Avec :

$\theta \sim Be\ (\alpha, \beta)$

$a = \alpha + s$ et $b = \beta + n - s$;

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}\ 1_{[0,1]}(\theta)$$

$$f(x|\theta) = \prod_{i=1}^{n} \mathbb{P}(X = x_i|\theta) = \ \theta^{s}(1-\theta)^{n-s}$$

n étant le nombre d'entreprises de la base Veris et $s = \sum_{i=1}^{n} x_i$.

Démonstration

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta \ = \int_{\theta} \theta^{s}(1-\theta)^{n-s}\frac{1}{B(\alpha,\beta)}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}\ 1_{[0,1]}(\theta)d\theta$$

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta \ = \frac{1}{B(\alpha,\beta)}\int_{\theta} \theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}d\theta$$

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta \ = \frac{B(s+\alpha-1, n-s+\beta-1)}{B(\alpha,\beta)}\int_{\theta} \frac{\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}1_{[0,1]}(\theta)}{B(s+\alpha-1, n-s+\beta-1)}d\theta$$

En posant $a = s + \alpha - 1$ et $b = n - s + \beta - 1$, on a :

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta \ = \frac{B(a,b)}{B(\alpha,\beta)}\int_{\theta} \frac{\theta^{a-1}(1-\theta)^{b-1}1_{[0,1]}(\theta)}{B(a,b)}d\theta$$

$\int_{\theta} \frac{\theta^{a-1}(1-\theta)^{b-1}1_{[0,1]}(\theta)}{B(a,b)} d\theta$ est l'intégrale de la densité d'une loi de bêta de paramètre a,b. Alors, $\int_{\theta} \frac{\theta^{a-1}(1-\theta)^{b-1}1_{[0,1]}(\theta)}{B(a,b)} d\theta = 1$

Par conséquent,

$$\int_{\theta} f(x|\theta)\pi(\theta)d\theta = \frac{B(a,b)}{B(\alpha,\beta)}$$

Cqfd.