IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Mangement
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le 8 janvier 2021

Par : Thomas Bastard

Titre : Modélisation du risque Cyber de perte de Données à Caractère Personnel, modèle de tarification, inclusion dans le BGS et proposition de scénarios de stress pour l'ORSA.

Confidentialité : ☐ NON OUI ☒ (Durée : ☐ 1 an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires

.....

.....

.....

Membres présents du jury de l'Institut du Risk Management

.....

.....

.....

.....

Secrétariat :

Bibliothèque :

Entreprise :
Nom : ADDACTIS FRANCE

Cachet :

Directeur de mémoire en entreprise :
Nom : Benjamin Poudret

Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Remerciements

Je tiens à remercier les personnes m'ayant soutenu tout au long de cette longue aventure de recherche, de réflexion et de rédaction. En premier lieu mes collègues pour leurs relectures attentives, leurs encouragements et leurs idées : Laurent Devineau, Emmanuelle Huguet, Romain Nobis, Auriol Wabo, Thomas Lallement et plus particulièrement Benjamin Poudret qui a encadré ce mémoire.

Je tiens également à remercier mes camarades du CEA ainsi que les autres professionnels avec qui j'ai pu échanger et mes proches : Julie, Anne-Céline, Aude, Sandrine, Cécile, Maxime, Hichem, Guillaume, Yannick, Karine.

Résumé

Une Cyber-attaque est une atteinte à des systèmes informatiques réalisée dans un but malveillant et ciblant différents dispositifs informatiques, comme par exemple des ordinateurs, des imprimantes ou encore des objets connectés : drones, serrures, *pacemakers*...

L'objet de ce mémoire est l'étude du risque Cyber de perte de Données à Caractère Personnel (DCP) devenu plus important depuis la mise en application dans l'Union Européenne du Règlement Général sur la Protection des Données entré en vigueur le 25 mai 2018, la définition de bonnes pratiques d'atténuation du risque et l'étude de la possibilité de transfert de risque auprès de l'entreprise à un assureur.

Des rappels et des exemples concrets d'attaques sont présentés pour permettre au lecteur de s'approprier les enjeux de ce risque émergent, puis deux bases de données publiques recensant essentiellement des incidents ayant eu lieu aux Etats-Unis sont étudiées en détails.

Deux sous-modèles stochastiques séquentiels construits et calibrés avec ces deux bases et prenant en compte les caractéristiques propres à une entreprise sont proposés. Tant les choix de modélisation que les résultats obtenus sont mis en perspective avec plusieurs études externes.

Quatre applications sont proposées : une modélisation brute pouvant être utilisée pour le Pilier 1 de Solvabilité 2, un modèle de tarification de couvertures non-proportionnelles, l'évaluation du BGS (Besoin Global de Solvabilité) de l'ORSA avec une proposition d'ajustement de la Formule Standard pour prendre en compte le risque Cyber et enfin la construction de scénarios dans une approche ERM pour le maintien des exigences réglementaires de l'ORSA.

Mots-clés : *Cyber, DCP, RGPD, ORSA, Jacobs, ERM, BGS, Scénario, Solvabilité 2, PRC, VERIS, loi tronquée, Weibull, Ponemon, CESIN, Tarification, Couverture, Traité, Réassurance, Non-proportionnel, SCR, Opérationnel, Formule Standard*

Abstract

A Cyber attack is an assault using one or more computers against a single or multiple computers, networks or devices. A Cyber attack can maliciously disable computers, steal data or use a breached computer as a launch point for other attacks. The emergence of the Internet of Things (IoT) devices raises serious concerns in the areas of privacy and security. Global industry and governmental moves to address these concerns have begun, starting in the USA then rapidly followed by other developed regions in the world.

This actuarial thesis investigates a major Cyber risk : data confidentiality breach incidents through the loss of personal records. This risk has become more serious since the application of the General Data Protection Regulation (GDPR) since May 28^{th} 2018. This thesis also tackles the issue of Cyber risk mitigation and Cyber risk transfer through insurance solutions.

First, several historical Cyber incidents are presented to support the understanding of the reader of both Cyber risk and its related issues. Two open databases that describe historical Cyber incidents which mostly occured in the USA are then studied.

Two stochastic sequential sub-models are built and calibrated based on the data to provide a global modeling which depends on the core characteristics of the company at stake. Special attention is paid to the coherence of the modeling and the results through the study and comparison with several external studies.

Four insurance applications are eventually presented : a raw modeling which suits the Solvency 2, Pillar 1 requirements, a non-proportionnal cover pricing model, the evaluation of the Overall Solvency Need (OSN) of the ORSA through a Standard Formula adjustement to include Cyber risk. The last application is the proposal of several Cyber stress scenarios to include in the ORSA.

Key words : *Cyber, Privacy breach, Personal records, GDPR, ORSA, Jacobs, ERM, OSN, Scenario, Solvency 2, PRC, VERIS, Truncated distribution, Weibull, Ponemon, CESIN, Pricing, Cover, Treaty, Reinsurance, Non-proportional, SCR, Operational, Standard Formula*

Sommaire

Introduction

Telle que définie sur le site du gouvernement français, une Cyber-attaque est une atteinte à des systèmes informatiques réalisée dans un but malveillant ciblant différents dispositifs informatiques : des ordinateurs, des serveurs, isolés ou en réseaux, reliés ou non à Internet, des équipements périphériques tels que les imprimantes, ou encore des appareils communicants comme les téléphones mobiles, les smartphones ou les tablettes. Les entreprises sont principalement exposées à quatre types de Cyber-attaques : l'exfiltration de données (en particulier les Données à Caractère Personnel), la perte de données, l'attaque par déni de service et la Cyber-extorsion.

Les risques Cyber ont émergé à la suite de plusieurs phénomènes : une tertiarisation des sociétés occidentales depuis le début du XXiè siècle, une numérisation en marche forcée concernant les individus et les entreprises et l'arrivée massive de matériel informatique dans les entreprises depuis le début des années 2000, suivis par de nouvelles dépendances, de nouveaux usages et de nouveaux enjeux liés aux données depuis le début des années 2010 avec la montée en puissance des architectures réseaux multi-sites et externalisées : serveurs distants loués à la demande, modèles économiques basés sur des requêtes ou de la bande passante, hébergement à distance intercontinentaux avec les offres d'informatique en nuage proposés par les géants américains du net tel que le Saas (*Software as a Service*) devenant une nouvelle norme de modèle économique.

Les moyennes et grandes entreprises sont généralement sensibilisées à la Cyber-sécurité, c'est-à-dire la mise en œuvre de moyens de protection contre le risque Cyber : une approche *a priori* avec l'atténuation de la Cyber-vulnérabilité et une approche *a posteriori* avec des mesures de gestion de crise et de résilience adaptées.
Cependant, la mise en œuvre de ces moyens est coûteuse et chaque entreprise est contrainte de conserver un risque résiduel après sa politique d'atténuation du risque.

L'objet de ce mémoire est l'étude des risques Cyber auxquels sont exposées les entreprises, la définition de bonnes pratiques d'atténuation du risque Cyber ainsi que l'étude de la capacité de transfert de risque de l'entreprise à un assureur.

Cette étude a une ambition double : sensibiliser les assureurs aux nouveaux enjeux du risque Cyber en leur apportant des éléments de réponse concrets pour l'évaluation de leur risque dans leurs processus ERM (ORSA), ainsi que fournir le coeur d'un modèle de tarification permettant la modélisation de ce risque et de nombreuses applications dont la tarification pour l'assurance et la réassurance, l'intégration dans des scénarios pour une approche Catastrophe pour les exigences réglementaires du Pilier 1 de Solvabilité 2 et le suivi de l'appétence au risque.

Après quelques rappels essentiels dans un premier chapitre sur le risque Cyber, la directive Solvabilité 2 et l'ORSA, le deuxième chapitre de ce mémoire portera sur l'étude fine de deux bases de données historiques listant des incidents Cyber de type perte de Données à Caractère Personnel (DCP).

L'étude comparée de deux bases de données est un choix innovant. En particulier, la deuxième base a été très peu étudiée dans la littérature et est de qualité supérieure à la première, largement étudiée. De plus, elle possède une variable clé que la première ne contient pas : le coût financier d'un incident ! C'est cette deuxième base qui offre des perspectives nouvelles avec la construction d'un modèle de sévérité stochastique.

Puis le troisième chapitre portera sur la modélisation du risque : la recherche d'une structure adaptée et la calibration de chaque sous-modèle retenu en mettant en perspective chaque choix retenu avec les bases historiques ainsi qu'avec d'autres études externes.

Enfin, quatre applications seront proposées dans un quatrième et dernier chapitre. Tout d'abord une quantification du risque Cyber de perte de Données à Caractère Personnel au quantile à 99,5 % en ligne avec la réglementation Solvabilité 2. Ensuite, la deuxième application portera sur la tarification de couvertures non-proportionnelles avec application sur des cas fictifs. Puis la troisième partie proposera de modifier la Formule Standard pour mieux prendre en compte le risque Cyber de manière analogue à l'évaluation du BGS (Besoin Global de Solvabilité) de l'ORSA. Enfin, des scénarios utilisables pour le respect du maintien des exigences réglementaires de l'ORSA seront définis et quantifiés.

Chapitre 1

Présentation du risque Cyber et rappels essentiels

1.1 Le risque Cyber

Cette section présente les différents types d'attaques Cyber au travers de quelques études de cas.

1.1.1 Quelques études de cas

1.1.1.1 Le ver Morris : premier virus de l'histoire (1988)

En 1988, Robert Morris, étudiant de 23 ans à l'université Cornell (État de New-York, États-Unis), lance le premier ver informatique sur le réseau internet. Seules les machines sous Unix étaient vulnérables. Morris n'a pas d'intention malveillante en tant que tel, son programme d'à peine une centaine de lignes de code n'a pour seul objectif que de pouvoir se propager sur le réseau. Jusqu'à 6 000 machines auraient été infectées, soit 10% du réseau Unix.

Malgré l'absence d'intentions réellement malveillantes, le ver Morris a causé d'importants dommages aux machines infectées : le ver étant capable d'infecter, par erreur, plusieurs fois chaque machine, et chaque infection ralentissant la machine. Un nombre conséquent d'entre elles ont été tellement ralenties au point de devenir inutilisables.

Cette erreur a transformé un exercice intellectuel en une puissante attaque de type déni de service.

Robert Morris est professeur au MIT (Massachusetts Institute of Technology) en sécurité informatique depuis 2006.

1.1.1.2 2. TJ Maxx : Vol de données de cartes bancaires (2005 à 2007)

En 2007, TJ Maxx, une chaîne de grands magasins opérant principalement aux États-Unis révèle une fuite de données de ses serveurs ayant débuté en 2005 et concernant des transactions remontant à 2003, en particulier des coordonnées de cartes de débit et de crédit liées à des transactions bancaires passées. Les hackers ont opéré de juin 2005 à janvier 2007. Les hackers ont pu voler via une simple connexion Wifi des données de TJ Maxx, mais également de toutes les autres filiales du groupe auquel l'entreprise appartenait : Marshalls, Winners, HomeSense, HomeGoods,... et ce dans plusieurs pays différents (États-Unis, Royaume-Uni, Canada, Irlande). Plus de 100 millions de clients ont été exposés à l'attaque et 46 millions ont été effectivement victimes. TJ Maxx a dû payer des frais administratifs et des remboursements liés à la fraude très conséquents (environ 800 millions £) et s'est fait attaquer en justice lors d'une *class action* menée par une association de banques victimes, ayant dû réémettre

un très grand nombre de cartes bancaires à la suite de l'attaque.

Il est possible de retenir de cette attaque que TJ Maxx conservait des données de transactions bancaires sur de très longues périodes, que les hackers ont pu voler des données pendant une durée de plus de 18 mois et qu'une intrusion via TJ Maxx permettait de récupérer des données issues d'autres filiales du groupe et des transactions issues d'autres régions du monde.

L'auteur principal de la fraude, Alberto Gonzales, a été condamné aux États-Unis à 20 ans de réclusion criminelle en 2010.

1.1.1.3 Saint-Gobain (2017)

En juin 2017, le groupe multiséculaire Saint-Gobain est victime du rançongiciel NotPetya, ce dernier s'étant infiltré dans son réseau interne via sa filiale ukrainienne. Le nom NotPetya provient d'un autre rançongiciel, Petya, datant de 2016 et dont NotPetya est une variante. L'infiltration provient d'un logiciel du site de l'administration fiscale ukrainienne ayant infecté la filiale de Saint-Gobain. Plusieurs systèmes d'information ont été bloqués à la suite d'un chiffrement par le virus, immédiatement suivi par la suppression massive de données, en à peine quelques minutes. De fait, les réseaux de distribution du groupe ont dû revenir temporairement au crayon et au papier pour la gestion des commandes. L'activité n'a été perturbée qu'une dizaine de jours mais pourtant, les pertes pour Saint-Gobain ont été considérables et estimées à plus de 220 millions d'euros de chiffres d'affaires, soit 0,5% du chiffre d'affaires annuel.

Les dirigeants du groupe ont présenté à la suite de cette attaque un plan de Cyberdéfense en quatre points : identification des risques, détection des failles, réaction et résilience ; un plan de continuité d'activité en mode dégradé et ont fait passer le budget Cyberdéfense à environ 15% du budget IT du groupe. De plus, Saint-Gobain a assuré une partie de son risque Cyber à la suite de l'attaque.

Les actions de Saint-Gobain, bien que tardives, illustrent parfaitement les bonnes pratiques de Cyber-sécurité.

Ces trois exemples illustrent les trois types de Cyber-attaques les plus impactantes : le déni de service (ver Morris), la perte de données (TJ Maxx) et la Cyber-extorsion (Saint Gobain). De plus, la chronologie de ces attaques permet de se rendre compte de la prise de conscience du risque Cyber et des politiques de sécurité informatique au cours des décennies : inexistantes dans les années 80, balbutiantes dans les années 2000 puis au coeur de la stratégie de l'entreprise illustré par Saint Gobain et sa politique de sécurité mise en place *a posteriori* de l'attaque.

1.1.2 Les types de Cyber-attaques

1.1.2.1 Exfiltration de données

Une attaque de type exfiltration de données a pour objectif de récupérer des données confidentielles, le plus souvent cela concerne des données personnelles conservées par des entreprises mais cela concerne également des données professionnelles confidentielles. Les catégories de données pouvant être volées sont :

— **Données d'Identité Personnelles**, telles que le nom complet, les détails de contact (adresse physique, adresse courriel, date de naissance, numéro de passeport ou de permis de conduite, numéro de sécurité sociale).

— **Données de carte bancaire**, en complément des données d'identité personnelles, telle que le numéro ou l'empreinte de carte bancaire de débit ou de crédit, code PIN, RIB, code d'accès aux services bancaires.

— **Données médicales**, en complément des données d'identité personnelles, des données concernant une pathologie, une posologie, des rendez-vous pris avec des experts médicaux ou des données relatives à des opérations de santé, des identifiants biométriques (empreintes digitales, d'iris), des résultats de tests médicaux.

— **Données professionnelles confidentielles** : informations sensibles liées à des propriétés d'entreprises, de secrets industriels, d'informations confidentielles au sujet de contreparties.

— **Données de propriété intellectuelle** : brevets, croquis industriels, recettes de fabrication.

La sévérité d'une attaque de type exfiltration de données est fortement corrélée au nombre de données perdues (pour les données personnelles), ainsi qu'au type de données et donc implicitement à la taille et au secteur d'activité de l'entreprise attaquée.

Les causes peuvent être accidentelles (perte d'un ordinateur professionnel avec accès au réseau de l'entreprise par exemple), malveillantes externes ou malveillantes internes.

1.1.2.2 Pertes de données

L'attaque de type perte de données vise à effacer, de manière temporaire ou définitive, des données d'une entreprise. Dans ce type d'attaque, l'auteur est moins susceptible de tirer profit de son attaque, mais l'entreprise victime peut subir des pertes très importantes, notamment à cause d'une interruption d'activité.

L'entreprise attaquée peut reconstituer ses données dans le cas où elle a prévu des copies de sauvegarde de celles-ci.

1.1.2.3 Attaque par déni de service

L'attaque de type déni de service correspond à la mise hors service temporaire ou définitive d'un élément opérationnel d'une entreprise. Cela concerne le plus souvent des attaques dématérialisées (attaque d'un site web) mais cela peut également avoir pour objectif des interruptions physiques, par exemple la mise hors service d'un élément de production physique d'une entreprise via la mise hors service de son système informatique.

En 2016, la moitié des grandes entreprises états-uniennes ont subi au moins une attaque par déni de service, et pour environ une attaque sur 8 cela a conduit à une interruption d'activité.

Les attaquants procèdent le plus souvent en saturant la bande passante d'un site internet en coordonnant un nombre important d'ordinateurs.

La plupart des attaques de type déni de service ne sont pas motivées par un gain pécuniaire direct mais seulement par la volonté de nuire ou d'intimider la victime.

Les différents types d'attaque vont se différencier, entre autres, par la couche du modèle de communication attaquée. Tous les systèmes informatiques utilisent la norme de communication réseau OSI

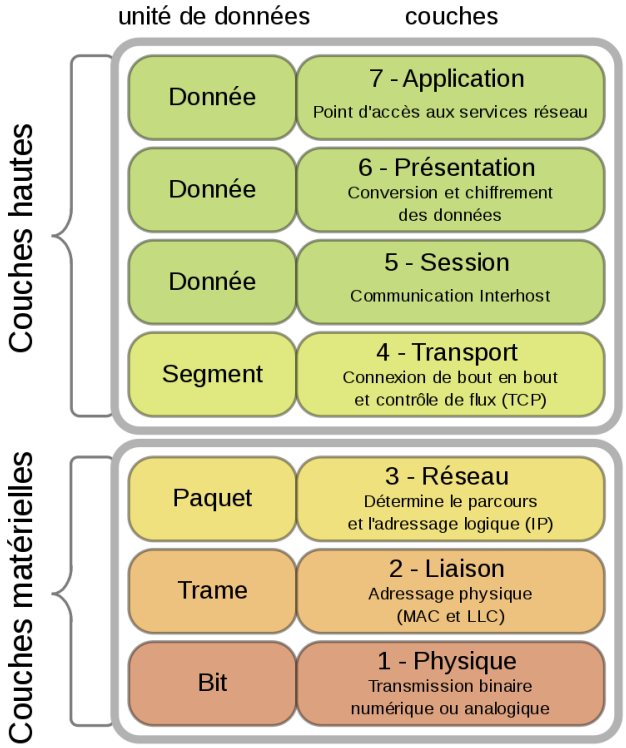


FIGURE 1.1 – Diagramme du modèle OSI

(*Open Systems Interconnection*) représenté en figure 1.1 - et décrivant les fonctionnalités nécessaires à la communication et l'organisation de ces fonctions.

Les différentes attaques de type déni de service sont :

— **Les attaques volumétriques**, ciblant les couches réseau et transport (couches 3 et 4) et les bombardent de requêtes pour saturer la bande-passante. Les utilisateurs légitimes subissent des ralentissements importants du réseau attaqué ou un arrêt complet du service.

— **Les attaques layer 7**, ciblant la couche 7 (applicative) et utilisent l'asymétrie de temps de calcul entre certaines requêtes utilisateurs et celui du serveur. Par exemple, la création d'un compte de messagerie auprès d'un fournisseur de service ne nécessite généralement que le remplissage d'un formulaire côté utilisateur. Étant donné que le fournisseur de service doit allouer beaucoup plus de ressources pour la création du compte utilisateur, il devient possible de saturer le serveur à faible coût. Une seule machine peut suffire à saturer un site internet, mais des protections relativement simples à mettre en œuvre (par exemple l'utilisation de CAPTCHA pour éviter le remplissage de formulaire de manière automatique) permettent de se prémunir de ce type d'attaques.

— **L'attaque SYN flood**, ciblant la couche 4 (Transport). Lors de l'initialisation d'une connexion TCP entre un client et un serveur, un échange de trois messages a lieu afin d'initialiser la connexion : 1- Client – Serveur ; 2 : Serveur – Client ; 3 – Client – Serveur. Si le client malveillant s'arrête au deuxième message, le serveur reste en attente du troisième message et prend

un certain temps avant de libérer les ressources réservées au client, typiquement une minute environ. En saturant de cette manière le serveur, le client légitime se retrouve dans l'incapacité d'établir une connexion avec le serveur.

Dans le cas des attaques volumétriques, les paramètres clés sont le volume de l'attaque (exprimé en Go par secondes – typiquement de 1 à plus de 1000 Go/s). Des centaines de milliers d'attaques volumétriques ont lieu tous les ans.

La moitié des attaques volumétriques durent moins de deux heures, 70% moins de six heures et 16% durent plus de 12 heures. La norme dans les contrats d'assurance Cyber est une durée de couverture jusqu'à 12h.

Il est possible de louer au marché des capacités d'attaques, l'ordre de grandeur est de 200 000 dollars US pour un volume de 50 à 100 Go/s pendant 24 heures, suffisant pour attaquer parmi les plus gros sites web.

1.1.2.4 Cyber-extorsion

Une attaque de type Cyber-extorsion consiste à bloquer une fonctionnalité d'un site internet ou à crypter des données sur une machine, et à demander une somme d'argent en échange du déblocage. Cela peut également concerner le fait de menacer une entreprise d'une attaque, et de demander une somme d'argent contre la levée de la menace.

Les principaux types d'attaques de Cyber-extorsion sont :

— **Le rançongiciel**, consistant à infecter une machine (un ordinateur personnel le plus souvent) et, dans un premier temps, à crypter les données présentes sur la machine de manière imperceptible pour la victime puis dans un second à bloquer la machine en proposant simultanément à la victime de payer une rançon afin de débloquer la machine et décrypter les données. Afin de continuer à être crédible, les attaquants décryptent généralement les données après paiement de l'utilisateur. Les montants ne dépassent généralement pas quelques centaines d'euros ou de dollars US, et sont le plus souvent exigés en crypto-monnaies (e.g. Bitcoin).

— La menace d'autres Cyber-attaques telles que le déni de service, en échange d'une rançon.

1.2 La directive Solvabilité 2 : rappels succints pour l'étude

Dans cette section, des rappels concernant le Pilier 2 de la directive Solvabilité 2 seront donnés, en se concentrant sur les points essentiels et d'intérêt pour notre étude. Pour des informations complémentaires, plus globales ou plus précises concernant d'autres aspects que ceux traités ici, le lecteur intéressé est invité à se référer à la littérature existante et souvent foisonnante sur le sujet ou aux textes réglementaires de référence.

1.2.1 Solvabilité 2 : les fondements

Un organisme assureur de l'Union Européenne (une compagnie d'assurance, un institut de prévoyance ou une mutuelle) est solvable s'il est capable de faire face à ses engagements dans la durée, c'est-à-dire être en mesure de verser les montants dus à ses cotisants et assurés.

FIGURE 1.2 – Machine bloquée par un rançongiciel CryptoLocker

De nombreux facteurs peuvent fragiliser et remettre en cause cette solvabilité, tels qu'une mauvaise maîtrise de ses risques assurantiels (par exemple taux garantis trop élevés, dégradation des sinistres), une insuffisance de provisionnement ou de tarification, des pertes sur les marchés financiers ou des risques opérationnels se réalisant : processus défaillants, disparition d'« homme-clé », risques émergents.

Solvabilité 2 a remplacé la directive Solvabilité 1 et a l'ambition de mettre en place des bonnes pratiques de marché concernant la gestion de tous ces facteurs de risque : suivi des portefeuilles et de leur rentabilité, devoir de prudence lors de la constitution des provisions techniques et réserve additionnelle pour faire face à des scénarios adverses, des règles de gestion prudente des actifs ainsi que l'implémentation d'une véritable culture de suivi du risque et de contrôle des différents processus de l'entité.

Le projet Solvabilité 1, signé dans les années 1970, a été conçu pour instituer un cadre réglementaire commun pour les assureurs et réassureurs au sein de l'Union Européenne, en évaluant les engagements de « manière prudente » et en garantissant une marge de solvabilité en se basant sur une approche comptable. Les principales critiques de Solvabilité 1 étaient la mauvaise quantification des risques portés par l'entreprise, en particulier une mauvaise évaluation des profils de risque spécifiques des entreprises. Solvabilité 1 a été amélioré au début des années 2000 puis une nouvelle réglementation a été développée et votée par la Commission Européenne le 22 avril 2009 : Solvabilité 2.

Solvabilité 2 a été mise en place au 1er janvier 2016 et concerne l'ensemble de l'Union Européenne.

Au-delà de la protection des assurés et cotisants, Solvabilité 2 a également pour objectif de renforcer l'intégration du marché européen de l'assurance et de favoriser la compétitivité des assureurs et réassureurs européens basée sur une vision « économique » et intégrant la notion de profil de risque spécifique à chaque entreprise.

Les principes de Solvabilité 2 sont :

— le principe d'évaluation en « juste valeur »,

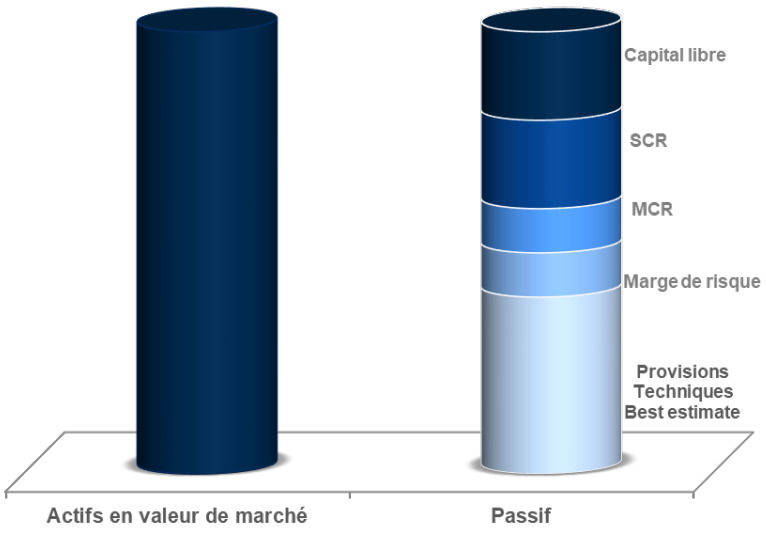


FIGURE 1.3 – Bilan simplifié S2

— des exigences réglementaires basées sur le risque,
— plus de gestion des risques et de contrôle interne,
— plus de principes et moins de règles (une approche « *principles based* » privilégiée à une approche « *rules based* » : les entreprises s'engagent à une gestion saine plutôt qu'à des règles potentiellement arbitraires),
— une meilleure information au public.

1.2.2 Les trois piliers de la directive Solvabilité 2

1.2.2.1 Pilier 1 : exigences quantitatives, l'entreprise doit être suffisamment capitalisée

Le Pilier 1 définit le principe de capital économique, ainsi que le type d'actifs éligibles à sa couverture. Le Pilier 1 correspond à la définition des exigences quantitatives, la directive imposant l'évaluation de plusieurs types de capitaux : provisions techniques, notions de SCR et de MCR.

L'une des innovations principales de Solvabilité 2 est l'évaluation du bilan en valeur économique, contrairement à une approche comptable. Cela revient à évaluer la valeur de marché des actifs et des passifs, c'est-à-dire le « montant pour lequel ils pourraient être échangés dans le cadre d'une transaction conclue, dans les conditions de concurrence normales, entre des parties informées et consentantes ». (Référence article 75 directive 2009/128/CE).

Le bilan S2 simplifié est présenté sur la figure 1.3

Plusieurs approches sont possibles : la formule standard et l'évaluation du capital économique en modèle interne.

1.2.2.2 Pilier 2 : exigences qualitatives, l'entreprise doit être bien gouvernée

Le Pilier 2 définit clairement la structure organisationnelle de l'entreprise afin d'expliciter les fonctions minimales attendues, il définit des fonctions clés ainsi que les responsabilités et les tâches de chacune. Le Pilier 2 définit les règles de prise de décision, de reporting interne, de communication, de coopération, de rémunération et de supervision.

L'objectif de ce pilier est la mise en œuvre de moyens adéquats par l'entreprise pour garantir un bon niveau de compréhension et de pilotage de l'entreprise afin de maîtriser sa solvabilité. Solvabilité 2 définit donc des exigences en matière de gouvernance et de gestion des risques : l'organisation du système de gouvernance doit comprendre une structure organisationnelle adéquate et une répartition transparente des responsabilités clés.

Le Pilier 2 définit quatre fonctions clés, devant être occupées par des dirigeants de l'entreprise ayant un niveau d'autorité suffisant, indépendants, compétents et honorables.

Les fonctions clés sont :

— la fonction de gestion des risques,
— la fonction actuarielle,
— la fonction de conformité,
— la fonction d'audit interne.

De plus, le Pilier 2 définit et cadre l' « ORSA » (Évaluation interne des risques et de la solvabilité - *Own Risk and Solvency Assessment*). L'ORSA lie des exigences quantitatives et qualitatives.

1.2.2.3 Pilier 3 : exigences de reporting, l'entreprise doit bien communiquer sur sa capitalisation

Le Pilier 3 énonce les exigences en matière de communication financière et de transparence vis-à-vis des autorités de contrôle et de communication au public, dans un format harmonisé défini au niveau européen. Certains pays définissent des exigences additionnelles adaptées à leurs spécificités nationales.

Les objectifs du Pilier 3 sont alors :

— harmoniser le reporting au niveau européen,
— standardiser et diffuser les pratiques sur le marché,
— simplifier le contrôle des autorités de contrôle.

Il existe trois types d'exigences :

— les rapports quantitatifs :
  — QRT (*Quantitative Reporting Templates*),
  — ENS (États Nationaux Spécifiques).

— Les rapports narratifs :
  — SFCR (*Solvency Financial Condition Report*),
  — RSR (*Regular Supervisory Report*).

— Un rapport ORSA.

Les QRT sont des templates harmonisés au niveau européen. Ils sont soit trimestriels soit annuels, certains sont rendus publics et d'autres sont uniquement destinés à l'Autorité de Contrôle.

Les ENS sont des templates quantitatifs définis au niveau national complémentaires des QRT. Certains pays membres n'exigent aucun ENS.

Le RSR n'est pas public et est remis à l'Autorité de Contrôle nationale. Il doit être produit au moins tous les 3 ans mais il peut être exigé plus régulièrement, en particulier en cas de changement important de profil de risque (acquisition d'une entité, arrêt ou forte croissance d'une activité, fusion, système de gouvernance...)

Le SFCR est exigé à une fréquence annuelle, et doit être rendu accessible au public.

Le rapport ORSA doit préciser comment l'ORSA a été réalisé et comment il a été intégré aux décisions stratégiques.

1.3 L'ORSA

L'ORSA (Évaluation Interne des Risques et de la Solvabilité – *Own Risk and Solvency Assessment*) désigne l'évaluation interne des risques et de la solvabilité. Dans la directive Solvabilité 2, les articles 44 et 45 traitent spécifiquement de la gestion des risques et de l'ORSA.

L'ORSA a la particularité de faire le lien entre les exigences quantitatives et qualitatives de la directive dans la mesure où il impose aux organismes d'assurance un suivi et un pilotage de son exposition au risque. C'est un outil d'aide à la décision se focalisant sur la solvabilité de l'entreprise.

L'ORSA possède les propriétés générales suivantes :

— l'ORSA repose sur le principe de proportionnalité : les moyens mis en œuvre doivent être proportionnés et adaptés aux risques des différentes activités de l'entreprise,

— l'ORSA est spécifique à l'entreprise,

— l'ORSA doit pouvoir fournir une vision prospective à l'horizon du plan stratégique (ou *Business Plan*),

— l'ORSA, en opposition au SCR et sa vision run-off, doit considérer la stratégie de l'entreprise concernant la souscription future,

— l'ORSA s'inscrit dans un cadre stratégique et doit servir d'aide à la décision en étant à la fois un outil de risk management et de pilotage,

— avec l'ORSA, les métriques de risque et les seuils de confiance ne correspondent pas forcément à la vision SCR considérant un seuil de confiance de 99,5% à horizon 1 an.

L'ORSA évalue trois composantes :

— **le Besoin Global de Solvabilité (« BGS »)**, compte tenu du profil de risque spécifique, de son appétence, de sa tolérance et de ses limites de risque ainsi que de sa stratégie commerciale,

— **le respect permanent** des exigences de capital et des exigences concernant les provisions techniques du Pilier 1,

— **la mesure dans laquelle le profil de risque de l'entreprise s'écarte des hypothèses sous-tendant le capital de solvabilité requis du Pilier 1**, calculé à l'aide de la formule standard ou avec un modèle interne partiel ou intégral.

1.3.1 Le Besoin Global de Solvabilité (BGS)

Le Besoin Global de Solvabilité (BGS) est estimé en fonction du profil de risque spécifique, de son appétence, de sa tolérance et de ses limites de risque ainsi que de sa stratégie commerciale.

1.3.1.1 Profil de risque

Le profil de risque est le niveau de risque auquel l'entreprise est actuellement exposée. D'après l'article 44 de la directive, l'ORSA doit couvrir l'ensemble des risques significatifs pouvant avoir une influence sur le niveau actuel des fonds propres ou des garanties offertes aux assurés. L'ORSA complète donc l'évaluation quantitative du Pilier 1 avec les risques non pris en compte, quantifiables ou non.

En particulier, le risque Cyber n'est pas explicité dans l'évaluation du Pilier 1. En outre, la formule standard du Pilier 1 ne prend pas en compte le risque de liquidité, le risque des spreads souverains, le risque de réputation et les risques stratégiques, le risque d'évolution de l'environnement légal, le risque d'inflation.

L'ORSA présente également les moyens d'atténuation et de transfert de risque tels que par exemple un plan d'urgence, un suivi d'indicateurs de vulnérabilité, le recours à la co-assurance ou à la réassurance, par une allocation d'actifs spécifique.

Surtout, l'ORSA propose une vision intégrée du profil de risque, par opposition à une vision en silo dans laquelle chaque équipe ou département évalue son profil de risque de manière indépendante.

1.3.1.2 Appétence, tolérance et limite au risque

L'appétence au risque est le niveau de risque agrégé que la compagnie est prête à prendre en vue de la poursuite de son activité et afin d'atteindre ses objectifs stratégiques (rentabilité, croissance, mix produit...) à l'horizon de son plan stratégique. L'appétence au risque est déterminée par les instances de gouvernance de l'organisme et s'exprime sous la forme de métriques de risque.

L'appétence au risque est directement associée à la définition des orientations stratégiques, au niveau du Conseil d'Administration.

La tolérance au risque représente le niveau de risque que l'entreprise accepte de prendre en vue de la poursuite son activité et de son développement au niveau de chaque famille et catégorie de risque. Ce niveau de risque est déterminé en fonction de l'ampleur et les types de risques que l'organisme est prêt à tolérer afin de réaliser ses objectifs stratégiques, dans le respect du cadre d'appétence pour le risque global. Cette déclinaison peut être réalisée aussi bien par famille de risques voire sous-risques,

par zone géographique, par ligne d'activité, par compagnie ou entité (dans le cas d'un Groupe). La tolérance au risque sera exprimée selon les mêmes métriques que celles de l'appétence au risque. Les tolérances au risque doivent pouvoir être agrégées pour être vues de façon globale (appétence).

La tolérance au risque est directement associée à la mise en œuvre de la stratégie, au niveau de la Direction générale et de la Direction des risques.

La limite opérationnelle au risque est le niveau de risque que l'entreprise est prête à prendre par classe de risque. C'est la déclinaison de la tolérance au risque au niveau opérationnel.

La limite opérationnelle est directement associée à la gestion opérationnelle des activités, par exemple au niveau de la souscription, de la tarification, de la réassurance, de la gestion actif/passif, de l'allocation d'actifs.

1.3.1.3 Métriques de risque

Pour pouvoir mesurer et suivre l'appétence au risque de l'entreprise, il convient de définir les métriques à retenir. Chaque entreprise définit ses propres métriques, pouvant être qualitatives ou quantitatives. Les métriques quantitatives sont préférées, les mesures qualitatives étant réservées aux métriques difficiles à quantifier (par exemple pour le risque de réputation et d'image).

Les métriques de risque doivent :

— permettre de mettre en place une communication globale efficiente entre les parties prenantes. Pour cela, les métriques de risque retenues devront être intelligibles par l'ensemble des parties prenantes et en nombre limité afin de ne pas compliquer la prise de décision,

— être cohérentes avec les risques sous-jacents,

— pouvoir se ventiler et s'agréger par catégorie de risque (des métriques trop spécifiques à certains risques pourraient donc se révéler inadaptées à une généralisation),

— être adaptées aux ressources de l'organisme.

Pour chaque métrique, l'entreprise doit ensuite associer un seuil de confiance et un horizon temporel.

Les métriques sont définies au niveau de l'appétence au risque puis déclinées au niveau de la tolérance au risque (on parle d'approche *top-down*).

Les limites opérationnelles sont ensuite déterminées via l'identification et la sélection d'indicateurs de risque. Ces indicateurs doivent être en nombre limité pour garantir leur compréhension et quantifiés (avec la définition de limites hautes et basses).

1.3.1.4 Évolution du profil de risque et modification des limites

Une fois l'appétence et les tolérances au risque déterminées et leurs limites fixées, il conviendra d'évaluer le profil de risque actuel et de vérifier qu'il est conforme à ces limites. Dans le cas contraire, des actions managériales correctrices sont à envisager (par exemple réduction de l'exposition, utilisation de couvertures financières - *hedging* - ou techniques - réassurance).

Cependant, dans le temps, le profil de risque va naturellement évoluer (en fonction par exemple des marchés financiers, du mix-produits, de la richesse de l'organisme, des décisions stratégiques). Il s'avère donc nécessaire de contrôler régulièrement qu'il reste en ligne avec les limites de tolérance et d'appétence fixées.

En cas d'évolution significative, l'organisme pourra alors au choix :

— décider de maintenir ses limites et de mettre en œuvre des actions correctrices,

— déterminer de nouvelles cibles et limites pour ses tolérances ou son appétence au risque. Dans un tel cas, les limites de niveau(x) inférieur(s) devront également être revues.

En pratique, à chaque production de l'ORSA il conviendra, sur base de l'évaluation du profil de risque (et de la mesure a posteriori des indicateurs), de réfléchir à d'éventuelles modifications de l'appétence, de la tolérance et/ou des limites opérationnelles.

Entre deux processus ORSA, le suivi permanent de la solvabilité et de l'appétence au risque est généralement effectué via l'analyse des indicateurs opérationnels. Lorsque l'un de ces indicateurs montre un niveau de risque élevé ou en cas de dépassement de limites importantes validé par les organes dirigeants, la mise à jour du processus ORSA pourra s'avérer nécessaire. Le cas échéant, il faudra éventuellement revoir la politique de gestion des risques voire lancer un ORSA exceptionnel.

1.3.2 Le respect des exigences du Pilier 1

Deux points sont concernés par cette évaluation : les exigences de capital et les provisions techniques.

Concernant l'exigence de capital, l'objectif est de veiller à son respect compte tenu de l'évolution du profil de risque de l'activité, en considérant les impacts stratégiques et environnementaux potentiels (marché, réglementation) qu'il est possible d'anticiper. L'évaluation devra être réalisée à un horizon de temps au moins égal à celui du business plan. Des stress tests et des scénarios seront mis en œuvre afin d'analyser la sensibilité et la robustesse du profil de risque. Il peut par exemple s'agir de chocs sur le rendement des actifs, sur les ratios sinistres sur primes.

Concernant les provisions techniques, une revue doit être effectuée de telle sorte à justifier l'adéquation du niveau de provisions techniques. Cette revue est placée sous la responsabilité de la fonction actuarielle.

1.3.3 Écart entre le profil de risque et les hypothèses sous-jacentes

Cette étude doit justifier de la pertinence de l'application, par exemple, des hypothèses de la formule standard au regard du profil de risque de l'entreprise.

En cas de déviation suffisamment justifiée, des paramètres spécifiques à l'entreprise peuvent venir remplacer certains paramètres de la formule standard (exemple : coefficient de volatilité concernant le risque de réserves non-vie) : les USP (*Undertaking Specific Parameters*).

Chapitre 2

Étude de bases de données Cyber

Il existe deux bases de données publiques et gratuites recensant des évènements de types Cyber-attaques. Ce chapitre rappelera les critères de qualité d'une base de données au sens Solvabilité 2 puis évaluera chacune des deux bases à la lumière de ces critères de qualité.

2.1 Qualité des données - Notions générales

De manière générale et tel que rappelé par la directive Solvabilité 2, la qualité des données s'apprécie au regard de trois critères :

— l'exhaustivité,
— l'exactitude,
— la pertinence.

2.1.1 Exhaustivité

Au sens de Solvabilité 2, l'exhaustivité désigne :

— l'identification des principaux groupes de risques,

— une granularité suffisante pour identifier les tendances et l'évolution des risques sous-jacents,

— des historiques suffisants et disponibles.

Dans le cadre de la problématique Cyber, cela se traduit par une granularité suffisante pour *discriminer* le comportement de certains risques en fonction de certains critères (par exemple le secteur d'activité d'une entreprise assurée).

2.1.2 Exactitude

L'exactitude désigne le fait pour des données :

— d'être adaptées et appropriées à l'usage qui leur est destiné (i.e. l'établissement d'hypothèses, etc..) et utilisables pour le risque assuré,

— de refléter les risques auxquels est exposé l'assureur.

Dans le cadre de la problématique Cyber, cela se transpose en des données représentant bien le risque. Pour la sévérité cela peut correspondre à une absence de biais lié à la sur-représentation d'un secteur d'activité. Pour la fréquence, la question est de savoir si les données reportées sont représentatives du nombre de sinistres subis dans une région au cours d'une période donnée, la question de la fréquence est donc directement liée à la constance des sources d'acquisition au cours du temps.

2.1.3 Pertinence

La pertinence signifie que des données :

— sont exemptes d'erreurs matérielles et d'omissions,

— contiennent des informations stockées de manière adéquate et avec une mise à jour fréquente, ainsi que l'absence de doublons,

— satisfont à un niveau général de confiance.

Pour les données Cyber, cela se traduira par la présence ou non d'erreurs et d'ommissions, et par une attention particulière portée aux sources et à l'origine des données.

2.2 La base PRC

La base PRC (*Privacy Rights ClearingHouse*) est disponible gratuitement sur le site de l'association *Privacy Rights* `https://privacyrights.org/about`. C'est une association à but non lucratif fondée en 1992 ayant l'ambition de protéger la vie privée des citoyens états-uniens en fournissant des informations concernant les droits des individus ainsi que des moyens de défendre leurs droits.

Dans ce mémoire, la base a été extraite le 23 janvier 2019. C'est la même extraction utilisée pour la rédaction de la première version du papier [Farkas *et al.*, 2019]. La base contient 8860 incidents.

2.2.1 Des sources hétérogènes alimentent la base PRC

PRC a recueilli des données issues de différentes sources :

— des agences gouvernementales états-uniennes au niveau fédéral. Ces données publiques sont disponibles grâce à des lois obligeant les entreprises subissant des pertes de plus de 500 Données à Caractère Personnel (DCP) à communiquer publiquement des informations précises sur ces évènements. Le principal contributeur de cette catégorie est le *Department of Health and Human Services (HHS)*. Il reporte les incidents des entreprises du domaine de la santé,

— des agences gouvernementales états-uniennes au niveau de chacun des 50 États. Depuis 2018, les États reportent les pertes de données ayant eu lieu dans l'État, mais selon des règles propres à chaque État. Le choix du seuil de reporting est, au niveau fédéral, de 500 DCP,

— les pertes de données reportées dans les médias,

— les pertes de données reportées par d'autres organisations à but non lucratif, telles que Databreaches.net.

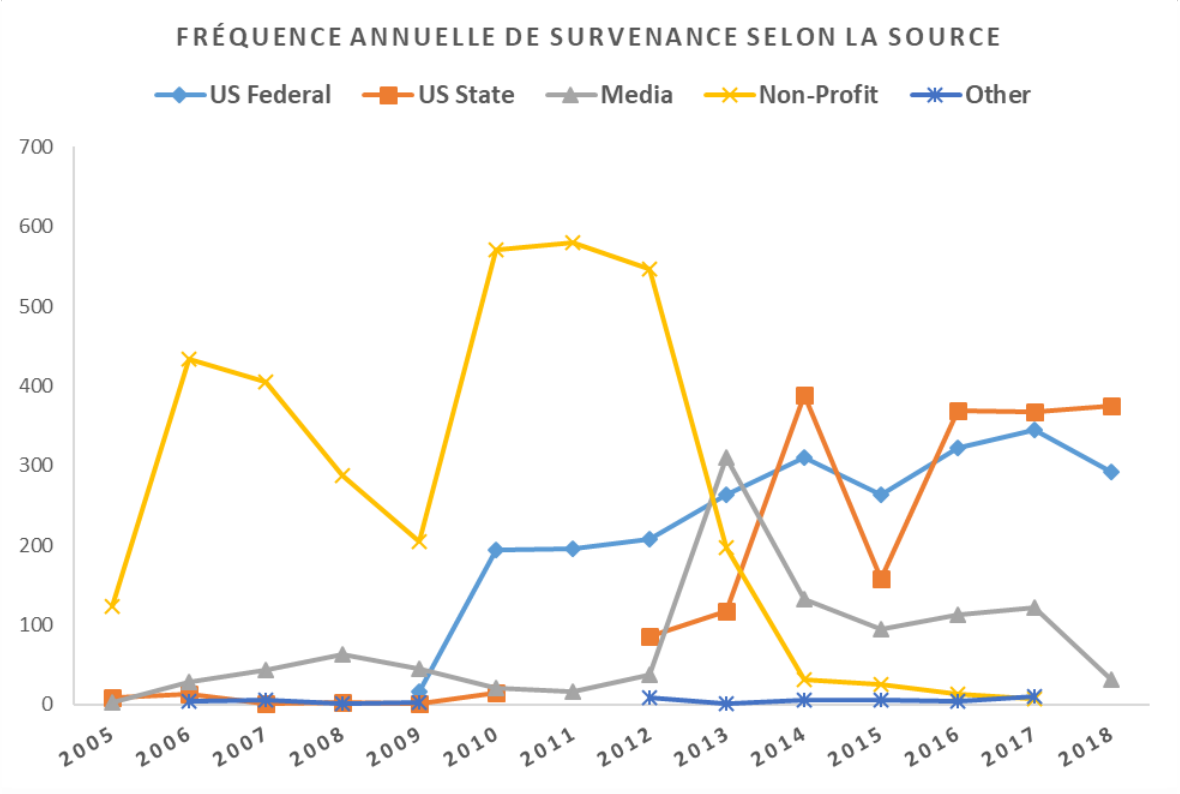


FIGURE 2.1 – Fréquence annuelle de sinistres dans la base PRC, selon la source

Dans la base PRC, la source de données est directement disponible pour chaque incident. De grandes variations de fréquence selon les sources sont observables.

La base PRC a dont été alimentée par différentes sources, et la contribution de chacune de ces sources a énormément varié au cours du temps.

2.2.2 Champs disponibles

La base PRC contient 13 colonnes, et les champs sont remplis de manière directement interprétable, ils possèdent des taux de présence de la donnée conséquents. Cela en fait une base **facile d'accès** sur laquelle presque aucun retraitement n'est nécessaire.

La base PRC contient un seul champ quantitatif (*Total.records*), représentant le nombre de DCP en jeu pour un incident donné. L'information est assez bonne car 6641 incidents contiennent un nombre de DCP non nuls.

2.3 La base VERIS

La *VERIS Community Database* ou VCDB est une base collaborative reportant des incidents de sécurité et respectant un reporting standardisé : VERIS.

La collecte d'information dans la base VCDB a débuté en 2013. La base se concentre exclusivement sur les incidents de type perte de données (*Data breaches*).

Champ_retraité	Champ_original	Modifications / retraitements nécessaires	Usage possible
Date.Made.Public	Date.Made.Public	Aucun	Date
Company	Company	Aucun	Calibration de la fréquence
City	City	Aucun	Zone géographique et réglementaire
State	State	Aucun	Zone géographique et réglementaire
Type.of.breach	Type.of.breach	Aucun	Classifie selon le type d'attaque
Type.of.organization	Type.of.organization	Aucun	Classifie selon le secteur
Total.Records	Total.Records	Aucun	Information quantitative
Description.of.incident	Description.of.incident	Aucun	Recherche d'informations complémentaires
Gather.Information.Source	Information.Source	Regroupement selon les 4 macro-catégories de sources	Classifie selon la source
Source.URL	Source.URL	Aucun	Recherche d'informations complémentaires
Year.of.Breach	Year.of.Breach	Aucun	Date
Latitude	Latitude	Aucun	Information géographique, pertinent pour les USA seulement
Longitude	Longitude	Aucun	Information géographique, pertinent pour les USA seulement

TABLE 2.1 – Les champs de la base PRC

VERIS signifie *Vocabulary for Event Recording and Incident Sharing*. C'est un ensemble de métriques définissant un langage commun pour la description d'incidents de manière structurée.

La base VERIS est disponible gratuitement sur le site de la communauté VERIS : `http://veriscommunity.net`.

Les données proviennent des deux souces : le *Department of Health and Human Services (HHS)* qui est une organisation fédérale états-unienne, et des différents outils mis à la disposition du public par le *State Attorney General (SAG)* (ou procureur général d'État) de chacun des 50 états américains. Chaque SAG est le principal conseiller juridique du gouvernement et est chargé de l'application de la loi dans l'État. Enfin, la base est complétée d'informations issues des médias. Sur le site de *veris community*, il est précisé que près de la moitié des incidents reportés proviennent du HHS.

Ainsi les différentes sources sont presque exclusivement états-uniennes. Les incidents reportés sont issus d'informations fédérales (environ 50%), étatiques ou des médias. De plus, il existe un biais conséquent de telle sorte que les incidents liés au secteur de la santé sont sur-representés dans la base VERIS.

La source n'est pas précisée par incident dans la base de donnée.

Pour ce mémoire, la base d'étude a été extraite le 1er avril 2019. A cette date, la base contient 8198 incidents ayant eu lieu entre 1971 et 2019. Ils ont été reportés à partir de 2013.

Pour ne pas confondre *reporting year* avec année de déclaration, un terme assurantiel, l'anglicisme année de reporting sera conservé.

Année de survenance	Année de reporting 2013	2014	2015	2016	2017	2018	2019	Total
1971	1	0	0	0	0	0	0	1
1984	1	1	0	0	0	0	0	2
1994	0	1	0	0	0	0	0	1
1998	1	0	0	0	0	0	0	1
1999	1	0	1	0	0	0	0	2
2000	1	0	0	0	0	0	0	1
2001	3	1	0	1	0	0	0	5
2002	1	5	0	0	0	0	0	6
2003	6	4	0	1	1	0	0	12
2004	7	7	1	2	0	0	0	17
2005	10	8	0	2	0	1	0	21
2006	11	10	1	0	0	0	0	22
2007	22	22	5	2	0	0	0	51
2008	38	29	5	9	0	1	0	82
2009	39	30	8	13	2	1	0	93
2010	278	281	17	9	2	0	0	587
2011	230	247	44	19	5	1	0	546
2012	770	403	58	30	5	3	1	1270
2013	1143	586	128	55	13	11	0	1936
2014	0	606	282	83	22	10	1	1004
2015	0	0	38	822	37	11	2	910
2016	0	0	0	592	197	29	1	819
2017	0	0	0	0	456	85	1	542
2018	0	0	0	0	0	234	32	266
2019	0	0	0	0	0	0	1	1
Total	2563	2241	588	1640	740	387	39	8198

TABLE 2.2 – Incidents de la base VERIS répartis par années de survenance et de reporting

La base VERIS a la particularité de contenir un très grand nombre de colonnes. Pour cette même raison, de nombreuses colonnes ne sont pas utiles. La notion d'*enumeration* ou "énumération" est spécifique à ce standard et est essentielle à la compréhension de la structure des données. C'est un type de champ. Nous différencierons les énumérations des variables, et nous considérons que ce sont chacun des types de champ. De même, un champ ne peut être qu'une énumération ou une variable.

2.3.1 Notion d'énumération

La base VERIS contient 2 441 colonnes. Cela semble énorme, mais en réalité il est possible de regrouper la plupart de ces colonnes en groupes de colonnes. Ces groupes correspondent à la notion d'énumération définie par le standard VERIS. Une énumération correspond à un champ pouvant prendre un nombre fini de modalités. Chaque modalité sera reportée comme une colonne de booléens (Vrai ou Faux). Ainsi, s'il existe 5 énumérations contenant chacune 3 modalités, 15 colonnes seront créées. Un exemple d'énumération : *victim.orgsize*, contenant 2 modalités : *Small* et *Large*, pour les 10 premiers incidents de la base, représenté en table 2.3.

Sur cet exemple, il est possible d'observer qu'un incident ne peut-être à la fois *Small* et *Large* (bien sûr, cette observation est à vérifier sur la totalité des incidents et non pas qu'avec les dix premiers). Un incident peut aussi n'être ni *Small* ni *Large*.

	victim.orgsize.Small	victim.orgsize.Large
1	0	1
2	1	0
3	0	1
4	0	0
5	0	1
6	0	1
7	1	0
8	0	0
9	1	0
10	0	0

TABLE 2.3 – Un exemple d'énumération

actor.External	asset.variety.Server
actor.Internal	asset.variety.Network
actor.Partner	asset.variety.User.Dev
actor.Unknown	asset.variety.Media
action.Malware	asset.variety.Person
action.Hacking	asset.variety.Kiosk.Term
action.Social	asset.variety.Unknown
action.Physical	asset.variety.Embedded
action.Misuse	attribute.Confidentiality
action.Error	attribute.Integrity
action.Environmental	attribute.Availability
action.Unknown	

TABLE 2.4 – Composition des macro-catégories de la grille A^4

De manière générale, la qualité de l'information au sein des énumérations est à étudier au cas par cas.

2.3.2 Grille A^4

La grille A^4 est un concept introduit par VERIS, et est spécifique à cette base d'incidents. Cela correspond à la combinaison de quatre macro-catégories : *Actor, Action, Asset* et *Attribute*. Les valeurs possibles pour chaque macro-catégorie sont représentées en figure 2.2.

Il y a donc 3 valeurs pour *Actor* (*Unknown* étant exclu), puis 7 pour *Action*, 7 pour *Asset* et 3 pour *Attribute*.

Il existe donc $3 \times 7 \times 7 \times 3 = 441$ combinaisons possibles. Or sur le site de VERIS, il est présenté la grille A^4 de taille $3 \times 7 \times 5 \times 3 = 315$ combinaisons possibles (en effet certaines modalités de asset ne sont pas présentes sur l'illustration du site internet, la base ayant très certainement évolué entre temps).

La grille A^4 est la représentation de ces combinaisons dans un plan en deux dimensions :

N'importe quel incident dont les quatre macro-catégories sont correctement remplies peut donc être associé à une des cases de la grille A^4.

2.3.3 Exploration de la base et retraitements

Les retraitements sur la base VERIS sont détaillés en annexe A.

	External.Malware	External.Hacking	External.Social	External.Misuse	External.Physical	External.Error	External.Env	Internal.Malware	Internal.Hacking	Internal.Social	Internal.Misuse	Internal.Physical	Internal.Error	Internal.Env	Partner.Malware	Partner.Hacking	Partner.Social	Partner.Misuse	Partner.Physical	Partner.Error	Partner.Env
Server.Conf	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Server.Integ	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Server.Avail	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
Network.Conf	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
Network.Integ	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
Network.Avail	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126
User.Conf	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147
User.Integ	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168
User.Avail	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189
Media.Conf	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210
Media.Integ	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231
Media.Avail	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252
People.Conf	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273
People.Integ	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294
People.Avail	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315

FIGURE 2.2 – Grille A^4 à 315 combinaisons telle que présentée sur le site VERIS

La base VERIS contient 20 champs retraités, tels que représentés figure 2.5.

Champ_retraité	Champ_original	Modifications / retraitements nécessaires	Usage possible
incident_id	incident_id	Aucun	Certains enregistrements (5 en tout) sont dupliqués, ce champ permet de retraiter
plus.github	plus.github	Aucun	Pas de signification explicite
reference	reference	Aucun	Lien vers des articles de journaux en ligne
incident_year	timeline.incident.year	Aucun	Date
incident_month	timeline.incident.month	Aucun	Date
incident_day	timeline.incident.day	Aucun	Date
victim.state	victim.state	Aucun	Information géographique, pertinent pour les USA seulement
Malware.name	campaign_id & action.malware.name	Retraitement sur-mesure. **Documenté spécifiquement**	Flag liés à la notion d'évènement d'accumulation
Actor.external.name	actor.external.name	L'information n'est conservée que si elle concerne au moins deux enregistrements. **Documenté spécifiquement**	Flag liés à la notion d'évènement d'accumulation
Actor	4 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Action	8 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Asset	8 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Attribute	3 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Discovery_method	discovery_method enumeration	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Ressemble à la grille A4
Size1	victim.employee_count	Size1 ne peut prendre que les modalités: "Small" ; "Large" or NA ; si l'information originale est "Small" ou dans une catégorie avec moins de 1000 employés alors "Small" ; si "Large" ou une catégorie à plus de 1000 employés alors "Large", sinon NA. **Documenté spécifiquement**	Classifie selon la taille de l'entreprise (en nb employés)
Size2	victim.employee_count	Reprend la même information que l'information initiale. Si la catégorie détaillé n'est pas disponible (si l'information originale est "Small", "Large" ou NA), alors "Unknown". Par conséquent, Size2 contient plus de NA que Size1.	Classifie selon la taille de l'entreprise (en nb employés)
sector	victim.industry.name	Aucun	Classifie selon le secteur de l'entreprise
number_records	attribute.confidentiality.data_total	Aucun	Information quantitative
revenue_USD	victim.revenue.amount & victim.revenue.iso_currency_code	Le montant en monnaie originale est simplement converti en USD (utilisant du taux FY18).	Information quantitative
loss_USD	impact.overall_amount & impact.iso_currency_code	Le montant en monnaie originale est simplement converti en USD (utilisant du taux FY18).	Information quantitative

TABLE 2.5 – Les 20 champs de la base VERIS retraitée

Avant / Après retraitement	Paramètre	PRC	VERIS
Base brute	Nombre d'enregistrements	**8860**	**8198**
	Nombre de colonnes	**13**	**2441**
	Nombre de champs	**13**	**191**
Base retraitée	Nombre d'enregistrements	**8860**	**8196**
	Nombre de colonnes	**13**	**20**
	Nombre de champs	**13**	**20**
	Nombre de variables quantitatives	1	4
	Liste des variables quantitatives	Nombre de données perdues	1. Nombre de données perdues 2. Taille de l'entreprise (nb employés) 3. Chiffre d'affaire 4. Perte pécuniaire associée à un incident
	Catégories de classification	3	3
	Champs de classification	1. Type d'attaque 2. Secteur 3. Source	1. Type d'attaque (Grille A4) 2. Secteur 3. Taille d'entreprise

TABLE 2.6 – Comparaison des deux bases de données

2.4 Synthèse des retraitements et correspondance entre les incidents des deux bases

Dans cette section, il est étudié dans quelle mesure les deux bases reportent ou non les mêmes incidents.

En effet, il est possible de remarquer que les deux bases ont été alimentées à partir de sources hétérogènes, mais que ces sources sont assez similaires selon les bases. Chaque base a été alimentée par des informations fédérales, des informations de chaque état et des informations issues des médias. La base PRC a de plus été alimentée par des organisations à but non lucratif. De plus, le nombre d'incidents dans chacune des bases est similaire. Il est donc naturel de s'intéresser aux correspondances qu'il pourrait y avoir entre les incidents des deux bases.

Les éléments de l'étude sont détaillés en annexe B.

La conclusion de cette étude est que les incidents des deux bases correspondent mal : les deux bases ne reportent pas les mêmes incidents.

De cette observation, il n'est non seulement pas possible d'étudier la fréquence d'occurrence d'incidents avec l'une ou l'autre des bases, mais cela prouve également que toute étude de fréquence basant sur l'une des deux bases n'est pas fondée.

2.4.1 Tableau de synthèse des bases retraitées

Après étude des bases, de manière globale et sur quelques cas particuliers, il apparaît qu'il est impossible d'enrichir l'une des deux bases avec les informations contenues dans l'autre tant les incidents reportés sont différents et tant la manière dont ils sont reportés différent également notamment au niveau des dates.

2.5 Conclusion et choix de la base d'étude

2.5.1 Limites des bases étudiées

2.5.1.1 Fréquence

Toute étude portant sur la fréquence d'évènements basée sur l'une ou l'autre des bases est fortement déconseillée et ce quelle que soit la base : il n'y a pas de cohérence entre les fréquences des deux bases, il manque des évènements dans chacune des bases, certains évènements sont comptés plusieurs fois, le nom des entreprises est reporté de manière hétérogène et d'importantes différences de nombre de sinistres reportés sont observables entre l'une et l'autre base pour les deux exemples retenus et étudiés en annexe (Walgreens et Facebook).

2.5.1.2 Sévérité

La sévérité d'une attaque de type perte de données peut être évaluée soit en termes de nombre de DCP perdues soit en termes de coût. L'étude du nombre de DCP perdues peut se faire à partir de la base PRC ou à partir de la base VERIS. Cependant, la modélisation d'un nombre de DCP n'est pas suffisante, l'étude du lien entre nombre de données perdues et coût d'une attaque est fondamentale.

Taux de présence des variables quantitatives dans les bases

La base PRC ne contient qu'une variable quantitative, le nombre de données perdues. 6641 incidents sur 8860 possèdent une valeur pour cette variable (taux de présence de 75%).

présence (en nombre)	Size1	number_records	revenue_USD	loss_USD
Size1	5479	2602	495	214
number_records		3748	233	145
revenue_USD			507	15
loss_USD				302

présence (en taux)	Size1	number_records	revenue_USD	loss_USD
Size1	66,8%	31,7%	6,0%	2,6%
number_records		45,7%	2,8%	1,8%
revenue_USD			6,2%	0,2%
loss_USD				3,7%

TABLE 2.7 – Taux de présence des variables quantitatives dans la base VERIS

Aide à l'interprétation de la figure 2.7 :
66,8% des incidents ont leur variable Size1 renseignée, et 31,7% ont à la fois l'information Size1 et l'information *number_ records* renseignée.

Information relative à la taille de l'entreprise

La taille de l'entreprise peut être exprimée en fonction du nombre d'employés (*Size1*) ou du chiffre d'affaires (*revenue_ USD*). La variable *Size1* sera privilégiée étant donné le faible taux de présence de la variable *revenue_ USD*.

2.5.2 Qualité des données

Dans cette section est présentée la synthèse de l'étude de la qualité des données au sens Solvabilité 2 des deux bases de données.

		Qualité des données PRC	Qualité des données VERIS
Exhaustivité	Identification de différents groupes de risques	**Bonne à moyenne** ; identifie selon le type d'attaque, le secteur d'activité et la source de l'information. Peu de champs sont disponibles mais ils semblent être les plus pertinents pour caractériser le risque et sont correctement remplis.	**Bonne** : Un nombre très important de champs sont présents ; cependant ils contiennent souvent peu d'information. Les champs essentiels sont correctement remplis.
	Granularité suffisante	**Bonne** granularité.	**Bonne** granularité
	Historique suffisant	**Moyenne à mauvaise** ; Environ 13 années d'historique, cependant absence de stabilité des sources d'acquisition des données	**Moyenne à mauvaise** : Quelques sinistres anciens mais la majorité des sinistres survenus à partir de 2005. Irrégularités dans la fréquence ne pouvant être expliquée par une évolution du risque mais par une manière d'acquérir les données. L'information clé du coût n'est présente que pour une faible proportion des enregistrements.
Exactitude	Données adaptées au risque	**Moyenne à mauvaise** : Les données représentent bien des sinistres de type pertes de données, mais il existe un fort biais sectoriel (sur-représentation du secteur de la santé), un biais lié à la source (les pertes majeures *ont des raisons* d'être sur-représentées) ; absence pure et simple d'information concernant le coût des sinistres : la seule variable quantitative est le nombre de données perdues. Les sinistres reportés représentent une faible proportion de l'ensemble des sinistres sur la période et du périmètre (USA), de plus les sources sont multiples : la base PRC ne semble donc pas adaptée à la capture de caractéristiques (structure de modèle et calibration) de la fréquence d'un processus.	**Moyenne** : Les données représentent bien des sinistres de type pertes de données, mais il y a des raisons de penser que l'information est biaisée avec une sur représentation des sinistres majeurs ; il existe plusieurs variables quantitatives dont le coût des données. Les sinistres reportés représentent une faible proportion de l'ensemble des sinistres sur la période et du périmètre (USA), de plus les sources sont multiples : la base VERIS ne semble donc pas adaptée à capture de caractéristiques (structure de modèle et calibration) de la fréquence d'un processus.
	Reflète les risques d'un assureur	**Moyenne à mauvaise** : Les coûts n'existent pas dans la base. Cette base peut permettre de comprendre le comportement du nombre de données perdues uniquement. Pas de possibilité de répartir les coûts (ceuxc-ci n'étant pas présents) par garantie.	**Bonne à moyenne** : Le nombre de données perdues et le coût associé existent dans la base mais il n'existe pas de possibilité pratique de répartir les coûts par garantie.
Pertinence	Absence d'erreurs	**Moyenne à mauvaise** : De nombreux retraitements sont nécessaires (ex: noms des entreprises), des incohérences sont présentes dans les variables de classification, notamment celle du secteur. Il est difficile de vérifier la variable quantitative, basée sur des déclarations d'entreprises induisant un problème de consistance méthodologique.	**Moyenne** : Certains retraitements essentiels sont nécessaires, il y a des incohérences sur certains champs. Des informations déclaratives avec un problème de consistance méthodologique.
	Stockage adéquat de l'information	**Excellente** : L'information étant stockée dans une base de données publique, il n'y pas de problématique liée à ce sujet.	**Excellente à bonne** : L'information étant stockée dans une base de données publique, il n'y pas de problématique liée à ce sujet. L'information est peu facilement accessible car il existe de nombreuses informations inutiles.
	Niveau de confiance général	**Moyenne à mauvaise** : Toute étude sur la fréquence semble inadaptée, de plus il existe des biais sectoriel et lié à la source pour l'étude du nombre de données.	**Bonne à moyenne** : Toute étude sur la fréquence semble inadaptée, il existe *probablement* un biais lié à l'acquisition des données, enfin une faible proportion des enregistrements possède un coût, et ce coût n'est pas bien ventilé par garantie.

TABLE 2.8 – Comparaison de la qualité des données des deux bases

En conclusion et pour synthétiser la figure 2.8, la base VERIS est de qualité générale équivalente à légèrement meilleure que la base PRC. Les deux bases étant caractérisées par la nécessité d'effectuer de nombreux retraitements pour corriger certaines incohérences, et par la présence significative d'erreurs.

L'avantage majeur de la base VERIS est qu'elle contient la variable du coût d'un incident, contrairement à la base PRC qui ne contient pas ce champ.

2.5.3 Choix de la base d'étude

La base PRC a déjà été étudiée de manière extensive dans la littérature scientifique, ces études ne seront pas remises en question dans le cadre de ce mémoire et la base PRC ne sera pas étudiée dans les chapitres suivants. Lorsque cela sera pertinent, des études existantes pourront être citées pour mettre en perspective le résultat des études basées sur la base VERIS.

La base VERIS est également privilégiée car la présence de la variable de coût ouvre le champ des applications actuarielles avec notamment la capture d'une incertitude et la calibration de modèles stochastiques.

Chapitre 3

Structure et calibration de la modélisation du risque Cyber de perte de Données à Caractère Personnel

3.1 Quelques études préliminaires

3.1.1 Ponemon

Ponemon Institute LLC est un institut de recherche publiant chaque année depuis 2006 un rapport sur le coût des Cyber-attaques de type perte de données basée sur des échanges approfondis avec des entreprises ayant subies ce type d'attaques au cours de l'année écoulée.

Le dernier rapport, [Ponemon et IBM, 2019], a été réalisé sur la base d'échanges ayant eu lieu entre Juillet 2018 et Avril 2019. 507 entreprises ont été interrogées pour l'édition 2019. La zone géographique d'activité la plus représentée parmi ces entreprises sont les États-Unis d'Amérique, avec 64 entreprises, puis l'Inde (45), le Royaume-Uni (45), l'Allemagne (36), le Brésil (35), le Japon (33), la France (32).

Ponemon ne propose pas réellement de modèle mais des données descriptives moyennes par taille d'entreprise, d'attaques, de secteur d'activité.

Dans cette section sera proposée une approche naïve de l'estimation du coût et de la fréquence d'une attaque de type Cyber subie par deux assureurs fictifs en se basant uniquement sur les chiffres de Ponemon :

— L'assureur *La Française*, de petite taille (2 000 employés et 800m € de chiffre d'affaires), évoluant sur le marché français et un total de 200 000 clients particuliers.

— L'assureur *L'Allemande*, de taille importante (100 000 employés dans le monde dont 15 000 en Allemagne, pour 70mds € de chiffres d'affaires) et un total de 10 millions de clients particuliers.

3.1.1.1 Coût d'une perte de taille standard

Ponemon traite séparément les pertes de données entre 10 000 et 100 000 Données à Caractère Personnel (DCP) (pertes "non-méga") des méga-pertes de données (plus de 1 million de DCP). Ponemon n'étudie pas la zone 100 000 à 1 million de DCP par manque de données issues de leur sondage marché.

Première estimation

Coût total moyen	$3,92$m \$
Taille moyenne	25 575 DCP
Coût par DCP	150 \$

TABLE 3.1 – Coûts moyens total et par DCP (Ponemon 2019)

L'approche la plus naïve reviendrait à estimer le coût moyen par simple lecture ($3.92m$ \$) ou en multipliant le coût moyen par DCP par le nombre moyen de DCP perdues : $25575 \times 150 = 3,84m$ \$. Ce calcul permet de retrouver un montant proche de $3,92m$ \$.

Avec cette première estimation, on modéliserait le même coût pour chacun des assureurs fictifs pris pour exemple.

Deuxième estimation : Prise en compte du pays

Ponemon fournit des données plus précises, par pays ou région : tables 3.2 et 3.3.

États-Unis	242 \$
Allemagne	193 \$
Canada	187 \$
Moyen-Orient	173 \$
France	163 \$
Afrique du Sud	155 \$
Royaume-Uni	155 \$
Corée du Sud	153 \$
Italie	146 \$
Japon	141 \$
Scandinavie	130 \$
ANASE	129 \$
Australie	110 \$
Turquie	95 \$
Inde	72 \$
Brésil	69 \$

TABLE 3.2 – Coût moyen par DCP selon le pays ou région (Ponemon 2019)

Moyen-Orient	38 800
Inde	35 636
États-Unis	32 434
Brésil	26 523
France	26 300
Allemagne	25 610
Italie	24 577
Royaume-Uni	23 636
Corée du Sud	23 600
Canada	23 071
Turquie	22 551
ANASE	22 500
Afrique du Sud	22 060
Scandinavie	21 663
Japon	20 445
Australie	19 800

TABLE 3.3 – Taille moyenne d'une attaque (en nombre de DCP) selon le pays ou région (Ponemon 2019)

Les deux tables 3.2 et 3.3 permettent d'obtenir les chiffres de la table 3.4.

La Française	26 300 $\times 163 = 4,3m$ \$
L'Allemande	25 610 $\times 193 = 4,9m$ \$

TABLE 3.4 – Deuxième estimation du coût moyen des entités fictives

Les chiffres obtenus en deuxième estimation sont plus élevés que le chiffre de la première estimation ($3,84m$ \$ - déduit la table 3.1). Étant donné que seule la variable pays a été utilisée, cela signifie que l'Allemagne et la France présentent des profils de risques plus risqués qu'en moyenne dans le monde, d'après Ponemon.

Troisième estimation : Prise en compte de la taille et du secteur

Ponemon fournit également des coûts moyens par taille d'entreprise et par type d'industrie (table 3.5).

Nombre d'employés	Coût total moyen (millions \$)
< 500	$2,74$
500 à 1 000	$2,65$
1 001 à 5000	$3,63$
5 001 à 10 000	$4,41$
10 001 à 25 000	$4,35$
$> 25\ 000$	$5,11$

TABLE 3.5 – Coût total moyen en fonction de la taille de l'entreprise (Ponemon 2019)

D'après les données Ponemon présentées table 3.5, le coût moyen croît avec la taille de l'entreprise,

exprimée en nombre d'employés. Le rapport entre le coût d'une très grande entreprise (plus de 25 000 employés) et le coût moyen pour une entreprise de moins de 500 employés reste toutefois inférieur à 2 ($5,11m$ \$ contre $2,74m$ \$).

Le coût moyen en fonction de la taille de l'entreprise tel qu'indiqué dans le tableau n'est pas strictement croissant car l'échantillon d'incidents est relativement limité et parce que le coût est très volatil.

Secteur	Coût total moyen (millions \$)
Santé	6,45
Finance	5,86
Énergie	5,60
Industrie	5,20
Pharma	5,20
Technologie	5,05
Éducation	4,77
Services	4,62
Loisirs	4,32
Transport	3,77
Communication	3,45
Consumer	2,59
Media	2,24
Hôtellerie	1,99
Vente au détail	1,84
Recherche	1,65
Secteur public	1,29

TABLE 3.6 – Coût total moyen en fonction du secteur (Ponemon 2019)

D'après la table 3.6, le secteur est une variable qui a un fort impact sur le coût moyen d'une attaque de type perte de données, qui varie de $1,29m$ \$ pour le secteur public à $6,45m$ \$ pour le secteur de la santé.

A partir des tables 3.5 et 3.6, il est possible d'introduire des coefficients d'ajustement pour affiner l'estimation du coût moyen des deux entreprises fictives :
$\alpha_{secteur} = \frac{coût\ moyen_{secteur}}{coût\ moyen\ global}$

D'où l'équation :

$$\alpha_{finance} = \frac{coût\ moyen_{finance}}{coût\ moyen\ global} = \frac{5,86}{3,92} = 1,49 \tag{3.1}$$

De même, pour la taille d'entreprise :

$$\beta_{2\ 000\ employés} = \frac{coût\ moyen_{1k-5k}}{coût\ moyen\ global} = \frac{3,63}{3,92} = 0,93 \tag{3.2}$$

$$\beta_{15\ 000\ employés} = \frac{coût\ moyen_{10k-25k}}{coût\ moyen\ global} = \frac{4,35}{3,92} = 1,11 \tag{3.3}$$

A partir des équations 3.1, 3.2 et 3.3, il est possible de proposer une troisième estimation plus affinée (3.7).

Entreprise	Nombre	Coût unit.	Coût total	Aj. Secteur	Aj. taille	Coût total ajusté
La Française	26 300	163 $	4 286 900 $	1,49	0,93	5 934 381 $
L'Allemande	25 610	193 $	4 942 730 $	1,49	1,11	8 199 392 $

TABLE 3.7 – Troisième estimation du coût à partir de l'étude Ponemon

A partir de l'étude Ponemon, trois estimations extrêmement simples du coût d'une Cyber-attaque de type perte de DCP ont été proposées.

L'étude Ponemon a permis d'identifier plusieurs variables explicatives de la loi de sévérité :
- le pays ou région
- le secteur
- la taille de l'entreprise

En revanche, il existe plusieurs limites à cette approche : l'approche ne considère que le coût moyen des attaques, aucune information concernant la distribution n'est proposée, il n'est pas possible de faire une hypothèse de forme de distribution ni de calibrer un modèle à partir de ces données.

3.1.1.2 Fréquence

Dans son étude, Ponemon reporte l'évolution dans le temps de la fréquence bi-annuelle d'occurence des sinistres pour une entreprise donnée. En effet, Ponemon effectue chaque année une étude de marché. Seuls les sinistres de plus de 10 000 DCP perdues sont utilisés pour l'estimation de cette fréquence.

Date	Probabilité bi-annuelle
2014	22,6%
2015	24,8%
2016	25,6%
2017	27,7%
2018	27,9%
2019	29,6%

TABLE 3.8 – Évolution de la probabilité bi-anuelle de survenance d'un sinistre

D'après les chiffres de l'étude Ponemon, la fréquence d'occurence de sinistres de type perte de données augmente significativement au cours du temps.

Pour l'estimation naïve, l'hypothèse d'une fréquence qui suit une loi de Poisson est raisonnable. Il est possible d'estimer le paramètre de la loi de Poisson annuelle qui permettrait de modéliser la fréquence d'un évènement Cyber de type perte de données.

On note $X_{2\ ans}$ et $X_{1\ an}$ les lois de Poisson représentant la fréquence de l'évènement à horizon 2 ans et 1 an.

On a alors :

$29,6\% = 0,296 = \mathbb{P}(X_{2\ ans} > 0) = 1 - \mathbb{P}(X_{2\ ans} = 0) = 1 - \mathbb{P}(X_{1\ an} = 0)^2 = 1 - e^{-\alpha_{2ans}} = 1 - e^{-2\alpha_{1an}}$

Soit :

$$\alpha_{2ans} = 0,351 \tag{3.4}$$

Et :

$$\alpha_{1an} = 0,175 \tag{3.5}$$

A partir des données Ponemon et sous l'hypothèse d'une fréquence suivant une loi de Poisson, l'espérance est de 0,175 attaque par an.

De plus, sous l'hypothèse d'indépendance entre loi de fréquence et loi de sévérité, l'espérance de perte annuelle sera de :

La Française	$5,9m\$ \times 0,175 = 1,03m\ \$$
L'Allemande	$8,2m\$ \times 0,175 = 1,435m\ \$$

TABLE 3.9 – Estimation du coût annuel moyen des pertes de données "non-méga"

L'étude Ponemon est intéressante car elle permet d'obtenir un premier ordre de grandeur quantitatif du risque auquel est exposé un assureur ou une entreprise quelconque, à la fois le coût moyen en cas d'attaque et le coût moyen annuel.

De plus, elle a l'avantage d'être facilement interprétable car suffisamment simple, et fournit plusieurs variables explicatives.

En revanche, la modélisation qu'il est possible de faire à partir de ce rapport est trop simpliste pour une utilisation dans le cadre du Besoin Global de Solvabilité de l'ORSA car il n'est pas possible d'estimer l'incertitude autour des coûts moyens présentés.

3.1.1.3 Coût d'une méga-perte

Ponemon définit les méga-pertes comme les pertes impliquant au moins 1 million de DCP.

En plus d'une exposition à un sinistre de taille classique, Ponemon introduit le concept de méga-perte de données : un incident concernant plus d'1m de DCP.

Il y a très peu de données disponibles pour ces sinistres et seules 14 entreprises sur 507 en ont expérimenté au cours de l'année d'étude de Ponemon en 2019. Ces 14 entreprises ont permis à Ponemon de déduire la table 3.10. Les chiffres de l'étude de 2018 sont aussi reportés.

Nombre de DCP	Coût moyen total (millions $)	
	2018	2019
1 million	39	42
10 millions	148	163
20 millions	200	225
30 millions	279	308
40 millions	325	345
50 millions	350	388

TABLE 3.10 – Coût moyen total (millions $) selon la taille de la méga-perte de données

Dans le cadre des deux entités fictives sélectionnées, en faisant l'hypothèse que La Française a moins de 1 million de clients, on déduit qu'elle n'est pas exposée à ce type de perte. De même avec une hypothèse raisonnable l'Allemande est exposée au risque de méga-perte mais de 1 million de DCP seulement.

Étant donné que Ponemon ne donne pas d'information sur la fréquence, en estimant une période de retour à 1 fois tous les 1000 ans, une contribution moyenne de l'ordre de 42m*0,1%= 42k $ serait obtenue.

Dans le cadre d'un scénario ORSA, ce type de scénario ne serait probablement pas envisagé car la période de retour est trop faible.

3.1.1.4 Conclusion

Tout d'abord, concernant les **incidents de taille "non-méga"**, 10 000 à 100 000 DCP perdues, il a été vu à partir d'un raisonnement simple qu'il existe trois variables explicatives de **la loi de sévérité** :

— le pays ou région
— le secteur
— la taille de l'entreprise

et il a été obtenu un premier ordre de grandeur quantitatif de **la loi de fréquence**, sous une hypothèse de modélisation naturelle et simple : la loi de Poisson.

Enfin, l'information de l'étude Ponemon concernant les **méga-pertes de données** a été analysée et présentée : seule l'information de la sévérité est pertinente car basée sur un nombre extrêmement limité d'incidents.

Il est déjà possible de modéliser le risque à partir de l'étude Ponemon uniquement : un modèle déterministe, très simple, basé sur des observations moyennes et des ajustements proportionnels adaptés au profil de risque de l'entreprise considérée.

Cependant, la sévérité de ce risque étant extrêmement volatile, nous verrons dans la suite du mémoire avec les travaux de Jacobs qu'il est nécessaire et possible de capter l'incertitude, la volatilité de la loi de sévérité plutôt que d'utiliser une approche basée sur la moyenne.

3.1.2 Jay Jacobs

Jay Jacobs est le nom d'un analyste senior qui exerçait au sein du département des risques de Verizon de New York en 2014 et qui a publié un article qui est rapidement devenu une référence sur le sujet de la modélisation du coût d'une perte de données ainsi que la **quantification de l'incertitude** du coût, en décembre 2014 : [Jacobs, 2014].

Jacobs ne s'intéresse qu'au lien entre nombre de DCP perdues et coût, c'est-à-dire la modélisation du coût sachant le nombre de DCP perdues connu.

3.1.2.1 Premier modèle (2014)

En 2014, Jacobs challenge l'étude Ponemon et remet complètement en question la relation déterministe et proportionnelle entre le coût d'un sinistre et le nombre de DCP perdues.

Il conserve le nombre de DCP comme variable explicative mais essaie des formes de modèles différentes en se basant sur les données utilisées par Ponemon pour les éditions 2013 et 2014. Ces données ne sont pas publiques et n'ont pas pu être récupérées et étudiées dans ce mémoire.

Tout d'abord (figures 3.1 et 3.2), Jacobs enrichit l'approche de Ponemon
$Coût\ total = Coût\ unitaire \times Nombre\ de DCP$
en ajoutant une constante (un *intercept*) au modèle. Il se base sur les données 2013 et les données 2014. Sur ces figures sont représentées l'information telle que présentée sur son blog (capture d'écran de code R, commenté).

Alternative 1: Simple Linear Regression

Since we have the data, we can explore the relationship between the number of records and output from the model for 2013 data.

```
##
## Call:
## lm(formula = total ~ records, data = y3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6725789 -2085298  -828787  1930669 13515451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.33e+06   7.87e+05    2.96   0.0046 **
## records     1.07e+02   2.23e+01    4.78  1.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3330000 on 52 degrees of freedom
## Multiple R-squared:  0.306,  Adjusted R-squared:  0.292
## F-statistic: 22.9 on 1 and 52 DF,  p-value: 1.46e-05
```

There is a lot going on in this output. First the model estimated by the linear regression is:

```
<Losses> = 2,330,000 + $107*<Records>
```

FIGURE 3.1 – Régression linéaire sur les données Ponemon de 2013 (Jay Jacobs)

And 2014:

```
##
## Call:
## lm(formula = total ~ records, data = y4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6383308 -2228903  -938958  2154815 14767865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.86e+06   8.08e+05    3.54  0.00079 ***
## records     1.03e+02   2.29e+01    4.48  3.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3570000 on 59 degrees of freedom
## Multiple R-squared:  0.254,  Adjusted R-squared:  0.242
## F-statistic: 20.1 on 1 and 59 DF,  p-value: 3.43e-05
```

This model is:

```
<Losses> = 2,862,000 + $103*<Records>
```

FIGURE 3.2 – Régression linéaire sur les données Ponemon de 2014 (Jay Jacobs)

Les modèles obtenus sont représentés 3.3. L'étude des résidus étant très peu satisfaisante, Jacobs va proposer une nouvelle forme de modèle.

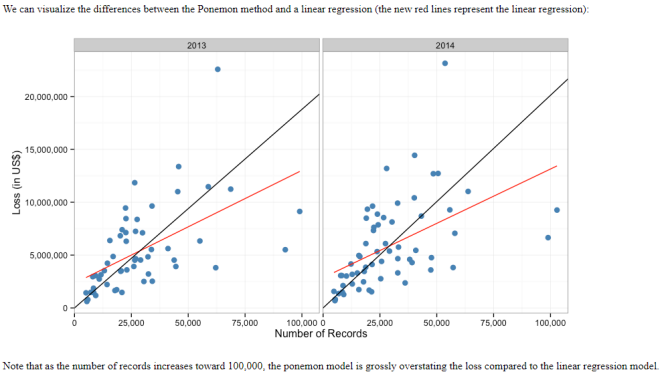


FIGURE 3.3 – Modèle Ponemon vs Régression linéaire avec intercept

Alternative 2: Log-Log Regression

After some trial and error, I found a fairly good model to describe the data, but it's at the expense of simplicity. If we take the `log()` of both the impact and loss prior to modeling and add in a polynomial value as well, we get just about as good of a fit as we will get from this data.

```
##
## Call:
## lm(formula = log(total) ~ log(records), data = ponemon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0243 -0.3792  0.0204  0.4197  1.0188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.6800     0.7013    10.9   <2e-16 ***
## log(records)   0.7584     0.0697    10.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.523 on 113 degrees of freedom
## Multiple R-squared:  0.512,  Adjusted R-squared:  0.508
## F-statistic:  119 on 1 and 113 DF,  p-value: <2e-16
```

Notice the R-squared is now around 50%, which still isn't great, but it's certainly an improvement over the other two models.

This model is:

```
log(impact) = 7.68 + 0.76*log(records)
```

FIGURE 3.4 – Modèle log-log de Jacobs (2014)

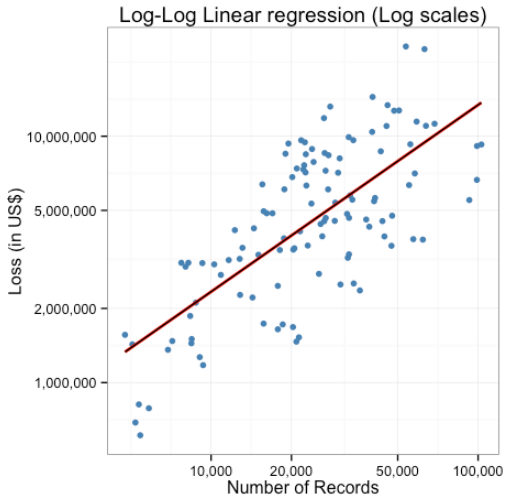


FIGURE 3.5 – Représentation du modèle Jacobs (2014) dans le plan log-log

C'est le modèle log - log qui retenu, c'est-à-dire la régression linéaire du logarithme du coût sur le logarithme du nombre de DCP perdues.

Jay Jacobs conclut de la manière suivante : *"Even though none of the models presented here performed particularly well with this data, we were able to improve on the simplistic method employed by Ponemon. But even with the improved results, it is painfully clear that there are a lot more factors contributing to loss than just a count of records lost. As George Box famously said, "All models are wrong, but some are useful." After looking at this data, I would caution anyone using these models to take them all with a grain of salt. While using something like the log-log model above may be able to provide a frame of reference where there is currently a lot of uncertainty, the amount of variance in*

the model is a serious challenge to adoption."

Jay Jacobs reconnaît donc que son modèle est largement perfectible, mais qu'avec les données à disposition c'est le meilleur qu'il soit possible de proposer. En analysant la figure 3.4 , il y a 113 degrés de liberté. Jay Jacobs a donc utilisé un **échantillon de seulement 115 points** pour calibrer son modèle.

3.1.2.2 Deuxième modèle (2015)

Jay Jacobs a publié une version mise à jour de son modèle quelques mois après avoir publié sa première version. Le modèle mis à jour a été publié dans le rapport annuel de Verizon : [Verizon, 2015].

Dans ce rapport de Jacobs, certains passages proposent des résultats tout à fait différents de l'étude Ponemon ! L'un des enjeux de ce mémoire sera de mener des analyses critiques de chaque étude pour aboutir à une modélisation pertinente.

En particulier, concernant les méga-pertes de données, pour lesquelles d'après le rapport Verizon le coût par DCP peut descendre aussi bas que 1 ou deux centimes de dollar.

Ce n'est pas du tout en ligne avec le rapport Ponemon, figure 3.10, pour qui même pour des méga-pertes de données de plus de 50 millions de DCP, le coût total est de 388m \$ et donc le coût par DCP est proche de 8 \$.

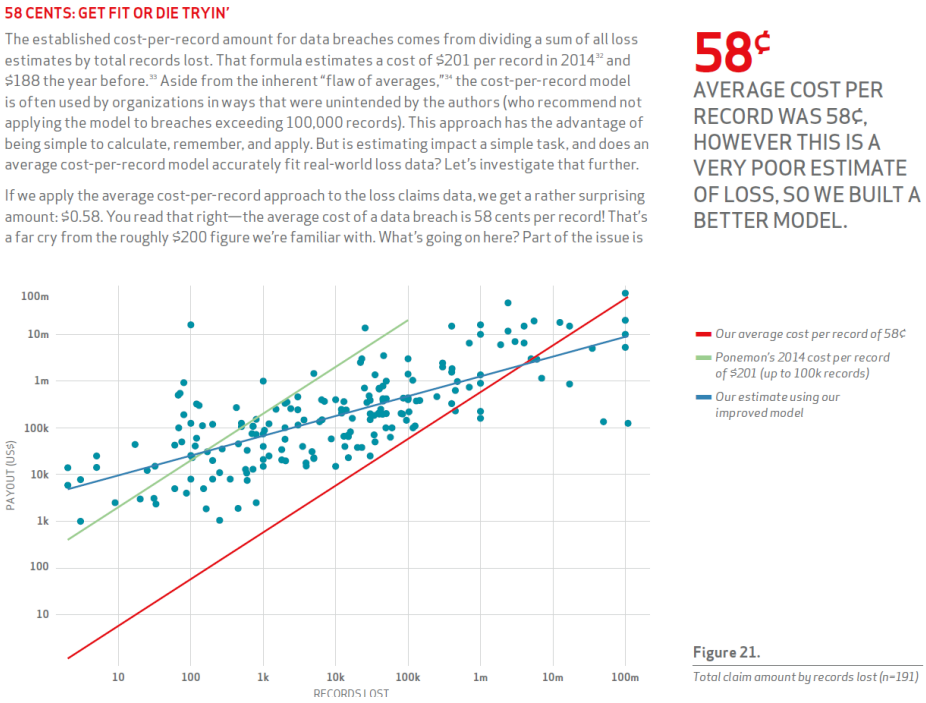


FIGURE 3.6 – Modèle Ponemon vs Régression linéaire avec intercept

Sur la figure 3.6, c'est la droite bleue qui est intéressante : c'est le modèle log-log de Jay Jacobs recalibré, et supposé valide également pour les méga-pertes de données. Pour rappel, Ponemon (courbe

verte) précisait que le domaine de validité de leur modèle était de 10 000 à 100 000 DCP.

En réalité, ce graphique présente une information non seulement partielle mais aussi trompeuse : l'information est partielle car la courbe bleue n'est que le scénario médian du modèle Jay-Jacobs, l'incertitude n'est pas disponible en lecture sur le graphe (on aurait pu imaginer un intervalle de confiance). De plus, le graphe compare des coûts moyens (courbes rouges et courbes vertes) à une représentation de coûts médians. Enfin, la courbe verte est représentée sur l'intervalle de 1 à 100 000 DCP mais elle n'est valide que sur l'intervalle 10 000 à 100 000.

Afin d'avoir à disposition toutes les informations du modèle de Jacobs, il convient d'obtenir les trois paramètres de la régression : la pente et l'intercept de la droite sont disponibles facilement (graphiquement). Pour le paramètre d'incertitude, le tableau représenté en figure 3.7 issu du rapport Verizon permet de le retrouver, ainsi qu'obtenir numériquement la pente et l'intercept.

LET IT GO, LET IT GO

The cold (cost-per-record) figure never bothered us anyway, but we think it's time to turn away and slam the door. To that end, we wrap up this section with a handy lookup table that includes a record count and the single-point prediction that you can use for "just give me a number" requests (the "expected column" in the middle). The rest of the columns show 95% confidence intervals, first for the average loss and then the predicted loss. The average loss should contain the mean loss (if there were multiple incidents). The predicted loss shows the (rather large) estimated range we should expect from any single event.

RECORDS	PREDICTION (LOWER)	AVERAGE (LOWER)	EXPECTED	AVERAGE (UPPER)	PREDICTION (UPPER)
100	$1,170	$18,120	$25,450	$35,730	$555,660
1,000	$3,110	$52,260	$67,480	$87,140	$1,461,730
10,000	$8,280	$143,360	$178,960	$223,400	$3,866,400
100,000	$21,900	$366,500	$474,600	$614,600	$10,283,200
1,000,000	$57,600	$892,400	$1,258,670	$1,775,350	$27,500,090
10,000,000	$150,700	$2,125,900	$3,338,020	$5,241,300	$73,943,950
100,000,000	$392,000	$5,016,200	$8,852,540	$15,622,700	$199,895,100

Figure 23.
Ranges of expected loss by number of records

FIGURE 3.7 – Modèle Ponemon vs Régression linéaire avec intercept

Estimation des paramètres du modèle recalibré

Première étape, pente et intercept :

Il faut reporter les chiffres médian de la figure 3.7 (colonnes records et expected), et prendre leur log.
Ces points sont alignés sur une droite, dont la pente est : $a_{JJ2} = 0,423568$ et l'ordonnée à l'origine est : $b_{JJ2} = 8,194$.

Deuxième étape, volatilité :

Le problème est de déterminer l'erreur standard du paramètre d'incertitude ϵ centré et d'erreur standard σ^2, avec l'hypothèse de normalité des résidus :

$$log(coût) = 8,194 + 0,4236 \times log(nb\ DCP) + \epsilon \tag{3.6}$$

En reportant la différence des logs des colonnes Predicion (lower) et Prediction (Upper), il est

possible de calculer : $F_X(0,975) - F_X(0,025)$

Avec : $X \sim Normale(0\ ;\ \sigma^2)$

Prediction (lower)		Prediction (upper)		$F_X(0,975) - F_X(0,025)$
Value	log	Value	log	
1 170	7,06	555 660	13,23	6,16
3 110	8,04	1 461 730	14,20	6,15
8 280	9,02	3 866 400	15,17	6,15
21 900	9,99	10 283 200	16,15	6,15
57 600	10,96	27 500 090	17,13	6,17
150 700	11,92	73 943 950	18,12	6,20
392 000	12,88	199 895 100	19,11	6,23

TABLE 3.11 – Utilisation des données du rapport Verizon pour estimer le paramètre d'incertitude du modèle de Jacobs

En résolvant l'équation suivante de manière analytique (l'inconnue est σ) :
$F_X(0,975) - F_X(0,025) = 6,17$

La solution est $\sigma = 1,57$.

Cette erreur standard est à mettre en perspective avec celle du premier modèle de Jay Jacobs, qui valait 0,523.

3.1.2.3 Conclusion

L'étude des modèles de Jay Jacobs permet de mettre en évidence la pertinence de la modélisation du coût à partir du nombre de DCP perdues basée sur un modèle de régression linéaire log-log.

Une limite de ces modèles est qu'ils ont été calibrés sur un nombre restreint d'incidents d'une part et que la calibration est devenue ancienne au regard du caractère fortement évolutif du risque. De plus, la comparaison entre les paramètres de volatilité des deux modèles, 0,523 en 2014 contre 1,57 en 2015 met en évidence une très forte révision à la hausse de l'incertitude de la prédiction. Cela n'est pas dû à une évolution du risque en si peu de temps mais à un affinement de la modélisation.

Enfin, Jay Jacobs ne s'intéresse pas à la modélisation du nombre de DCP perdues, essentielle dans le cadre de la problématique de ce mémoire.

3.1.3 Analyse croisée des études Ponemon et Jacobs

Tout d'abord, les études de Jacobs et de Ponemon ne sont pas souvent cohérentes : Jacobs ayant considéré que l'approche "en moyenne" de Ponemon n'était pas du tout représentative du risque et ayant proposé une forme de modèle de type régression linéaire log-log et des calibrations de l'incertitude dans ses modèles pour pallier la faiblesse de l'étude Ponemon et la spécificité de ce risque. Ce risque étant hyper volatil, il n'est pas adapté de proposer des analyses basées uniquement sur des observations

moyennes.

Ensuite, la forme log-log induit une modélisation log-normale du coût à nombre de DCP perdues fixé. Jacobs estimait le paramètre σ de loi log-normale à 0,523 en 2014 et 1,57 en 2015, c'est-à-dire à nombre de DCP fixé un rapport moyenne sur médiane de 1,69 en 2014 contre 4,8 en 2015.

Le premier enjeu de la suite du mémoire sera de choisir entre une modélisation directe du coût ou une modélisation séquentielle en deux étapes : modélisation du nombre de DCP puis modélisation du coût.

3.2 Lien direct entre le coût et les variables explicatives

L'enjeu de cette section est d'étudier la faisabilité et la pertinence d'une modélisation directe du coût, en fonction des caractéristiques propres à une entreprise, par opposition à la modélisation Jay Jacobs qui modélise le coût à partir d'une variable intermédiaire : le nombre de DCP perdues.

La base VERIS contient l'information du coût de sinistres pour 302 incidents.

Dans la section 3.1.1, trois variables explicatives ont été mises en évidence : le pays ou région, la taille de l'entreprise et le secteur d'activité.

Ces variables explicatives sont présentes dans la base VERIS, cependant étant donné que la majorité des DCP concernent les États-Unis, il n'est pas possible de discriminer les données en se basant sur le pays ou région : dans cette section, les variables explicatives de la taille de l'entreprise et du secteur seront testées.

3.2.1 Relation entre coût et taille d'entreprise

Tout d'abord, la variable Size2 est considérée car elle est plus fine que Size1. D'après l'étude de la base, il a néanmoins été observée que Size2 était moins fiable que Size1. Size2 est composée d'uniquement 8 modalités (des intervalles) mais représente une variable quantitative.

En première approche, il est possible de s'inspirer de l'approche de Jacobs, en proposant une régression linéaire dans le plan log-log.

Une transformation des modalités est effectuée de telle sorte à obtenir une variable quantitative : la moyenne du logarithme en base 10 des bornes de chaque intervalle est retenue. Pour l'intervalle « 100 000 employés et plus », une borne supérieure 500 000 employés est choisie, correspondant à l'ordre de grandeur des plus grandes entreprises du monde en terme de nombre d'employés.

Size1	Size2	log10(Size2)
Small	A.1.10	0,50
Small	B.11.100	1,50
Small	C.101.1000	2,50
Large	D.1001.10k	3,50
Large	E.10k.25k	4,20
Large	F.25k.50k	4,55
Large	G.50k.100k	4,85
Large	H.100k+	5,35

TABLE 3.12 – Transformation de Size2 en variable quantitative

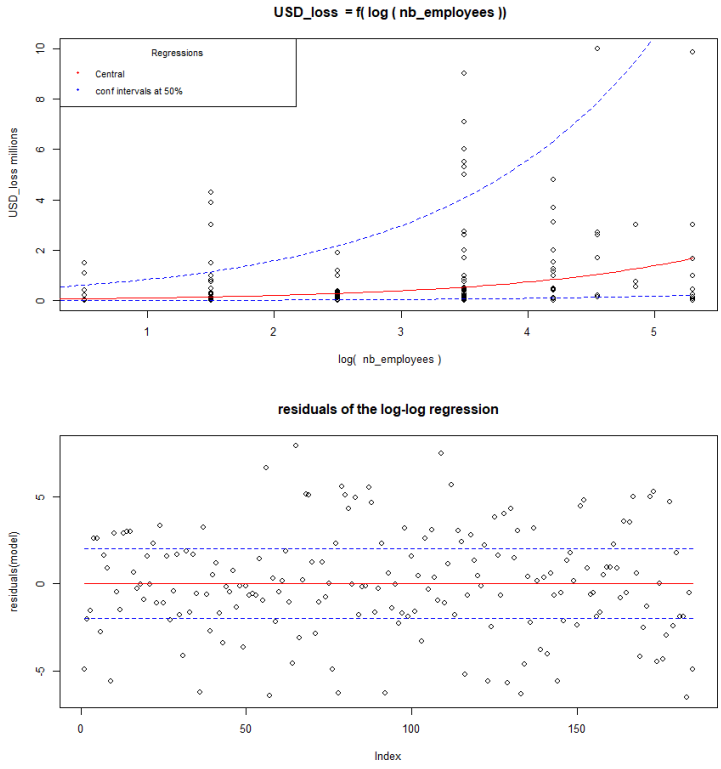


FIGURE 3.8 – Lien entre le nombre d'employés et le coût - en bleu les bandes de confiance à 95%, en rouge la médiane

On positionne les données dans le plan $log(coût) \times log10(Size2)$, ce qui permet d'observer une légère tendance positive, statistiquement significative et en ligne avec l'étude de Ponemon.

	Régression log-log du coût sur *Size2*
Intercept	$11,01^{***}$
Erreur standard	$(0,55)$
Coef Size2	$0,63^{***}$
Erreur standard	$(0,16)$
R^2	0,08
R^2 ajusté	0,07
Nombre obs.	185
RMSE	3,00

$^{***}p < 0,001$, $^{**}p < 0,01$, $^{*}p < 0,05$

TABLE 3.13 – Régression linéaire du coût sur le nombre d'employés

La régression linéaire du log du coût sur le $log(size2)$ n'est cependant pas pertinente, les points au sein de chaque catégorie étant trop dispersés. Cela se traduit par un faible coefficient R^2, proche de zéro.

En conclusion, la taille de l'entreprise semble être une bonne variable explicative car les paramètres *Intercept* et *Size2* ont des p-values suffisamment proches de zéro, cependant le modèle étudié dans cette section, la régression linéaire dans le plan log-log pour expliquer le coût directement, n'est pas adapté car le pouvoir de prédiction (représenté par le coefficient R^2 est particulièrement faible).

3.2.2 Relation entre coût et secteur d'activité

L'étude de la relation entre le coût et le secteur d'activité permet également de souligner le caractère explicatif du secteur d'activité mais il n'est pas possible de proposer de modélisation suffisamment prédictive, de manière analogue à la relation entre coût et taille d'entreprise.

3.2.3 Conclusion

Les études de la relation directe entre le coût et les variables explicatives permettent de confirmer le caractère significativement explicatif de la taille et du secteur d'activité. Cependant, l'utilisation directe de ces variables n'est pas pertinente car elle implique une modélisation grossière du coût car chaque variable explicative n'est pas quantitative et ne possède qu'un nombre limité de modalités. Le fait qu'il n'existe qu'un nombre relativement modeste d'incidents dans la base ne permet pas de capter les spécificités de chaque modalité.

Enfin, le pouvoir prédictif des modèles construits avec la même structure que le modèle de Jacobs est très inférieur au pouvoir prédictif du modèle de Jacobs.

Une modélisation séquentielle en deux étapes semble donc préférable et sera étudiée dans la section suivante.

3.3 Modélisation du coût en deux étapes

3.3.1 Modélisation du lien entre le coût et le nombre de DCP perdues

Dans cette section seront confrontés les modèles ayant déjà été publiés dans la littérature aux données de la base VERIS : modèle de Jacobs et modèle de Farkas. De plus, il est possible de proposer un backtest issu des chiffres Ponemon.

Il a déjà été mis en évidence que la relation proportionelle entre le nombre de DCP perdues et le coût (Ponemon) n'était pas pertinente.

3.3.1.1 Structure Jay Jacobs

Tout d'abord, la base VERIS est mise à contribution pour recalibrer le modèle Jay Jacobs, sur les 145 points disponibles dans la base (figure 3.9).

Dans toutes les figures de cette section, les **bandes bleues** en pointillés sont les **bandes de confiance à 95%** (sauf dans certains où cela est explicité dans la légende ce sont les bandes de confiance à 50%) des différents modèles, et les **courbes rouges** représentent **la médiane**.

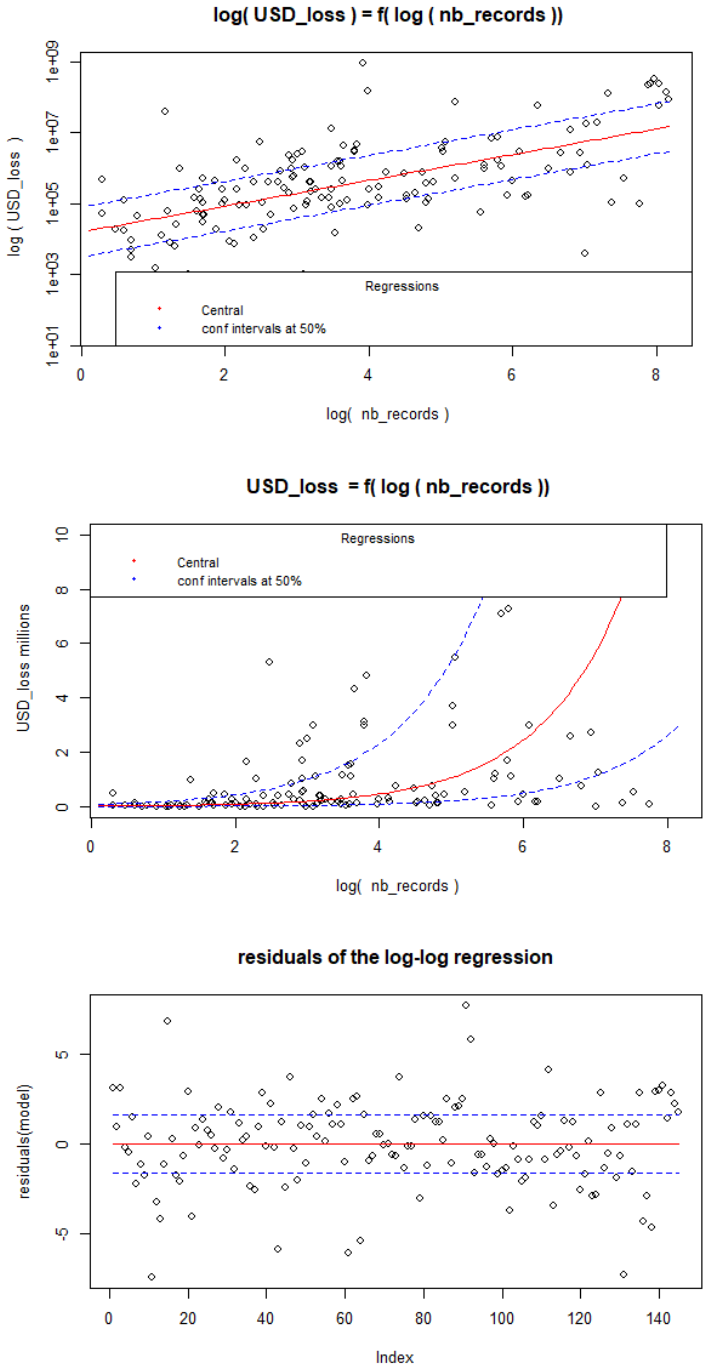


FIGURE 3.9 – Calibration du modèle de Jacobs à partir de la base VERIS

	Régression log-log du coût sur le nombre de DCP
Intercept	$9,67^{***}$
Erreur standard	$(0,41)$
Coef. nb de DCP	$0,84^{***}$
Erreur standard	$(0,10)$
R^2	0,35
R^2 ajusté	0,34
Nombre obs.	145
RMSE	2,41

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.14 – Paramètres de la régression linéaire log-log du coût sur le nombre de DCP

Utiliser le nombre de données perdues plutôt que la taille d'entreprise permet d'obtenir un modèle avec un pouvoir prédictif beaucoup plus important (le R^2 est nettement plus élevé).

Graphiquement (figure 3.9), le modèle ne semble pas bien fonctionner pour les très hautes valeurs. Plus précisément, le modèle recalibré sous-estime le coût des très grosses pertes de données.
C'est cohérent avec les remarques de Ponemon et de Jay Jacobs relatives au pertes de plus d'1 million de DCP.

Si ce modèle n'est pas pertinent pour les méga pertes de données, il faut le recalibrer en excluant les points correspondant à de très grosses pertes de données.

Le modèle est recalibré après exclusion des quelques points qui se comportent mal. Le seuil de 10 millions de DCP est retenu plutôt que l'exclusion des points particuliers afin d'éviter le *cherry picking*. Cela correspond à l'exclusion de 15 points.

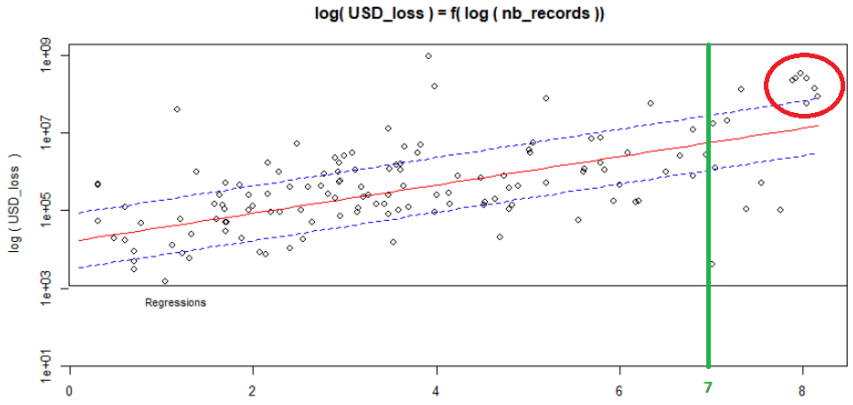


FIGURE 3.10 – Points problématiques et insertion d'un seuil

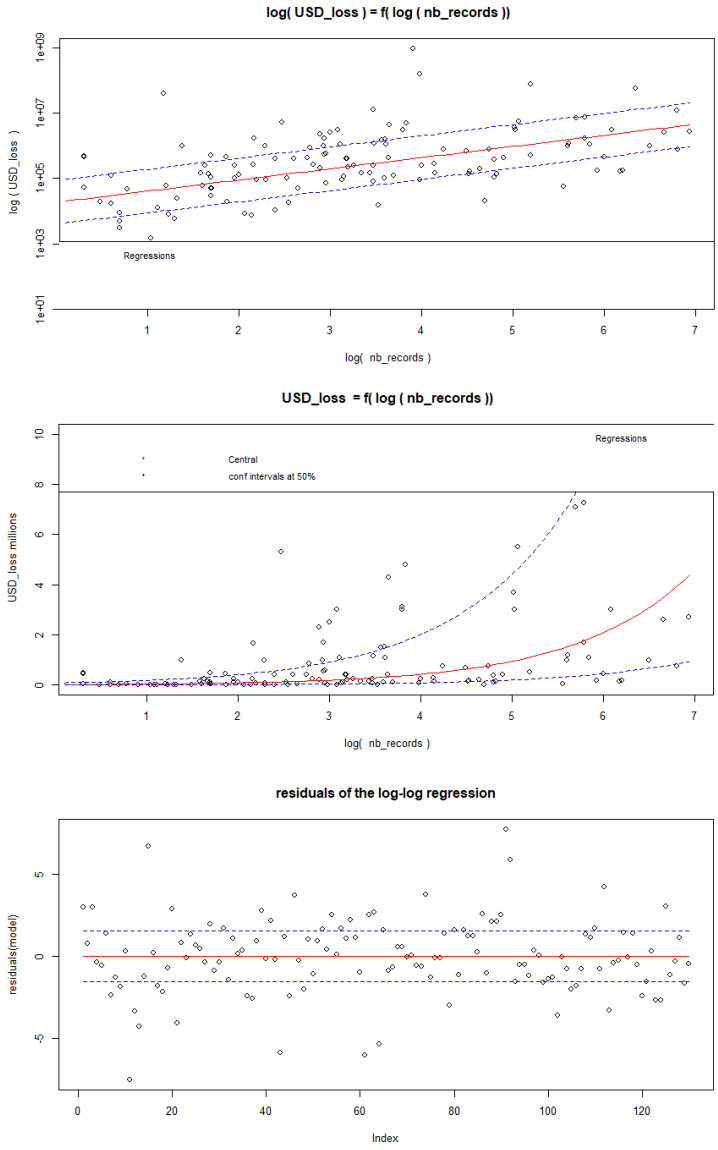


FIGURE 3.11 – Calibration de la structure de Jacobs, jusqu'à 10m DCP

	Même modèle mais limité à 10 millions de DCP
Intercept	$9,83^{***}$
Erreur standard	$(0,44)$
Coef. nb de DCP	$0,79^{***}$
Erreur standard	$(0,12)$
R^2	0,25
R^2 ajusté	0,25
Nombre obs.	130
RMSE	2,29

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.15 – Paramètres du modèle, jusqu'à 10 millions de DCP

Graphiquement et statistiquement, les résidus se comportent de manière satisfaisante.

Les résidus sont leptokurtiques, c'est-à-dire avec des extrêmes plus dispersés qu'une loi normale de même écart-type, mais le choix d'une modélisation par une loi normale plutôt qu'avec une loi de Student sera retenue pour être homogène avec le modèle de Jay Jacobs.

	Observation des résidus
Nombre obs.	130
Minimum	-7,55
Maximum	7,75
1. Quartile	-1,22
3. Quartile	1,34
Moyenne	-0,00
Médiane	-0,08
Somme	-0,00
Variance	5,19
Ecart-type	2,28
Coef. asymétrie	-0,01
Kurtosis	1,71

TABLE 3.16 – Statistiques concernant les résidus

Avec les données VERIS, il est possible de calibrer le modèle de Jay Jacobs de telle sorte à obtenir un modèle pertinent jusqu'à 10 millions de DCP.
Ainsi, la limite entre méga-pertes et pertes "non-méga" est différente de celle retenue par Ponemon. Dans son étude Ponemon appelle méga-perte de données les pertes de plus d'1 million de DCP.

3.3.1.2 Structure Farkas

Dans le papier [Farkas *et al.*, 2019], Farkas propose un autre modèle pour pallier le mauvais comportement du modèle de Jay Jacobs initial sur les méga-pertes de données.

Jacobs :

$$log(coût) = \alpha + \beta \times log(nb\ DCP.) + \epsilon \tag{3.7}$$

Farkas :

$$log(coût) = \alpha + \beta \times log(log(nb\ DCP.)) + \epsilon \tag{3.8}$$

Farkas calibre son modèle à partir de 4 points : deux points déduits du modèle Jay Jacobs, les bornes de l'intervalle où il est réputé correct (10 000 et 100 000 DCP), et deux points représentant le coût moyen de méga-perte de données récentes, issues du rapport Ponemon de 2018. Il n'a donc pas été possible de proposer une valeur pour le paramètre d'incertitude.

Calibration du modèle de Farkas dans [Farkas *et al.*, 2019] :

$$log(coût) = -1,998 + 7,503 \times log(log(nb\ DCP)) \tag{3.9}$$

Il est possible d'utiliser la base de données disponible et calibrer le modèle de Farkas dessus (figures 3.12 et 3.17).

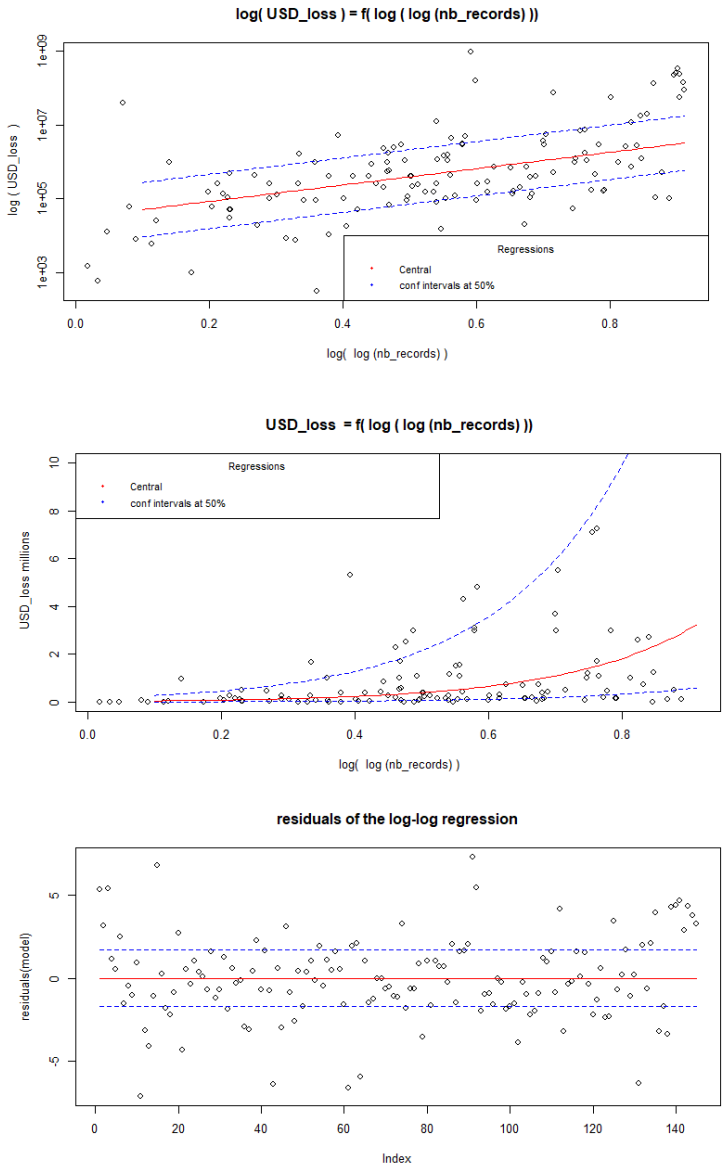


FIGURE 3.12 – Modèle de Farkas calibré sur les données VERIS

	Calibration du modèle de Farkas
Intercept	$10,33^{***}$
Erreur standard	$(0,38)$
Coef. nb de DCP	$5,1^{***}$
Erreur standard	$(0,67)$
R^2	0,29
R^2 ajusté	0,28
Nombre obs.	145
RMSE	2,52

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.17 – Paramètres du modèle de Farkas calibré sur les données VERIS

Ainsi (en considérant la pente, 5,1, ajustée du facteur 1/log(10), pour des raisons de base de logarithme) :

Calibration du modèle de Farkas avec VERIS :

$$log(coût) = 10,33 + 2,22 \times log(log(nb\ DCP)) + \epsilon, \quad \epsilon \sim Normale(\ 0\ ;\ 2,52) \tag{3.10}$$

Il est ainsi possible de représenter les valeurs médianes des différents modèles en fonction du nombre de données perdues, dans le plan log – log (figure 3.13).

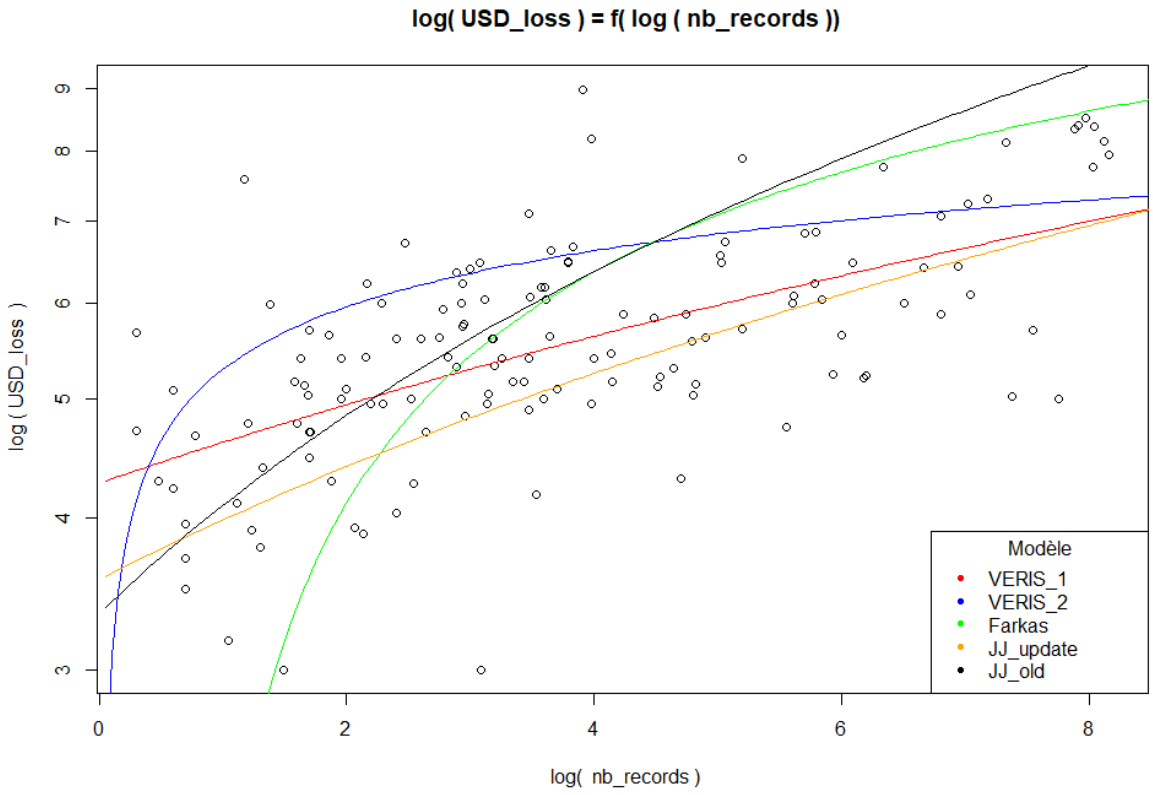


FIGURE 3.13 – Valeurs médianes des différents modèles dans le plan log-log (les deux axes en base 10)

Les données VERIS permettent d'obtenir un modèle assez proche du modèle de Jay Jacobs, notamment pour les pertes supérieures à 10 000 DCP, qui est la borne retenue par Ponemon dans son étude.

De plus, la structure du modèle de Farkas ne permet pas de capturer correctement la relation entre le coût et le nombre de DCP perdues.

3.3.1.3 Backtest avec Ponemon

Le *backtest* entre les chiffres moyens de marché recueillis proposés par Ponemon et ceux que l'on obtient avec les différents modèles proposés sont reportés dans cette section.

La différentiation entre la médiane et la moyenne n'est possible que lorsqu'on a à disposition le paramètre de volatilité du modèle.

Données Ponemon disponibles

L'étude Ponemon 2019 permet de lire les données suivantes, qu'il est possible d'ajuster pour obtenir des valeurs approchées aux bornes :

Nombre de DCP	Coût moyen total FY19 (milliers de $)
Moins de 10 000	2 200
10 000 à 25 000	3 300
25 000 à 50 000	4 700
50 000 à 100 000	6 400
1 000 000	42 000
10 000 000	163 000
20 000 000	225 000
30 000 000	309 000
40 000 000	345 000
50 000 000	388 000

TABLE 3.18 – Données issues du rapport Ponemon 2019

Nombre de DCP	Ponemon)
	Coût (milliers de $)
10 000	3 000
25 000	4 000
50 000	5 550
100 000	7 250
1 000 000	42 000
10 000 000	163 000
20 000 000	225 000
30 000 000	309 000
40 000 000	345 000
50 000 000	388 000

TABLE 3.19 – Données approchées, Ponemon 2019

Le modèle de type log – log(log()) (VERIS (Farkas)) ne semble pas adapté ici car la moyenne obtenue est tout à fait incohérente notamment pour les incidents de moins de 1 million de DCP perdues. Par exemple, ce modèle prédit un coût moyen de 100 millions $ pour une perte de 10 000 DCP.
A titre de comparaison, Ponemon prédit un coût moyen de seulement 42 millions $ pour les méga-pertes de 1 million de DCP.

3.3.1.4 Méga - pertes de données

Une approche de coût par DCP est pertinente pour les méga-pertes de données et est plus intuitive qu'une approche en coût global pour de gros montants.

Données Ponemon

Dans l'étude de 2019 dans laquelle les chiffres de coûts moyens de 42 millions $ et 388 millions $ pour des pertes de respectivement 1 et 50 millions DCP sont présentés, Ponemon précise que ces chiffres n'ont été calibrés que sur 14 entreprises ayant rencontré ce type de pertes, sur un historique de 3 ans.

Dans l'étude 2018, Ponemon est plus précis concernant le coût par DCP d'une perte de donnée. Ces chiffres sont en ligne avec les valeurs approchées de 2019. Les valeurs exactes de 2018 seront utilisées dans la suite de l'étude.

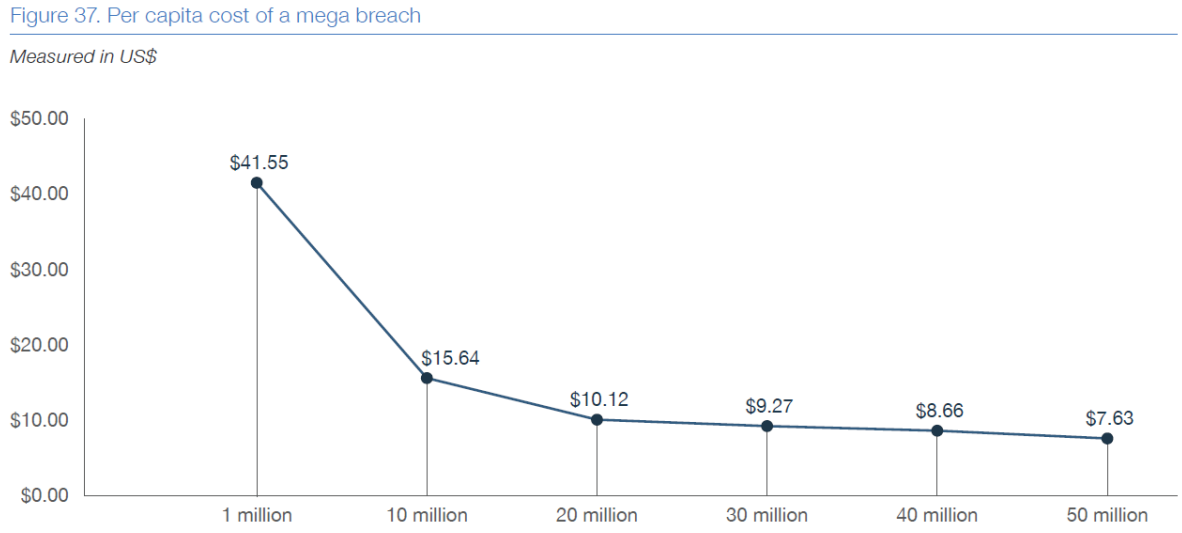


FIGURE 3.14 – Coût par DCP des méga pertes de données - Ponemon 2018

Données VERIS

La base VERIS contient 24 incidents sur la plage 1 à 150 millions de DCP perdus.
Représentation graphique

Nb de DCP	Ponemon	JJ 2014		JJ 2015		VERIS (JJ)		VERIS (Farkas)		Farkas
	Moy.	Moy.	Méd.	Moy.	Méd.	Moy.	Méd.	Moy.	Méd.	
10 000	3 000	2 721	2 373	618	179	5 912	432	100 803	4 212	2329
25 000	4 000	5 460	4 762	911	264	8 085	591	124 371	5 197	4 747
50 000	5 550	9 247	8 065	1221	354	10 244	749	144 017	6 018	7 799
100 000	7 250	15 659	13 658	1 638	475	12 981	950	165 255	6 905	12 427
1 000 000	42 000			4 344	1 259	28 502	2 085	247 494	10 342	48 804
10 000 000	163 000			11 520	3 338	62 582	4 578	348 234	14 551	155 151
20 000 000	225 000			15 452	4 477	79 300	5 801	382 279	15 974	212 792
30 000 000	309 000			18 347	5 316	91 079	6 663	403 004	16 840	254 456
40 000 000	345 000			20 724	6 005	100 484	7 351	418 074	17 470	288 145
50 000 000	388 000			22 779	6 600	108 442	7 933	429 972	17 967	316 875
100m				30 552	8 8852	137 411	10 052	468 102	19 560	422 540
200m				40 977	11 873	174 118	12 737	508 017	21 228	557 482
500m				60 408	17 503	238 105	17 418	563 545	23 548	792 157
1 milliard				81 022	23 476	301 712	22 072	607 657	25 391	1 022 505

TABLE 3.20 – Backtest entre les données Ponemon et les différents modèles (montants en milliers de $

id_incident	Nom_entreprise	DCP (10^6)	Coût (10^6)	Coût unit.
1948	AvMed, Inc.	1,22	3,00	2,46
6000	T-Mobile	1,50	0,16	0,11
6698		1,60	0,17	0,11
506	21st Century Oncology	2,20	57,00	25,91
8099	Hitachi Payments Services	3,20	1,00	0,31
5423	Scottrade	4,60	2,60	0,57
4580	TD Ameritrade Holding Corp	6,30	12,00	1,90
1934	LinkedIn	6,40	0,75	0,12
907	KT Corporation	8,70	2,73	0,31
2989	Interpark Corp	10,30	0,00	0,00
4798	Excellus Bluecross BlueShield	10,50	17,30	1,65
658	Nationwide Building Society	11,00	1,25	0,11
81	Experian Plc	15,00	20,00	1,33
7020	Office of Personnel Managmt	21,50	133,30	6,20
7468	Zappos	24,00	0,11	0,00
2997	SK Communications Co,, Ltd,	35,00	0,51	0,01
38	Uber	57,00	0,10	0,00
5265	Sony Online Entertainment	77,00	218,28	2,83
7886	JPMorgan Chase & Co	83,00	250,00	3,01
906	TJX Cos Inc	94,00	333,70	3,55
4892	The Home Depot	109,00	57,40	0,53
2470	Target Corporation	110,00	243,00	2,21
2384	Heartland Payment Systems	134,00	140,00	1,04
4782	Equifax	146,69	87,50	0,60

TABLE 3.21 – Données disponibles dans la base VERIS - de 1 à 150 millions DCP perdues

Nb de DCP	Ponemon		JJ 2015		VERIS (JJ)	
	Moy.	Moy./DCP	Moy.	Moy./DCP	Moy.	Moy./DCP
10 000	3 000	300	618	62	5 912	591
25 000	4 000	160	911	36	8 085	323
50 000	5 550	111	1221	24	10 244	205
100 000	7 250	73	1 638	16	12 981	130
1 000 000	42 000	42	4 344	4	28 502	29
10 000 000	163 000	16	11 520	1,2	62 582	6,3
20 000 000	225 000	11	15 452	0,8	79 300	4,0
30 000 000	309 000	10	18 347	0,6	91 079	3,0
40 000 000	345 000	9	20 724	0,5	100 484	2,5
50 000 000	388 000	8	22 779	0,5	108 442	2,2
100m			30 552	0,3	137 411	1,4
200m			40 977	0,2	174 118	0,9
500m			60 408	0,1	238 105	0,5
1 milliard			81 022	0,1	301 712	0,3

TABLE 3.22 – Coût moyen par DCP selon les études et les modèles disponibles - (milliers de $ et $ resp. pour la moyenne et moyenne / DCP)

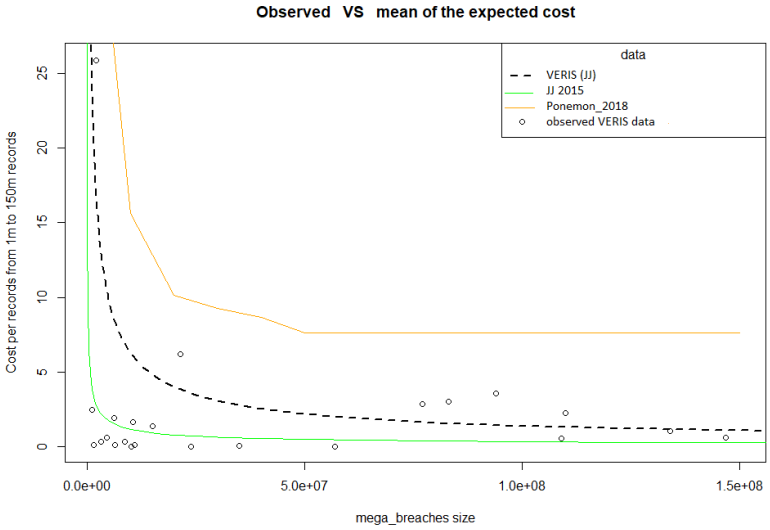


FIGURE 3.15 – Représentation des coûts moyens par DCP dans le plan coût x nombre de DCP, de 1 à 150m DCP

D'après la table 3.25, les coûts moyens reportés par Ponemon sont significativement plus élevés que les incidents reportés dans la base VERIS, et même complètement "hors sol", en effet même le sinistre de plus de 10 millions de DCP perdus ayant coûté le plus cher par DCP ne dépasse pas le coût par DCP de 8 $ sur lequel communique Ponemon.

Le modèle VERIS (JJ), calibré sur les données jusqu'à 10 millions de DCP sous-estime le coût des méga-pertes de données, mais moins que le modèle de Jacobs de 2015.

Données disponibles dans les médias

Les méga-pertes de données engendrant par nature plus de conséquences que les pertes de moins de 1 millions de DCP, elles sont plus susceptibles d'apparaître dans les médias.

L'information reste cependant difficile à obtenir, certains journalistes reportant même des chiffres inventés, ou plutôt reconstitués à partir d'informations mal interprétées, mais cela est aussi lié au fait que le coût est composé de plusieurs composantes distinctes.

Dans les médias sont souvent reportés une seule de ses composantes, par exemple le coût d'une procédure judiciaire qui est facile à obtenir. Les coûts liés aux pertes d'exploitation sont plus difficiles à estimer.

Entreprise	Date	Nb d'enregistrements	Coût total	Coût par enregistrement	Méthode d'estimation	Source
Yahoo	2014	3 milliards	350 millions	$0,12	Dans le cadre d'une offre de rachat par Verizon, le prix proposé a baissé de 350m (de 4,83mds à 4,48mds)	https://blog.primefactors.com/famous-data-breaches-what-they-cost
FriendFinder	2016	412,2 millions	239 millions	$0,58	Le journaliste s'est contenté d'utiliser l'estimation de 0,58	https://blog.primefactors.com/famous-data-breaches-what-they-cost
Linkedin	2012	165 millions	96 millions	$0,58	Le journaliste s'est contenté d'utiliser l'estimation de 0,58	https://blog.primefactors.com/famous-data-breaches-what-they-cost
Heartland Payment Systems	2009	130 millions	75 millions	$0,58	Le journaliste s'est contenté d'utiliser l'estimation de 0,58	https://blog.primefactors.com/famous-data-breaches-what-they-cost
Target Stores	2013	110 millions	64 millions	$0,58	Le journaliste s'est contenté d'utiliser l'estimation de 0,58	https://blog.primefactors.com/famous-data-breaches-what-they-cost
MySpace	2016	360 millions	209 millions	$0,58	Le journaliste s'est contenté d'utiliser l'estimation de 0,58	https://blog.primefactors.com/famous-data-breaches-what-they-cost
Heartland Payment Systems	2009	134 millions	140 millions	$1,04	Lecture dans la base VERIS - ne semble concerner que les compensations liées aux paiements frauduleux	VERIS https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html
Equifax	2017	143 millions	87,5 millions	$0,61	Lecture dans la base VERIS	VERIS https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html
Uber	2016	57 millions	100k (VERIS) à 148 millions (Reuters)	$0,00 à $2,6	Lecture dans la base VERIS et dans les médias	VERIS https://www.reuters.com/article/us-uber-databreach/uber-settles-for-148-million-with-50-us-states-over-2016-data-breach-idUSKCN1M62AJ

TABLE 3.23 – Recherche dans les médias de coûts associés aux méga-pertes de données sur quelques exemples

3.3.1.5 Synthèse et choix d'un modèle de coût

log - log	param1	param2	résidus	σ	Plage de validité
JJ 2015	8,194	0,424	normal	1,574	1 à 100m
JJ 2014	7,680	0,760	normal	0,523	1 à 100k
VERIS fit (1)	**9,831**	**0,342**	**normal**	**2,287**	**1 à 10m**
Ponemon fit	7,662	0,573	normal	2,000	10k à 1 milliard

TABLE 3.24 – Récapitulatif des différents modèles utilisés - Structure Jacobs

log - log(log)	param1	param2	résidus	σ	Plage de validité
Farkas	- 1,994	7,503	Non estimé	Non estimé	10k à 1 milliard
VERIS fit (2)	10,335	2,215	normal	2,52	1 à 10m

TABLE 3.25 – Récapitulatif des différents modèles utilisés - Structure Farkas

Parmi ces modèles, il a été montré que le modèle de la forme log-log calibré sur les données VERIS (**VERIS fit (1)**) est le meilleur modèle disponible, et valide sur la plage de [1 ; 10 millions] DCP. L'un des avantages de ce modèle est que les données sur lequel il a été calibré sont disponibles. De plus, le paramètre d'incertitude est plus élevé que celui du modèle calibré en 2015 par Jacobs, ce qui est plus conservateur. Dans le cadre du risque Cyber de perte de données, il est plus raisonnable de retenir des hypothèses conservatives lorsqu'on est amené à calibrer un modèle sur un historique de données étant donné la nature évolutive du risque : l'hypothèse selon laquelle le passé représente bien le futur est assez forte, c'est une spécificité de ce risque.

Des limites existent pour la modélisation du coût des incidents mettant en jeu plus de 10 millions de DCP, caractérisées par un manque de volume de données et un manque de fiabilité de ces données. Dans le cadre de l'ORSA, la modélisation d'incidents mettant en jeu jusqu'à 10 millions de DCP est supposée suffisante : peu d'assureurs stockent des données de plus de 10 millions d'assurés dans un même système informatique. Dans le cadre de la problématique du mémoire, il n'est pas nécessaire de modéliser spécifiquement ces très grosses pertes de données au delà de 10 millions de DCP, dont les limites de la modélisation ont été rappelées.

De plus, les modèles de type log – log(log()) ne semblent pas adaptés pour modéliser la relation entre le nombre de DCP et le coût en prenant en compte un paramètre de volatilité.

3.3.2 Modélisation du nombre de DCP perdues

La mise à disposition de deux bases de données contenant la variable *nombre de DCP* permet une comparaison entre les données des deux bases.

Dans une première section, il sera étudié sur la base VERIS la meilleure forme de modèle à retenir. Il sera essayé de modéliser différemment le nombre de données perdues en fonction de la taille de l'entreprise et du secteur.
Une fois un modèle retenu, il sera recalibré sur les données PRC et les résultats comparés.

Dans le cadre de l'ORSA, la modélisation du nombre de DCP perdues n'est pas essentielle car l'exercice peut être réalisé à partir de jugement d'expert, cependant ce modèle est important pour le modèle de tarification.

[Bessy-Roland et Boumezoued, 2019] proposent une modélisation du logarithme du nombre de données perdues en se basant sur la base PRC.

3.3.2.1 Nombre de DCP en fonction de la taille - approche log-log

Étude de tendance avec la variable Size2

En première approche, la classification la plus précise selon la taille (la variable Size2) est utilisée, et la régression linéaire du $log(nbDCP)$ sur le $log(Size2)$ est effectuée, de telle sorte à avoir une première idée de l'existence d'une tendance, de la pertinence de cette manière de modéliser cette relation.

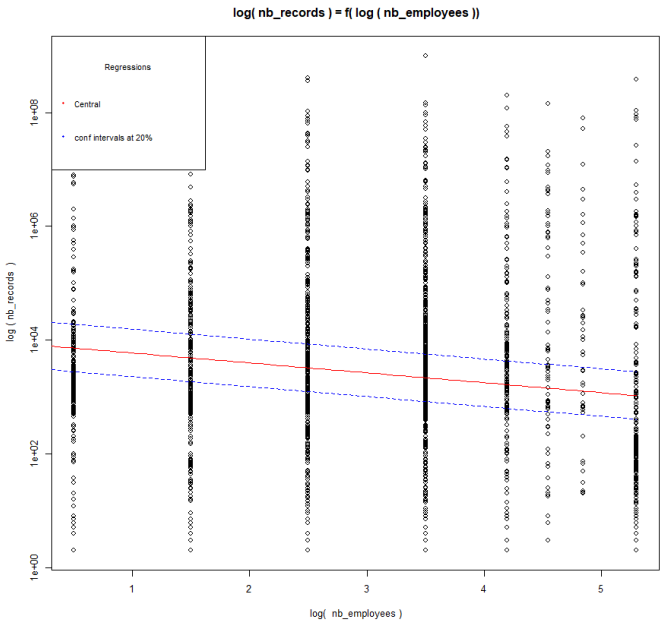


FIGURE 3.16 – Régression linéaire du $log(nb\ DCP)$ sur le $log(Size2)$

	Calibration sur *log(Size2)*
Intercept	$9,09^{***}$
Erreur standard	$(0,18)$
Coef. nb de DCP	$-0,40^{***}$
Erreur standard	$(0,05)$
R^2	0,024
R^2 ajusté	0,023
Nombre obs.	2339
RMSE	3,798

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.26 – Détails quantitatifs

Cette régression linéaire a une pente négative, ce qui signifie que plus l'entreprise est grande et mois elle a tendance à perdre de données lors d'un incident. Ce n'est pas conforme aux observations de marché.

De plus, les résidus (non représentés ici) ne suivent pas une loi normale.

Le pouvoir de prédiction du modèle est très faible : proche de zéro avec un R^2 de 2% environ.

Cette étude indique que le nombre d'employés (fortement lié à la taille de l'entreprise) n'apporte pas d'information utile pour la caractérisation du nombre de données perdues ou au moins que le modèle log-log n'est pas adapté ici.

De plus, le choix retenu de classifier les incidents de la base VERIS selon le nombre d'employés à la granularité la plus fine est aussi une hypothèse forte et trop ambitieuse, les classes étant trop fines par rapport au nombre de données et à leur variabilité, ce qui peut expliquer la pente négative. C'est aussi visible graphiquement sur la figure 3.16 : les classes élevées (à partir de la cinquième) contiennent moins de données que les classes 1 à 4, ceci étant très marqué pour les classes 6 et 7.

Dès à présent, une autre approche avec moins de classes peut être envisagée.

Étude avec la variable Size1

La variable *Size1* trie les incidents en deux classes : en-deça de 1 000 employés et au-delà de 1 000 employés.

Il est possible de calibrer le même modèle que précédemment (3.3.2.1) en remplaçant *Size2* par *Size1*.

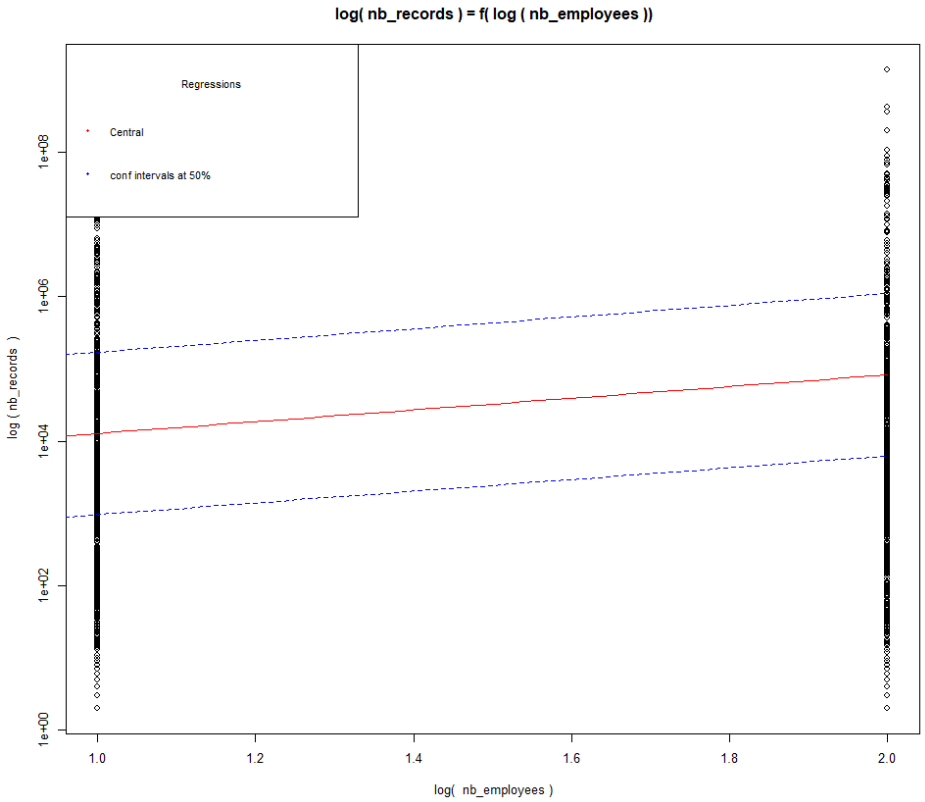


FIGURE 3.17 – Régression linéaire du $log(nbDCP)$ sur le $log(Size1)$

	Calibration sur *log(Size1)*
Intercept	$7,59^{***}$
Erreur standard	$(0,10)$
Coef. nb de DCP	$1,87^{***}$
Erreur standard	$(0,50)$
R^2	0,005
R^2 ajusté	0,005
Nombre obs.	2602
RMSE	3,837

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.27 – Détails quantitatifs

Cette régression linéaire a une pente positive ce qui est cohérent avec les observations du marché.

Cependant, le pouvoir de prédiction est extrêmement faible, avec un un R^2 de moins de 1%, de plus les résidus sont plus fortement dispersés qu'une loi normale.

Cette étude confirme que cette forme de modèle n'est pas pertinente, mais qu'une classification des incidents en seulement deux classes est intéressante.

Etant donné la forme du modèle, qui n'est composé que de trois paramètres (l'intercept, la pente

et l'écart-type des résidus), la modélisation est trop contrainte, une autre manière de modéliser avec plus de degrés de liberté doit être essayée.

L'approche régression linéaire log-log n'est pas adaptée.

3.3.2.2 Nombre de DCP en fonction de la taille - approche ajustement de distribution

Dans cette section, une nouvelle approche basée sur l'ajustement d'une loi pour chacune des classes est testée. La variable sur laquelle ajuster une loi est le logarithme du nombre de DCP perdues. En effet, sans passage au logarithme la variable est trop dispersée, à titre d'exemple le rapport entre la moyenne et la médiane est de l'ordre de 1 000 à l'intérieur de chaque classe.

Quatre distributions sont testées :

— Weibull
— Gamma
— Normale
— Log-normale

Avec la variable Size2

Un ajustement est effectué pour chacune des 8 classes, ainsi qu'un ajustement sur la totalité de l'échantillon (toutes classes confondues).

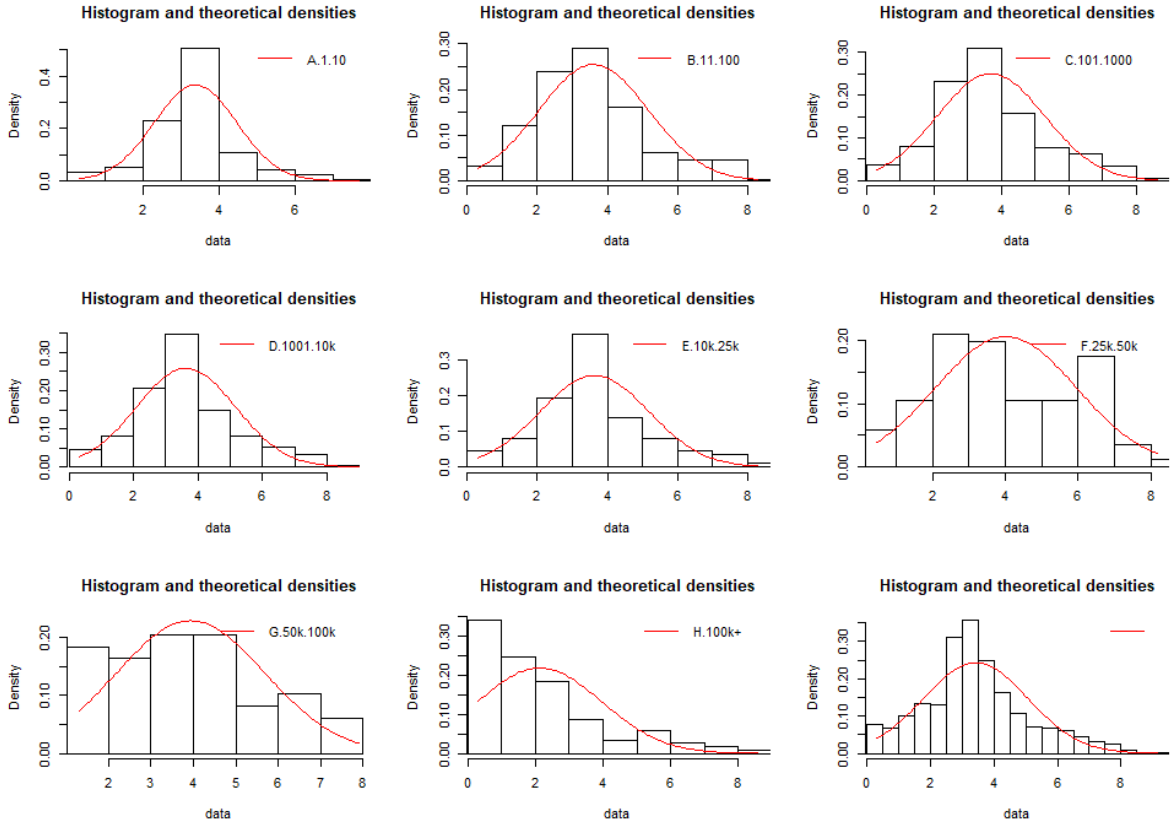


FIGURE 3.18 – Histogramme pour chacune des 8 classes (le 9iè graphe pour la totalité des données) et densité des lois normales ajustées

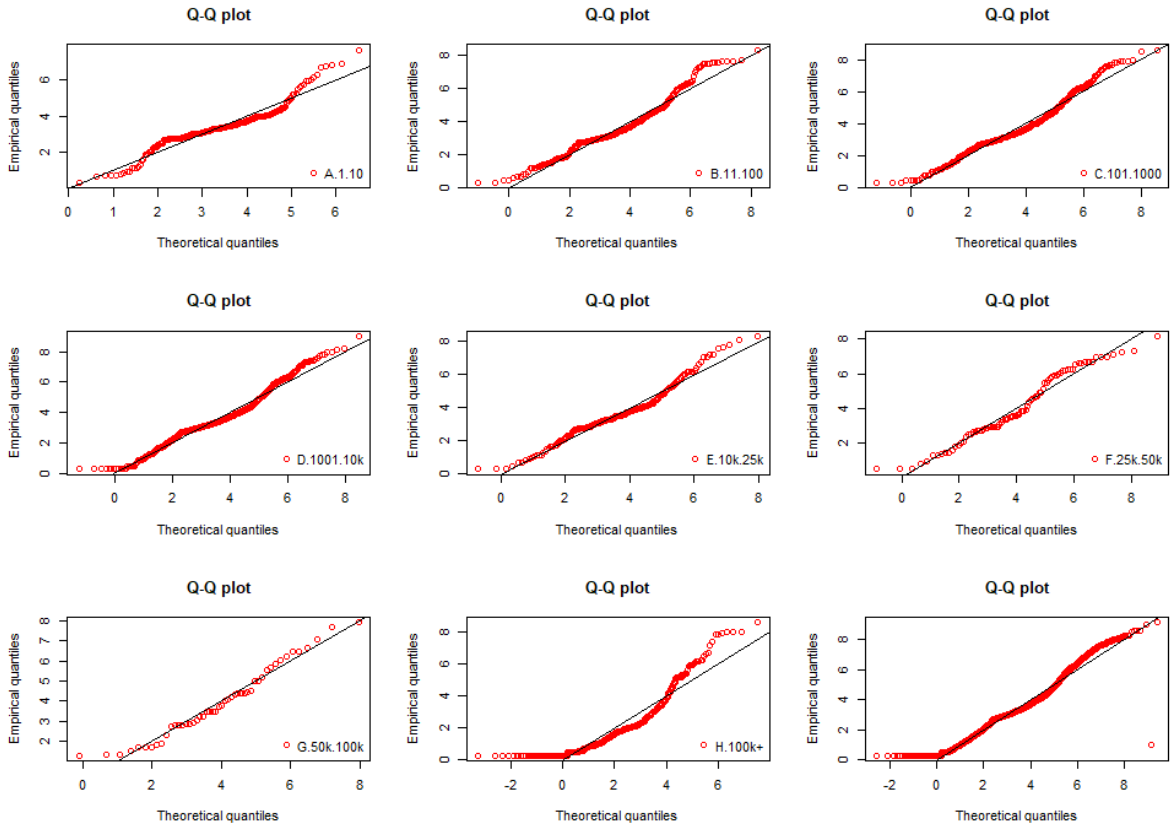


FIGURE 3.20 – Q-Q plot, lois normales ajustées

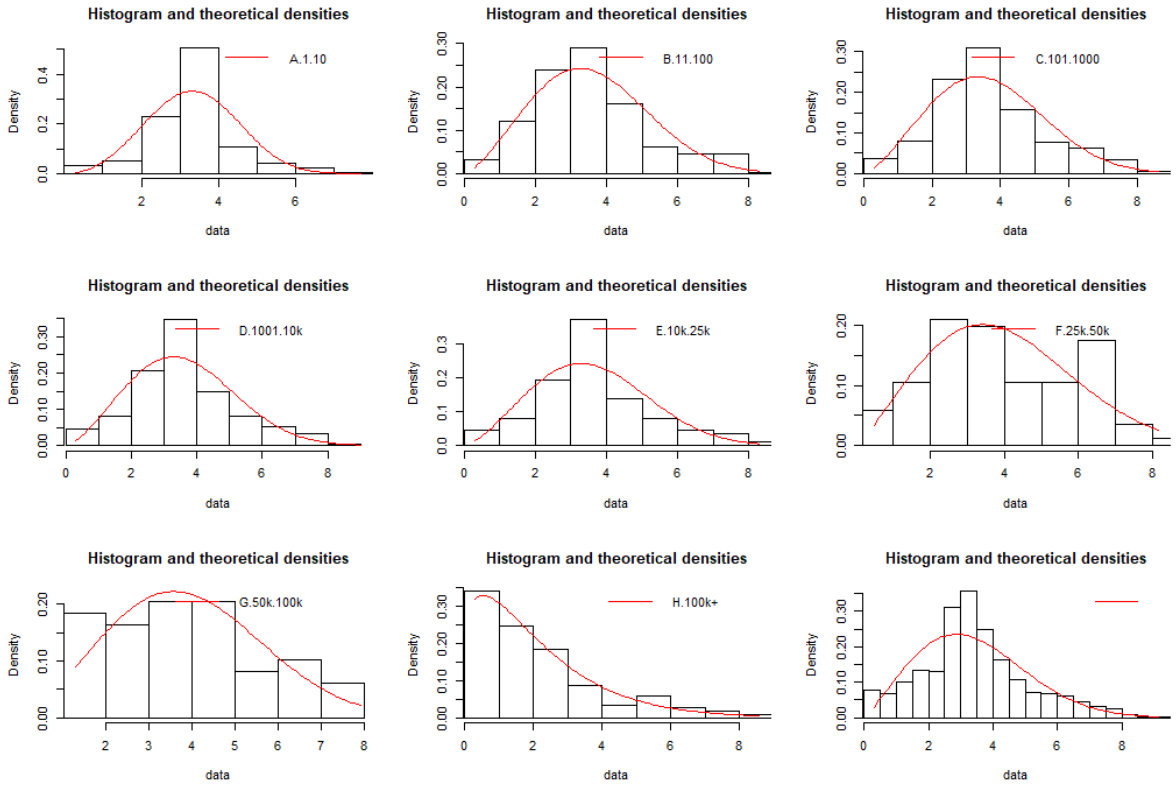


FIGURE 3.19 – Histogramme pour chacune des classes et densité des lois de weibull ajustées

Cette étude fait ressortir plusieurs éléments : le fait que la loi normale ne représente pas bien les valeurs aux quantiles très petits et très élevés pour certaines classes ; que la loi de Weibull ajuste mieux les valeurs aux quantiles élevés que la loi normale.

Cependant, cette modélisation se heurte au fait que les distributions ajustées ne sont pas *cohérentes* entre elles, au sens où les paramètres des distributions entre classes consécutives ont une tendance générale mais sont trop volatils.
Par exemple, la figure 3.28 représente les paramètres pour l'ajustement de loi normales : il semble y avoir une tendance avec des écart-types croissants, et le rapport écart-type / moyenne croissant également avec le nombre d'employés, mais il ne serait pas pertinent de retenir ces paramètres tels quels.

Deux approches sont possibles pour résoudre ce problème : effectuer un lissage sur les paramètres directement, ou refaire l'exercice avec moins de classes.

La solution retenue est d'utiliser la variable *Size1*, qui ne contient que deux classes.

	Classe	μ	σ	$\frac{\mu}{\sigma}$	$exp(\mu + \frac{\sigma^2}{2})$
1	A.1.10	3,37	1,09	0,32	1138877458,45
2	B.11.100	3,61	1,57	0,43	12814277449,06
3	C.101.1000	3,68	1,59	0,43	29585844942,56
4	D.1001.10k	3,63	1,54	0,43	15771525558,63
5	E.10k.25k	3,63	1,55	0,43	17032864730,20
6	F.25k.50k	4,02	1,94	0,48	1215383096163,81
7	G.50k.100k	3,93	1,74	0,44	462269706086,39
8	H.100k+	2,13	1,83	0,86	25046,96
9	Altogether	3,42	1,64	0,48	1791395763,93

TABLE 3.28 – Les paramètres des lois normales ajustées

Avec la variable Size1

Les histogrammes et les densités des distributions ajustées, les Q-Q plots, les fonctions de répartition et les P-P plots sont étudiés.

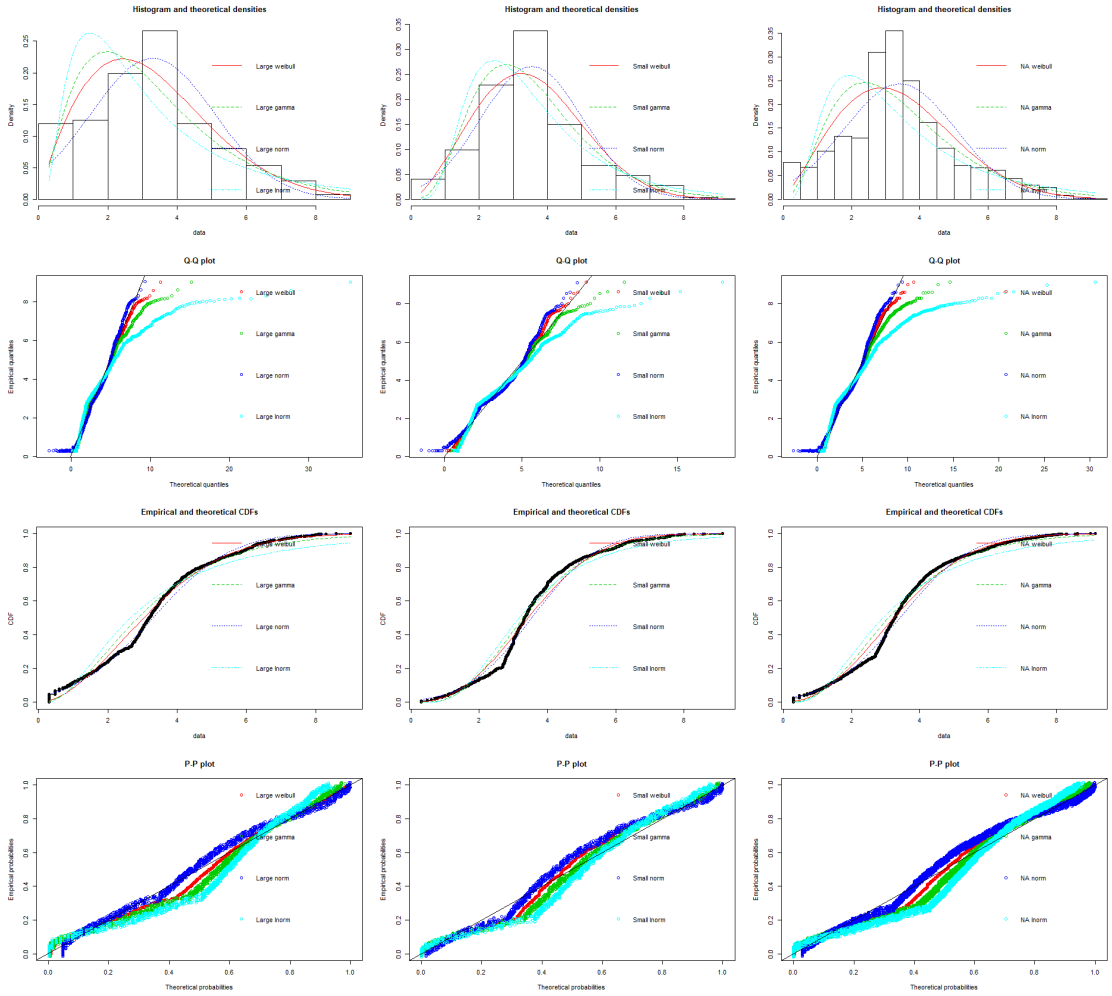


FIGURE 3.21 – Les graphes étudiés - à gauche pour les grandes entreprises, au milieu pour les petites, à droite pour l'ensemble de l'échantillon

En se basant sur ces graphes, les lois Log-normale et Gamma semblent être de mauvais candidats car elles ont des queues trop épaisses. En effet, modéliser le logarithme d'une variable avec une loi log-normale revient à appliquer deux fois la fonction logarithme ce qui est trop pénalisant et confirmé par l'étude graphique. Ces deux lois étant mises de côté, il reste les lois de Weibull et Normale comme candidates, qui sont les lois à queue les plus fines parmi les quatre lois testées.

On remarque aussi une étrangeté sur ces graphes, qui concernent toutes les lois : la représentation des fonctions de répartition et des P-P plots principalement témoignent d'une sorte de "pliure" ou d'irrégularité majeure autour de la valeur $log(nb\ DCP) = 2,7$ (en base 10). Cela correspond à $exp(log(10) * 2,7) = 500$ DCP.
On retrouve le nombre de 500 DCP, correspondant à la contrainte réglementaire qui oblige les entreprises des États-Unis à communiquer publiquement leurs incidents à partir de 500 DCP. Les statistiques et critères d'information seront donnés pour les lois ajustées de la figure 3.21, mais l'étude sera approfondie dans la suite de la section 3.3.2.2.

L'étude graphique n'étant pas suffisante pour choisir entre la loi Weibull et la loi Normale, une étude quantitative s'impose : les statistiques de Kolmogorov-Smirnov, Cramer-Von Mises et Anderson-Darling ainsi que les critères d'information d'Akaike et de Bayes sont estimés.

Statistiques de qualité de l'ajustement	Weibull	Gamma	Norm	LNorm
Kolmogorov-Smirnov	0,10	0,14	0,07	0,18
Cramer von Mises	5,73	11,2	5,79	23,81
Anderson-Darling	31,13	60,90	32,06	134,04
Critères d'Information				
Akaike	14220,30	14530,06	14360,85	15375,21
Bayésien	14232,76	14542,52	14373,31	15387,67

TABLE 3.29 – Statistiques et critères - toutes les entreprises, sur le support $[0; +\infty[$

D'après les statistiques, :

— les lois Gamma et Log-normale ne sont pas adaptées, ce qui confirme l'analyse graphique.

— la loi Normale est meilleure d'après la statistique de Kolmogorov-Smirnov : elle est légèrement plus adaptée pour modéliser le milieu de la distribution.

— La loi de Weibull est légèrement meilleure d'après les statistiques d'Anderson-Darling et Cramer von Mises : elle est plus adaptée pour modéliser les queues de distribution.

Les critères indiquent que la loi de Weibull est meilleure que les autres.

La distribution retenue est la **loi de Weibull**, car une attention plus particulière est portée aux quantiles élevés plutôt qu'au centre de la distribution.

En comparant les quantiles des lois de Weibull et des lois Normales ajustées sur les échantillons de données, il est possible de constater qu'utiliser la loi de Weibull est plus conservateur que la loi Normale aux quantiles élevés ($> 99\%$).

Dans cette section, il a été démontré que la meilleure loi de distribution est la loi de Weibull, qui se comporte mieux sur le support $[0; +\infty[$. La loi normale est également une loi ayant des caractéristiques intéressantes notamment une une queue fine qui permet de modéliser le log de la variable du nombre

Classe	param1	param2
Large	1,86	3,69
Small	2,47	3,98
All	2,10	3,83

TABLE 3.30 – Paramètres des lois de Weibull, ajustées sur les données de la base VERIS, sur le support $[0; +\infty[$

de données perdues.

Il faut encore réfléchir à la pertinence de ce modèle, en le mettant en perspective avec les données : est-ce pertinent d'utiliser un modèle, dans le cadre de simulations, susceptible de modéliser un nombre *infini* de données perdues, quel que soit les caractéristiques d'une entreprise ?
De plus, d'après l'étude fine des données dans la section 2, il a été mis en avant que les incidents de moins de 500 DCP sont moins susceptibles d'apparaître dans la base pour des raisons réglementaires.

Avec des loi tronquées

Dans cette section, des loi tronquées seront ajustées sur les données de la base VERIS. On retiendra le support $[500; 10\ millions]$ DCP, en ligne avec la plage de validité du modèle de coût (jusqu'à 10 millions) et en ligne avec la contrainte réglementaire qui impose aux entreprises de communiquer publiquement sur leurs incidents, à partir de 500 DCP perdus.

Le package *fitdistrplus* ([Delignette-Muller et Dutang, 2015]) a été utilisé dans *R* ([R Core Team, 2019]). Ce package sert à ajuster des distributions à des données. Les distributions peuvent soit celles déjà présentes dans les autres packages de base de *R* soit définies par l'utilisateur. Les distributions tronquées ont été implémentées afin de pouvoir utiliser ce package avec des lois tronquées, notamment de Weibull et Normale. Les fonctions recodées dans *R* l'ont été en repartant de la définition d'une loi tronquée : une loi tronquée est une loi conditionnelle à un intervalle dérivée d'une autre loi de probabilité où l'on ne garde que les tirages sur un intervalle défini.

On a ainsi, pour X la loi tronquée de $X_0 \sim Weibull(a, b)$ (de fonction de répartition F_0) sur l'intervalle $[500; 10\ millions]$:

$$F(x) = \mathbb{P}(X \leq x \mid 500 \leq X \leq 10m) = \frac{\mathbb{P}(500 \leq X \leq x)}{\mathbb{P}(500 \leq X \leq 10m)} = \frac{F_0(x) - F_0(500)}{F_0(10m) - F_0(500)} \tag{3.11}$$

L'équation 3.11 s'écrit avec R :

```
dtweibull <- function(x, shape, scale = 1, low, upp)
{
  PU <- pweibull(upp,shape, scale)
  PL <- pweibull(low,shape, scale)
  dweibull(x, shape,scale) / (PU-PL) * (x >= low) * (x <= upp)
}
```

FIGURE 3.22 – Loi tronquée dans R

De même, les fonctions *ptweibull*, *qtweibull*, *rtweibull*, *dtnorm*, *ptnorm*, *qtnorm* et *rtnorm* sont recodées dans R.

Les données sur lesquels l'ajustement est effectué sont filtrées sur la plage $[500; 10\ millions]$.

La figure 3.23 illustre le fait que les données tronquées permettent de bien mieux modéliser la variable du logarithme du nombre de données que les variables non tronquées. De plus il n'y a plus de

"pliure" à 500 DCP, et les lois tronquées sont confondues avec les données empiriques sur le graphe de la fonction de répartition, et parfaitement alignées sur le P-P plot.

De plus, graphiquement, les données de la colonne de gauche (modalité *Small* uniquement) semblent modélisées de manière très similaire à l'ensemble des données.
Ce point est vérifié statistiquement avec la table 3.31.

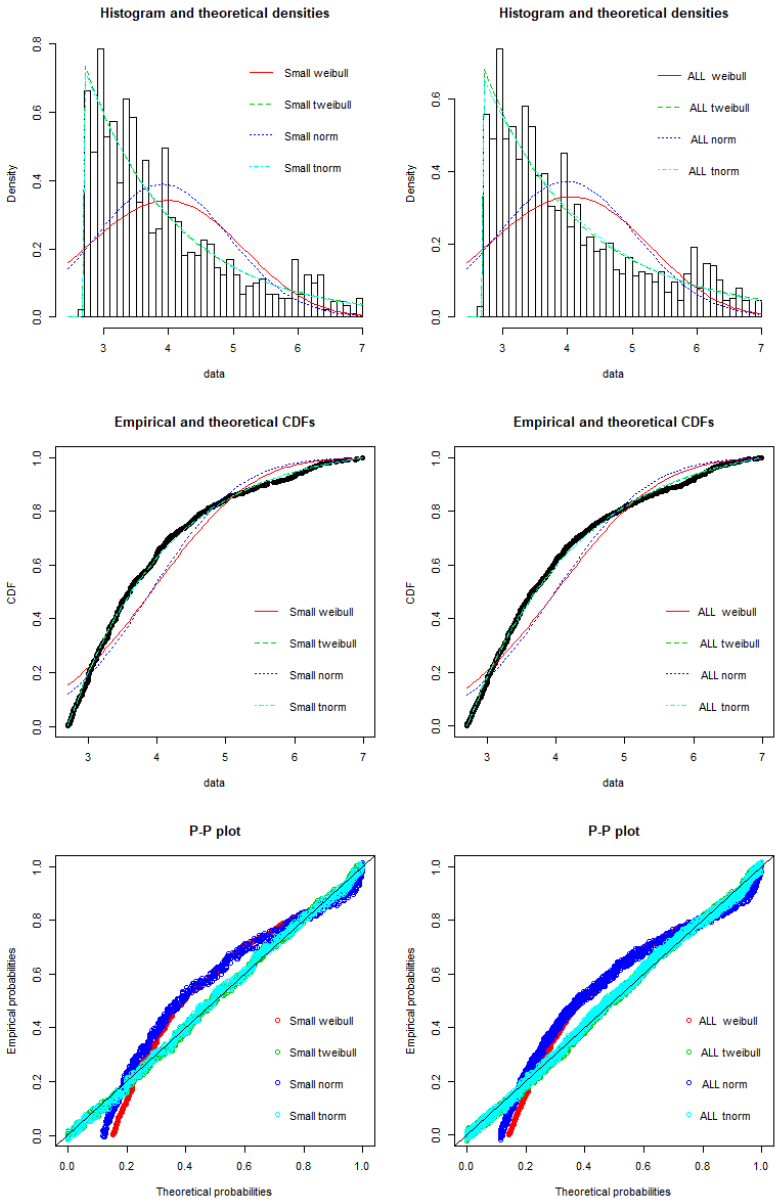


FIGURE 3.23 – Lois de Weibull et Normale ajustées aux données, tronquées et non tronquées et -Size1 = Small à gauche et toutes les données à droite

Statistiques de qualité de l'ajustement	Weibull	Weibull tronquée	Normale	Normale tronquée
Kolmogorov-Smirnov	0,14	0,03	0,13	0,03
Cramer von Mises	10,42	0,27	10,03	0,29
Anderson-Darling	64,01	1,89	61,71	1,95
Critères d'Information				
Akaike	5363,35	4267,76	5272,46	4269,37
Bayésien	5374,31	4278,72	5283,42	4280,33

TABLE 3.31 – Statistiques et critères - toutes les entreprises, sur le support $[log_{10}(500); log_{10}(10\ millions)]$

Distribution	Weibull tronquée		Normale tronquée	
Paramètres	forme (erreur)	échelle (erreur)	μ (erreur)	σ (erreur)
Large	0,74 (0,35)	1,02 (1,12)	-79,84 (0,02)	12,36 (0,01)
Small	0,93 (0,27)	1,25 (0,73)	-59,31 (0,03)	9,55 (0,03)
All	0,82 (0,22)	1,08 (0,64)	-44,56 (1,61)	8,86 (0,21)

TABLE 3.32 – Paramètres estimés et erreurs standard d'estimation des lois tronquées de Weibull et Normales sur le support $[log_{10}(500); log_{10}(10\ millions)]$

Tout d'abord, d'après la table 3.32, les paramètres de la loi normale tronquée sont peu interprétables, avec des μ négatifs de des *sigma* plus élevés que la longueur du support : $[log_{10}(500); log_{10}(10\ millions)]$ soit $[2,7; 7]$.

De plus, les statistiques et les critères d'information représentés en table 3.31 mettent bien en évidence la pertinence de l'utilisation de lois tronquées. Les chiffres de cette table diffèrent pour les loi de Weibull et Normale de la table 3.29 car les supports sont différents.

Enfin, les paramètres des lois de Weibull sont interprétables mais les erreurs standards prennent des valeurs trop élevées pour qu'il soit possible de conclure à une différence statistique significative du comportement des pertes des petites entreprises par rapport aux grandes entreprises.

Plusieurs raisons peuvent expliquer cela :

— La modalité *Small* contient en réalité de nombreuses grandes entreprises : d'après l'Union Européenne (`https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en`), une entreprise moyenne est composée d'entre 50 et 250 employés, les entreprises ayant entre 250 et 1 000 employés étant déjà de grandes entreprises. Cette définition permet d'avoir une idée du niveau de maturité, de sécurité et de vulnérabilité général d'une entreprise et cela permet de supposer que ces niveaux sont similaires pour des entreprises de 250 à 1 000 employés que pour des entreprises de plus de 1 000 employés.

— D'après les statistiques de l'OCDE (`https://stats.oecd.org/Index.aspx?QueryId=81354&lang=fr`), il y a 4 215 610 PME ($<$ 250 employés) aux États-Unis et 26 144 grandes entreprises ($>$ 250 employés) en 2015. En France en 2017, il y a 2 741 040 PME et 4 034 grandes entreprises. Il y a donc 163 fois plus de PME que de grandes entreprises aux États-Unis et 680 fois plus en France !

Or, la base VERIS reporte plus d'incidents concernant des grandes entreprises ($>$ 1 000 employés) que d'incidents concernant des petites entreprises ($<$ 100 employés), tel que représenté en annexe sur la figure A.9.
Cela signifie qu'une grande entreprise américaine a plus de 163 fois plus de risque de subir une attaque de type perte de données impliquant au moins 500 DCP qu'une petite ou moyenne entreprise.

Les deux points évoqués permettent de conclure que la base VERIS est une base représentative des incidents subis par les grandes entreprises, et que les petites et moyennes entreprises subissent très peu d'incidents ou que lorsque c'est le cas ce sont principalement des incidents impliquant moins de 500 DCP.

Pour cette étude, on conservera la loi Weibull tronquée avec un seul jeu de paramètres quel que soit la taille de l'entreprise.

Par contre, pour une entreprise donnée, la loi de Weibull sera tronquée de telle sorte à modéliser un nombre de données perdues sur la plage :

$[500; \textit{Exposition maximale de l'entreprise}]$

L'exposition maximale sera définie comme le plus grand nombre de DCP contenues dans un ou plusieurs systèmes d'information, en fonction de l'imperméabilité entre les systèmes d'information.

3.3.2.3 Comparaison base PRC vs base VERIS

Dans cette section, les points de la base PRC et de la base VERIS sont positionnés sur le même graphe, pour la variable nombre de DCP perdues. Seuls les points entre 500 et 10 millions DCP sont reportés et comme pour les sections précédentes, le logarithme en base 10 est choisi.
De plus, une loi Weibull tronquée est ajustée sur chacun des échantillons. Les deux lois sont également représentées sur le graphe.
La base VERIS contient 1 773 points entre 500 et 10 millions DCP contre 5 203 pour la base PRC. Sur la figure 3.24, les 1 773 points de la base VERIS sont représentés, et les points de la base sont lissés.

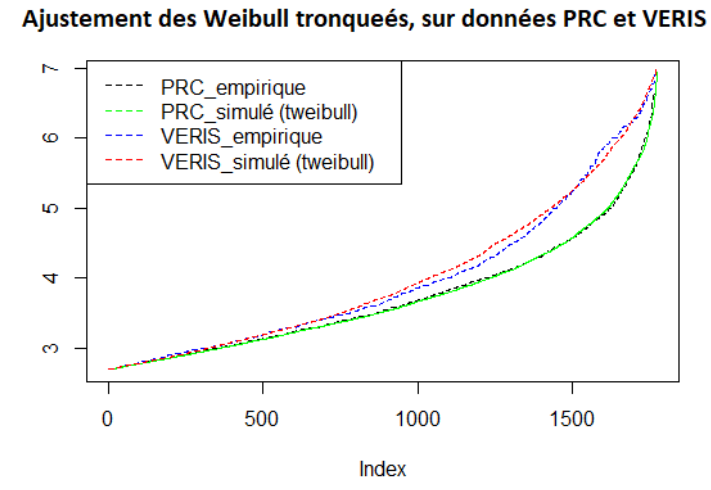


FIGURE 3.24 – Ajustement sur données PRC vs sur données VERIS

L'approche paramétrique basée sur une loi de Weibull tronquée est pertinente pour chacune des bases de données, cependant les incidents des deux bases sont significativement différents.
Utiliser la base PRC pour calibrer le modèle retenu aboutirait à des nombres de DCP simulés plus faibles qu'avec la base VERIS.

Les paramètres de la loi de Weibull tronquée ajustée sur les données PRC sont : $(2,028218; 2,781711)$.

On constate que la modélisation paramétrique avec une loi de Weibull tronquée permet de bien ajuster les données, quelle que soit la base.

De plus, ce graphe permet de bien cerner le fond du problème : se demander à quel point, pour chacune des bases, la distribution paramétrique retenue est intéressant théoriquement, mais le coeur du problème est ailleurs : c'est de savoir si les données que l'on utilise sont bien adaptées et représentatives du risque que l'on souhaite modéliser.

3.3.2.4 En fonction du secteur

De la même manière que pour la taille, aucune modélisation du nombre de DCP en fonction du secteur n'a pu être retenue.
Les modèles ajustés sur des données propres à un secteur ne permettent pas une différenciation statistiquement significative de la modélisation basée sur l'ensemble des données.

3.3.2.5 Synthèse et choix de la modélisation du nombre de DCP perdues

Il a tout d'abord été montré que l'ajustement de distribution sur le logarithme de la variable était adapté. L'ensemble des données a été divisé en classes pour plusieurs variables : *taille de l'entreprise* exprimée en nombre d'employés et *secteur*. Pour chaque variable, plusieurs ajustements, correspondant à plusieurs distributions ont été testés. Seule la variable *taille de l'entreprise* est suffisamment statistiquement explicative pour être intéressante avec cette approche. Cependant, la base de données VERIS contenant une proportion très faible de petites entreprises, la variable n'a finalement pas été retenue.

La modélisation a pu être largement améliorée en utilisant des distributions tronquées. Parmi les deux meilleurs distributions possibles (Weibull et Normale tronquées), seule la loi de Weibull tronquée est satisfaisante. L'utilisation de distributions tronquées permet de prendre en compte de manière naturelle la taille de l'entreprise exprimée en nombre de DCP à risque, qui est utilisé comme borne supérieure du support de la distribution tronquée.

Enfin, la calibration de la distribution de Weibull tronquée sur la base PRC a été comparée à la calibration sur la base VERIS : les deux distributions calibrées représentent bien les données sous-jacentes mais sont significativement différentes car les bases de données contiennent des incidents différents. Cela met en évidence que l'enjeu majeur de la calibration du risque Cyber de perte de DCP est la qualité des données, plutôt que la méthode de calibration.

La modélisation retenue est la loi de Weibull tronquée calibrée sur la base VERIS entière. Ses paramètres de forme et d'échelle sont $0,8152976$ et $1,0818893$. La borne inférieure est 500 DCP à cause de la manière dont a été constituée la base VERIS, et la borne supérieure est adaptée à l'entreprise en fonction de son nombre de DCP à risque, dans la limite de 10 millions de DCP.

3.4 Modélisation de la fréquence

3.4.1 Ponemon

Ponemon estime la probabilité à 29,6% de subir une attaque mettant au cause au moins 10 000 DCP à horizon deux ans (3.8), ce qui correspond, sous l'hypothèse du fréquence qui suit une loi de Poisson, à une espérance de 0,175 attaques par an et par entreprise.
Plus précisément, cette donnée issue de l'étude Ponemon représente la probabilité de subir une nouvelle attaque dans les deux ans, et est estimée à partir de l'échantillon d'environ 500 entreprises sélectionnées pour l'étude qui ont toutes subies des attaques entre Avril 2018 et Juillet 2019. Ce sont majoritairement des grandes entreprises ($>$ 250 employés).

L'échantillon de Ponemon est un ensemble d'entreprises réparties dans le monde entier.

A partir de la donnée de 29,6% associée au seuil de 10 000 DCP, et sous l'hypothèse de la loi de Weibull tronquée pour modéliser le nombre de DCP perdues, il est possible d'estimer la fréquence au seuil de 500 DCP.

En effet, d'après les paramètres de la modélisation retenue en 3.3.2.5 avec X la loi tronquée de $X_0 \sim Weibull(0,8152976, 1,0818893)$ sur l'intervalle $[500; 10\ millions]$, on a :

$$\mathbb{P}(X \geq 10000 \mid X \geq 500) = 1 - F(10\ 000) = 40\% \tag{3.12}$$

Cela signifie que 40% des incidents simulés avec la loi de Weibull tronquée dépassent 10 000 DCP.

La fréquence équivalente au seuil de 500 DCP est donc de $\frac{0,175}{40\%} = 0,4375$.

L'espérance de fréquence annuelle d'attaques entraînant des pertes de données d'au moins 500 DCP est de 0,4375.

3.4.2 CESIN

Le CESIN, ou Club des Experts de la Sécurité de L'information et du Numérique a réalisé un « Sondage OpinionWay pour le CESIN » puis publié son baromètre de la Cyber-sécurité des entreprises en Janvier 2019.

L'enquête a été réalisée auprès des membres du CESIN dont la liste est disponible ici : `https://www.cesin.fr/membres.html`. Ce sont quasiment exclusivement des grandes entreprises.

Le CESIN a publié que 80% des entreprises ayant répondu ont constaté au moins une Cyber-attaque au cours des 12 derniers mois. Malheureusement, le CESIN n'explicite pas le nombre d'attaques liées à des pertes de données.

Le chiffre de 80% correspond à une espérance de 1,61 attaques par an sous l'hypothèse d'une loi de Poisson.

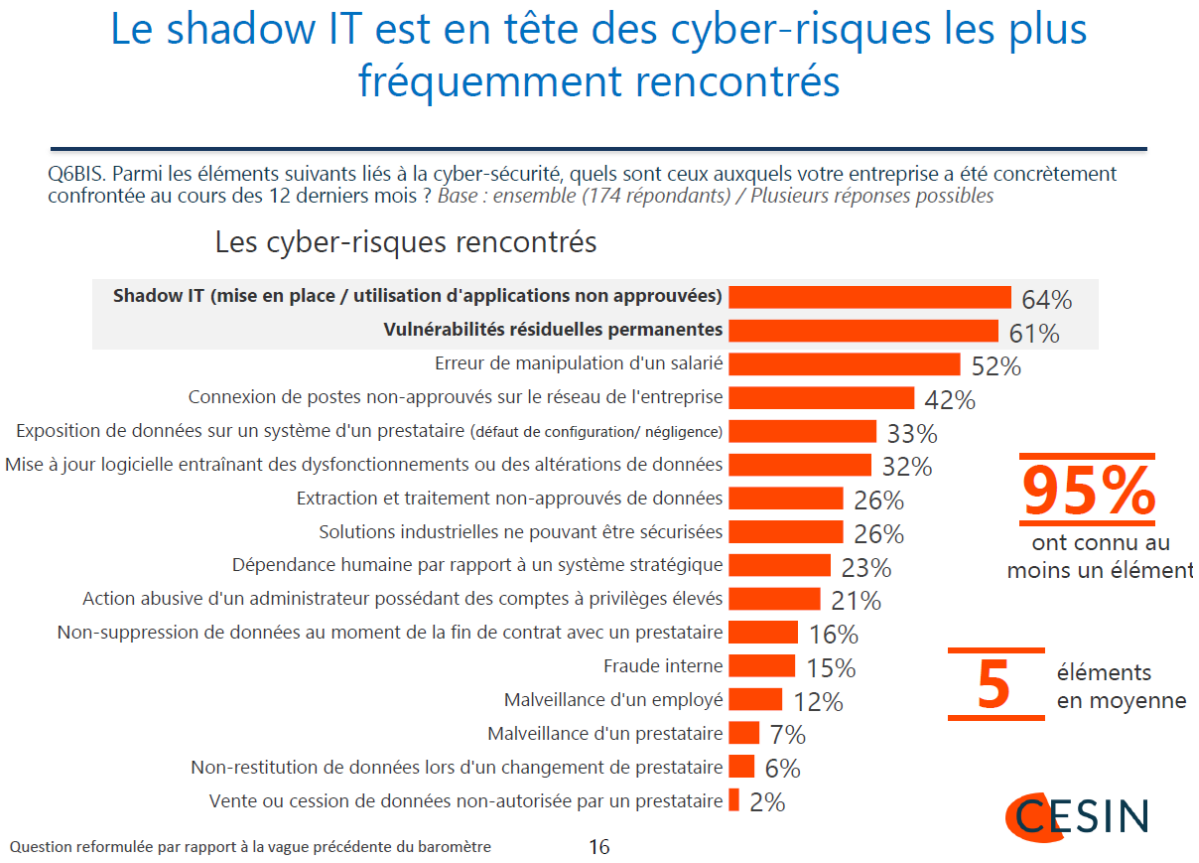


FIGURE 3.25 – Chiffres du CESIN permettant de déduire des fréquence d'occurrence de Cyber-attaques exposant les données d'une entreprise

Type	Probabilité de subir au moins 1 attaque	Espérance de fréquence	Perte de données ?
Toutes attaques confondues	80%	1,61	NA
Shadow IT	64%	1,02	0
Vulnérabilités résiduelles	52%	0,73	0
Erreur de manipulation	42%	0,54	0
Connexion de postes non approuvés	33%	0,40	0
Exp. de données sur un système d'un presta.	32%	0,39	1
M à j. logicielle - des altérations de d.	26%	0,30	1
Extraction/traitement non approuvés de d.	26%	0,30	1
Dépendance humaine à un système strat.	23%	0,26	0
Action abusive d'un administrateur	21%	0,24	0
Non-suppr. à la fin de contrat avec un presta.	16%	0,17	1
Fraude interne	15%	0,16	0
Malveillance d'un employé	12%	0,13	0
Malveillance d'un prestataire	7%	0,07	0
Non-rest. de d. - changement de presta.	6%	0,06	1
Vente ou cession de données non-autorisées	2%	0,02	1

TABLE 3.33 – Utilisation des données CESIN pour l'estimation d'une espérance de fréquence annuelle (1/2)

Périmètre	Espérance de fréquence
Total (somme des espérances de tous les types d'attaques)	4,81
Total - pertes de données uniquement	1,24
Ajustement du total - nb moyen de cases cochées par attaque	2,99
Perte de données ajustée	0,42

TABLE 3.34 – Utilisation des données CESIN pour l'estimation d'une espérance de fréquence annuelle (2/2)

Pour estimer une espérance de fréquence annuelle de subir une attaque de perte de données, les étapes suivantes sont effectuées :

— L'hypothèse que la loi de Poisson est applicable à tous les types d'attaques est retenue.

— L'espérance de fréquence est calculée au total (1,61) et pour chaque risque Cyber (colonne *Espérance de fréquence*).

— La somme des espérances des fréquences pour les attaques susceptibles d'engendrer des pertes

de données est reportée (1,24).

— En supposant que les répondants ont coché plusieurs cases pour chaque attaque effectivement subie, il est possible de déduire que 2,99 cases ont été cochées par attaque subie. C'est le rapport entre l'espérance toutes attaques confondues (1,61), et la somme des espérances de toutes les catégories d'attaques (4,81).

— L'espérance de fréquence « perte de données » est ajustée du facteur 2,99.

L'estimation est de 0,42 Cyber-attaque entraînant des pertes de données par an pour une grande entreprise. Cette estimation concerne la France, et les attaques à partir d'une seule DCP perdue.

Cette deuxième estimation n'est pas tout à fait en ligne avec la première estimation (0,4375 attaques d'au moins 500 DCP, pour les grandes entreprises). Ces deux estimations seraient cohérentes si l'on supposait que les grandes entreprises ne subissaient que des attaques de plus de 500 DCP perdues, ce qui est une hypothèse raisonnable pour une grande entreprise.

Il a été vu que les petites entreprises avaient nettement moins de risque de subir une attaque de plus de 500 DCP qu'une grande entreprise d'au moins un facteur 100 si on utilise le nombre d'entreprises existant aux États-Unis.

3.4.3 Synthèse pour la modélisation de la fréquence

Il a été montré avec l'étude des deux bases de données PRC et VERIS que ni l'une ni l'autre n'est adaptée pour la modélisation de la fréquence d'occurence pour le risque Cyber de perte de DCP.

Deux sources externes ont été étudiées : l'étude Ponemon et un sondage du CESIN. Les deux études permettre d'obtenir des estimations cohérentes, malgré certaines hypothèses qui ont du être retenues notamment sur la forme de la distribution.

La loi de fréquence retenue pour toutes les grandes entreprises est une loi de Poisson de paramètre $0,42$.

Chapitre 4

Application à la tarification de couvertures non-proportionnelles, pour la construction de scénarios et au BGS de l'ORSA

Dans ce quatrième chapitre, le modèle complet sera tout d'abord appliqué aux deux entités fictives de la section 3.1.1 du chapitre 3.

Ensuite, trois autres applications seront proposées : un modèle de tarification de couvertures non-proportionnelles, des scénarios centraux et stressés pour une démarche ERM et enfin une application au BGS de l'ORSA.

Une espérance de fréquence de 0,42 sera retenue afin de modéliser la survenance de Cyber-attaques entraînant des pertes de données, à partir d'une DCP perdue.

4.1 Application aux sociétés fictives

La structure complète du modèle de sévérité est présentée en figure 4.1.

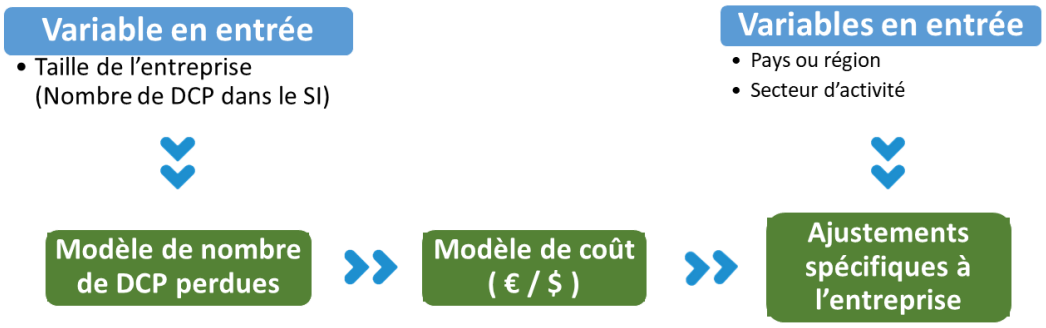


FIGURE 4.1 – Représentation du modèle de sévérité complet

Les deux sous-modèles sont exactement ceux choisis et calibrés dans les sections 3.3.2.5 pour le modèle de nombre de DCP perdues et 3.3.1.5 pour le modèle de coût. On utilise donc la loi de Weibull de paramètres 0,8152976 et 1,0818893 sur l'intervalle $[log_{10}(500); log_{10}(Exposition\ maximale)]$ pour modéliser le logarithme du nombre de données perdues en série avec le modèle VERIS_JJ pour estimer un coût en dollars.
L'étape finale, les ajustements spécifiques à l'entreprise, est basée sur les données marché de Ponemon présentés dans la section Ponemon 3.1.1.

Cependant, étant donné que les deux sous-modèles sont calibrés sur la base VERIS, ceux-ci permettent d'obtenir une modélisation du risque représentative des entreprises qui composent la base VERIS, et non représentative du marché global.
La calibration de la section 3.1.1 est donc légèrement ajustée pour en tenir compte à partir des observations de l'étude de la composition de la base VERIS (section 2.3 du chapitre 2).

Les deux hypothèses suivantes sont retenues :

— Les entreprises de la base VERIS sont toutes basées aux Etats-Unis,

— Les entreprises de la base VERIS sont composées à 50 % d'entreprises du domaine de la santé, les 50 % restants étant représentatifs de l'ensemble des secteurs.

Ces deux hypothèses se traduisent par :

Ajustement proportionnel en fonction du pays :

$$\alpha_{pays} = \frac{coût\ moyen_{pays}}{coût\ moyen_{Etats-Unis}}$$

A partir de cette hypothèse et de la table 3.2 du troisième chapitre, les facteurs d'ajustements sont les suivants :

États-Unis	1, 00
Allemagne	0, 80
Canada	0, 77
Moyen-Orient	0, 71
France	0, 67
Afrique du Sud	0, 64
Royaume-Uni	0, 64
Corée du Sud	0, 63
Italie	0, 60
Japon	0, 58
Scandinavie	0, 54
ANASE	0, 53
Australie	0, 45
Turquie	0, 39
Inde	0, 30
Brésil	0, 29

TABLE 4.1 – Facteurs d'ajustement en fonction du pays

Ajustement proportionnel en fonction du secteur :

$$\alpha_{secteur} = \frac{co\hat{u}t\ moyen_{secteur}}{0.5 \cdot (co\hat{u}t\ moyen_{sant\acute{e}} + co\hat{u}t\ moyen_{global})}$$

Ces facteurs sont appliqués aux coûts en sortie du deuxième sous-modèle.

A partir de cette hypothèse et de la table 3.6 du troisième chapitre, les facteurs d'ajustements sont les suivants :

Santé	$1,24$
Finance	$1,13$
Énergie	$1,08$
Industrie	$1,00$
Pharma	$1,00$
Technologie	$0,97$
Éducation	$0,92$
Services	$0,89$
Loisirs	$0,83$
Transport	$0,73$
Communication	$0,63$
Consumer	$0,50$
Media	$0,43$
Hôtellerie	$0,38$
Vente au détail	$0,35$
Recherche	$0,32$
Secteur public	$0,25$

TABLE 4.2 – Facteurs d'ajustement en fonction du secteur

Application aux entités fictives

On rappelle les caractéristiques utiles des deux entités fictives :

— L'assureur (secteur **finance**) *La Française*, évoluant sur le marché **français** et un total de **200 000** clients particuliers.

— L'assureur (secteur **finance**) *L'Allemande*, évoluant sur le marché **allemand** et un total de **10 millions** de clients particuliers.

Le modèle a été construit avec addactis ® Modeling qui est un progiciel de type DFA (*Dynamic Financial Analysis*). Une approche sévérité - fréquence a été retenue. Un total de 100 000 années ont été simulées, et la graine a été fixée afin de pouvoir isoler les impacts des paramètres d'entrée sur les résultats. Cette approche simulatoire est une approche fréquence - coût qui suppose l'indépendance entre le nombre de sinistres subis par une entreprise et le coût des sinistres.

Indicateur	La Française	L'Allemande
Coût moyen d'une attaque hors ajustements	$5,50m$ \$	$9,30m$ \$
Coût moyen d'une attaque ajusté	$4,18m$ \$	$8,38m$ \$
Coût moyen annuel ajusté	$1,72m$ \$	$3,30m$ \$
Quantile 99,5 % du coût annuel hors ajustements	$69,90m$ \$	$110,81m$ \$
Quantile 99,5 % du coût annuel ajusté	$53,21m$ \$	$99,88m$ \$

TABLE 4.3 – Résultats du modèle pour les deux entités fictives

La table 4.3 permet de se représenter la distribution du risque pour les deux entités fictives et de mesurer certaines sensiblités aux paramètres spécifiques à l'entreprise. La distribution est disponible avec addactis Modeling ®, ces indicateurs fournissent une information synthétique.

Tout d'abord, comparons les coûts moyens d'une attaque obtenus ici avec ce qui a été obtenu en première approche à partir des données Ponemon uniquement, représentées dans la table 3.7. Cette méthode ne permettait que d'estimer un coût moyen, et il a été obtenu $5,93m$ \$ pour La Française et $8,20m$ \$ pour l'Allemande, contre $4,18m$ \$ et $8,38m$ \$ avec le modèle stochastique. Ces estimations sont cohérentes entre elles, compte tenu des méthodes très différentes pour les obtenir ainsi que de la volatilité du risque. On pourra noter que le modèle permet d'obtenir une moyenne légèrement inférieure pour *La Française*, qui est de petite taille.

Ensuite, dans la table 4.3, la première ligne (coût moyen hors ajustements) permet d'isoler l'impact de la taille de l'entreprise (exprimée en nombre de DCP détenues par l'entreprise, ici le nombre de clients) sur les résultats du modèle. La différence entre $5,50m$ \$ et $9,30m$ \$ s'explique uniquement par le choix de la borne supérieure du premier sous-modèle (200 000 contre 10 millions ici). De même pour l'écart entre les quantiles à 99,5 % hors ajustements : $69,90m$ \$ contre $110,81m$ \$.

Ici, l'impact de la taille de l'entreprise sur la moyenne est de 69 % contre 58 % au niveau du quantile à 99,5 %. Cela illustre le fait que les grandes entreprises sont nettement plus susceptibles de subir de très grosses attaques.

Ces données montrent bien également la très forte incertitude du risque et du modèle, le rapport entre quantile à 99,5 % et moyenne est d'environ 30.

4.2 Mesures de transfert de risque et modèle de tarification

Le marché de l'assurance Cyber dans le monde est concentré à hauteur de 85% aux États-Unis en termes de primes émises (2019). C'est un marché en très forte croissance dans l'ensemble du monde, caractérisé par un rattrapage d'autres zones géographiques, l'Europe en particulier notamment avec l'effet de la mise en place du règlement général sur la protection des données (RGPD) entré en vigueur le 25 mai 2018 mais aussi la prise de conscience croissante des entreprises.

Le marché de l'assurance Cyber double tous les 3 à 4 ans dans le monde. Il est passé (en dollars US) de 2,6 mds en 2016 à 5,2 mds en 2018, est estimé à entre 6,3 mds et 8,2 mds en 2020 et 14 mds en 2022. Les acteurs économiques, en particulier les entreprises, doivent considérer, en plus des mesures de gouvernance et d'atténuation du risque, de souscrire des garanties contre les Cyber-attaques.

Les acteurs majeurs du marché dans le monde sont Chubb (environ 16% de part de marché), AXA (environ 13%) puis AIG et Beazley.

Le RGPD fait entrer le risque de pertes juridiques liées à des Cyber-attaques dans une nouvelle dimension. Il prévoit en effet des sanctions pécuniaires allant de 20m € à 4 % du chiffre d'affaires mondial. Avant l'entrée en vigueur du RGPD, les entreprises européennes ne risquaient pas plus de 500 000 € d'amendes. Par exemple, en juillet 2019, British Airways risquait une amende de 183m £ (environ 200m €), soit environ 1,5% de son chiffre d'affaire mondial, consécutivement à une attaque importante au cours de laquelle des informations personnelles (dont noms, prénoms, adresse, coordonnées bancaires, information de compte client, information sur les voyages effectués) d'environ 500 000 clients ont été dérobées.

4.2.1 Polices Cyber

Il existe principalement trois types de polices Cyber : les polices spécifiques (ou affirmatives), les polices silencieuses et les extensions de polices traditionnelles.

Les polices affirmatives sont exclusivement destinées à couvrir des risques Cyber, les polices silencieuses sont des polices de couverture de responsabilité générale qui n'excluent pas spécifiquement le risque Cyber, enfin les extensions de polices traditionnelles sont des contrats qui étendent la couverture d'une police traditionnelle au risque Cyber.

Le marché européen voit une augmentation de la proportion de couvertures affirmatives, et une baisse en proportion de couvertures silencieuses ou basées sur des polices traditionnelles.

4.2.2 Garanties Cyber

Les garanties Cyber peuvent être décomposées en garanties directes (1st party) ou au tiers (3rd party). Par exemple, si une entreprise qui fournit des services à d'autres entreprises subit un incident Cyber et devient dans l'incapacité d'assurer la qualité de service contractuelle qui la lie avec ses clientes, une garantie au tiers peut permettre de couvrir les dédommagements dus à ses clientes.

Garantie	1st / 3rd	Description
Atteinte à la confidentialité des données	1st	Coûts relatifs à la notification, aux amendes et indemnités.
Perte de données	1st	Le coût de reconstitution des données perdues
Défaillance de service à cause de la défaillance de systèmes d'informations	3rd	Compensation au tiers pour indisponibilité du service
Défaillance de systèmes d'information et perte d'exploitation	1st	Pertes liées à l'arrêt de la production
Protection et assistance juridique	1st	Coût liés aux frais juridiques
Atteinte médiatique	1st	Frais liés à la dégradation de sa réputation
Réponse à un incident	1st	Frais liés à la recherche de la cause d'un incident et à la gestion de crise
Propriété intellectuelle	1st	Coûts liés à la perte de valeur de propriété intellectuelle
Cyber-extorsion	1st	Coût lié au paiement d'une rançon

TABLE 4.4 – Les garanties Cyber les plus courantes sur le marché de l'assurance

4.2.3 Tarification d'une couverture non-proportionnelle - approche simulatoire

Dans cette section, la modélisation en deux étapes est mise en oeuvre de manière identique à la mise en oeuvre de la section 4.1, et la même graine est conservée. Le nombre d'années simulées est 100 000.

4.2.3.1 Structure de la couverture non-proportionnelle

La couverture non-proportionnelle utilisée a été choisie similaire à un traité de type Excédent de Sinistre (XS) car sa paramétrisation est souple et permet de tarifer des couvertures plus simples si besoin.

Pour cette exercice, il est choisi de tarifer une tranche peu travaillante pour mettre en évidence les différences de prix suivant le type d'entreprise.

Les différents paramètres du traité et paramètres de tarification sont :

Priorité	$5m$ \$
Portée	$40m$ \$
Nombre de reconstitutions	2
Prime de reconstitution	$1m$ \$
Coefficient de chargement pour risque (k)	10 %
Frais de chargement (θ)	20 %

TABLE 4.5 – Paramètres du traité

L'équation suivante est utilisée pour l'obtention des résultats :

$$Prime\ Commerciale = \frac{Prime\ Pure + k * \sigma_{charge\ cédée}}{1-\theta}$$

Avec $\sigma_{charge\ cédée}$ l'écart-type de la charge cédée simulée.

4.2.3.2 Résultats : distributions et observations graphiques

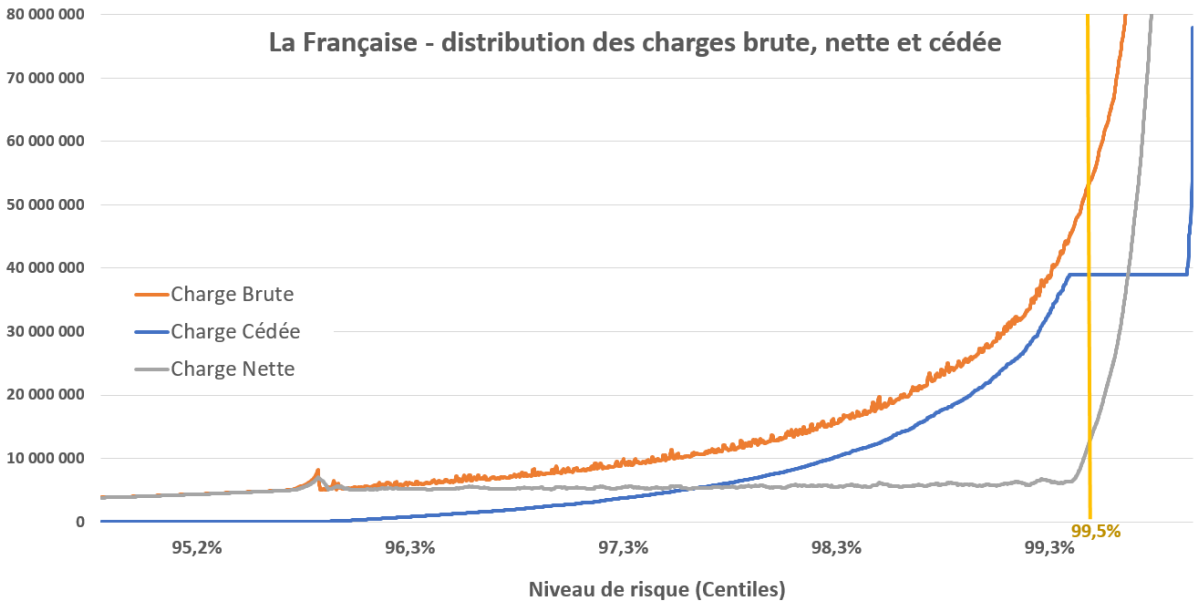


FIGURE 4.2 – Distribution des charges brute, nette et cédée pour La Française

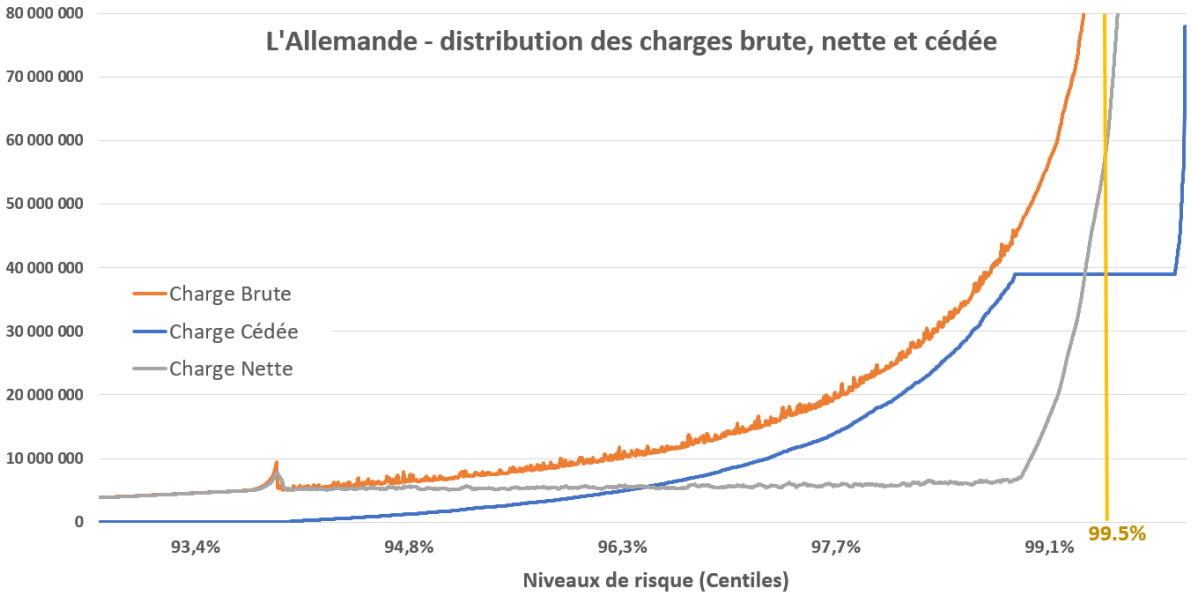


FIGURE 4.3 – Distribution des charges brute, nette et cédée pour L'Allemande

Pour La Française, seules 4,2 % des simulations ont donné lieu à une charge cédeé non nulle, contre 6,0 % pour l'Allemande. La couche est donc peu travaillante, et L'Allemande a environ 50 % de risque supplémentaire de recourir à sa couverture non-proportionnelle que La Française.

Sur ces graphiques, il est possible de retrouver les résultats de la table 4.3. On notera que la charge annuelle cédée n'est pas une fonction croissante de la charge annuelle brute. Sur cette représentation, les simulations ont été triées en fonction de leur charge cédée, et la charge brute ainsi que la charge nette ont été lissées pour améliorer la représentation graphique.

De plus, le quantile de la charge cédée associé au niveau de risque à 99,5 % est le même pour les deux entités et égal à $39m$ \$, ce qui correspond à la survenance d'un sinistre d'au moins $40m$ \$ et de zéro sinistre supplémentaire ou de sinistres supplémentaires inférieurs à la priorité du traité. Le chiffre de $39m$ \$ correspondant à une cession de $40m$ \$ et à l'émission simultanée d'une prime de reconstitution de $1m$ \$ conformément aux paramètres du traité.

4.2.3.3 Résultats quantitatifs : Gain en capital économique, Prime pure, Prime commerciale

Indicateur	La Française	L'Allemande
Quantile 99,5 % (charge brute) (1)	$53,21m$ \$	$99,88m$ \$
Quantile 99,5 % (charge nette) (2)	$13,97m$ \$	$59,29m$ \$
Quantile 99,5 % (charge cédée)	$39,00m$ \$	$39,00m$ \$
Gain en capital économique (1) - (2)	$39,23m$ \$	$40,59m$ \$
Prime Pure	$545k$ \$	$908k$ \$
Ecart-type de la charge cédée	$3,82m$ \$	$5,11m$ \$
Chargement pour risque	$382k$ \$	$511k$ \$
Prime Commerciale	$1,16m$ \$	$1,77m$ \$

TABLE 4.6 – Gain en capital et tarification de la couverture non-proportionnelle étudiée pour La Française et L'Allemande

Sur la table 4.6, le quantile de la charge nette n'est pas exactement égal à la différence entre les quantiles des charges brutes et cédées (qui est le gain en capital économique) car la charge cédée n'est pas une fonction croissante de la charge brute.

Cette couverture non-proportionnelle permettrait un gain en capital économique (hors diversification) similaire pour les deux entités fictives, de l'ordre de $39m$ \$, pour un coût de $1,16m$ \$ et $1,77m$ \$, ce qui en fait une couverture de risque opérationnel Cyber intéressante pour un assureur, en supposant que le risque Cyber rentre en compte dans le calcul du capital économique de ce dernier.

Cette application permet de mettre en évidence des niveaux de primes commerciales très différents, $1,16m$ \$ contre $1,77m$ \$, selon les caractéristiques de l'entreprise, une part importante de cette différence étant liée aux tailles différentes des deux entreprises.

4.3 Application au BGS de l'ORSA

L'objectif de cette section est de proposer un ajustement de la formule standard permettant de mieux prendre en compte le risque opérationnel Cyber, c'est-à-dire à des défaillances de systèmes d'information liés à un acte malveillant.

4.3.1 Le calcul en vigueur sous Solvabilité 2

Le risque opérationnel correspond à la mesure de l'impact sur l'activité d'incidents opérationnels auxquels l'entreprise peut avoir à faire face, dont la fraude, des incidents informatiques, RH, liés à la conformité, des erreurs de calculs, des incidents Cyber.

Dans le cadre du calcul de l'exigence réglementaire de la réglementation Solvabilité 2, le règlement délégué 2015/35, article 104, exige des compagnies européennes d'estimer le risque opérationnel selon la formule :

$$SCR_{Oprationnel} = min(0,3 \cdot BSCR; Op) + 0,25 \cdot Exp_{Ul}$$

Avec $BSCR$ pour désigner le captial de solvabilité requis de base, Op pour désigner le capital requis de base pour risque opérationnel et Exp_{Ul} pour désigner le montant des dépenses encourues au cours

des 12 derniers mois en ce qui concerne les contrats d'assurance vie où le risque d'investissement est supporté par les preneurs.

Dans le cadre de cet exercice, c'est le capital requis de base pour risque opérationnel qui est la métrique d'intérêt. Op s'écrit comme :

$$Op = max(Op_{Premiums}; Op_{Provisions})$$

Où :

$Op_{Premiums}$ et $Op_{Provisions}$ désignent le capital requis pour risque opérationnel sur base respectivement des primes acquises et des provisions techniques.

La composante $Op_{Premiums}$ est calculée comme la somme de 4 % du montant des primes pour les engagements d'assurance et de réassurance vie au cours des 12 derniers mois, sans déduction des primes des contrats de réassurance mais avec déduction des primes pour lesquels le risque est porté par l'assuré et 3 % des primes pour les engagements d'assurance et de réassurance non-vie au cours des 12 derniers mois.
De plus, en cas d'augmentation de plus de 20 % du montant de primes acquises tant en vie qu'en non-vie par rapport à l'année précédente, la contribution au calcul des primes acquises (tant en vie qu'en non-vie) au cours des derniers 12 mois supérieures aux primes acquises au cours des 12 mois précédant les 12 derniers mois augmentés de 20 % est doublée.

La composante $Op_{Provisions}$ est la somme de 0,45 % des provisions techniques pour les engagements d'assurance vie et de réassurance vie dont sont déduites les provisions pour les engagements dont le risque est porté par l'assuré, et de 3 % des provisions techniques d'engagements d'assurance et de réassurance non-vie.

4.3.2 Proposition de modification pour prendre en compte le risque Cyber

Pour prendre en compte le risque Cyber dans le calcul de la formule standard, il est proposé d'ajuster la composante Op :

$$Op_{ajustée} = max(Op_{Premiums}; Op_{Provisions}; Op_{Cyber})$$

La composante Op_{Cyber} pourrait être décomposée en trois sous-composantes pour prendre en compte les trois risques principaux en Cyber : l'attaque par déni de service, l'altération de données à cause d'un rançongiciel ou la perte de DCP. Dans le cadre de ce mémoire, seule une proposition pour la mesure du risque de perte de DCP est proposée. La formule devient donc :

$$Op_{ajustée} = max(Op_{Premiums}; Op_{Provisions}; Op_{Cyber_{DCP}})$$

La modification proposée repose sur :
— une estimation du quantile à 99,5 % du risque Cyber de DCP avant transfert de risque adaptée au profil de risque de l'entreprise,

— la prise en compte d'un transfert de risque à travers l'utilisation d'une couverture d'assurance.

4.3.2.1 Capital requis pour le risque Cyber de perte de DCP avant ajustements spécifiques et avant transfert de risque

Le capital requis avant ajustements spécifiques et avant transfert de risque est issue de l'approche simulatoire de la table 4.3. En effet, le quantile à 99,5 % du coût annuel hors ajustements de L'Allemande correspond à la quantité recherchée ($110,81m$ \$). Etant donné que toutes les entreprises soumises à Solvabilité 2 sont des entreprises évoluant dans le secteur financier, il est possible d'appliquer au montant de $110,81m$ \$ l'ajustement pour l'ensemble du secteur $(1,13)$, de telle sorte à obtenir un montant de capital requis avant ajustements spécifiques et avant transfert de risque de $125m$ \$. De plus, ce montant peut-être converti en euros pour plus de simplicité. En retenant le taux du 31 décembre 2019 (1 euro pour 1,12 dollar), ce montant devient $112m$ €.

4.3.2.2 Ajustements en fonction du pays et de la taille de l'entreprise

L'ajustement en fonction du pays est un ajustement proportionnel directement issu de la table 4.1.

L'ajustement en fonction de la taille de l'entreprise doit être appliqué en fonction du nombre d'individus dont les Données à Caractère Personnel sont hébergées par l'entreprise. Dans la cadre d'une compagnie d'assurance, le nombre d'assurés particuliers de l'entité est un bon indicateur. Cet ajustement doit être appliqué à la baisse uniquement, pour les petites entreprises étant donné que le capital requis de base de $112m$ € a été estimé sur la base d'une entreprise de taille très importante (10 millions de DCP).

Une autre contrainte existe, celle de la simplicité d'utilisation dans la cadre de la formule standard.

Il est rappelé ici que la sensibilité du modèle à la taille de l'entreprise est disponible dans la table 4.3 : pour La Française (200 000 clients), le quantile à 99,5 % est de $69,90m$ \$ contre $110,81m$ \$ pour L'Allemande, ce qui représente un rapport de 0,63.

Les facteurs d'ajustement pour la taille proposés (et calibrés avec le modèle) sont proposés dans la table 4.7 :

Nombre d'assurés	Facteur d'ajustement de la taille
Supérieur à 1 million	$1,0$
100 000 à 1 million	$0,7$
Inférieur à 100 000	$0,5$

TABLE 4.7 – Facteurs d'ajustement pour le BGS en fonction de la taille

Le montant ajusté (en €) devient donc :

$Op_{Cyber_{DCPbrut}} = 112m \cdot \alpha_{pays} \cdot \alpha_{taille}$

La formule peut être appliquée à La Française et à L'Allemande :

Indicateur	La Française	L'Allemande
α_{pays}	0,67	0,80
α_{taille}	0,7	1,0
$Op_{Cyber_{DCPbrut}}$	52,80m €	89,32m €
Quantile 99,5 % du coût annuel ajusté	47,51m €	89,18m €

TABLE 4.8 – Comparaison des résultats du modèle stochastique avec la proposition de calcul pour le BGS

Les montants de $Op_{Cyber_{DCPbrut}}$ et de quantile 99,5 % du coût annuel ajusté en € sont proches et cohérents, car les facteurs de la formule de $Op_{Cyber_{DCPbrut}}$ ont été calibrés à partir des résultats du modèle. La formule approchée permet de calculer presque instantanément la quantité.

4.3.2.3 Prise en compte d'une couverture du risque Cyber

Dans le cadre de la proposition de formule pour le BGS, il est nécessaire de prendre en compte les mesures de transfert de risque mise en oeuvre par l'entreprise considérée afin de valoriser le risque restant à la charge de l'entreprise.

L'estimation de la charge brute prend en compte l'ensemble des coûts auxquels une entreprise pourrait être exposée, dont des coûts inassurables tels que des coûts liés à des amendes, qui sont dans de nombreux pays légalement inassurables.

Dans la formule : $Op_{ajustée} = max(Op_{Premiums}; Op_{Provisions}; Op_{Cyber_{DCP}})$, la quantité $Op_{Premiums}$, dont le calcul a été précisé précédemment, correspond environ à l'amende maximale que la réglementation RGPD prévoit : la réglementation prévoit un maximum de 4 % du total des primes émises, et la quantité $Op_{Premiums}$ correspond à la somme de 4 % des primes acquises en assurance vie et de 3 % des primes acquises en assurance non-vie, majoré dans le cas d'une augmentation fortes des volumes de primes par rapport à l'année précédente.

Cela veut dire que $Op_{Premiums}$ est toujours supérieure à la part non assurable de $Op_{Cyber_{DCP}}$ et qu'il n'a pas besoin d'ajuster la formule proposée pour tenir compte de la part non assurable de $Op_{Cyber_{DCP}}$.

De plus, l'hypothèse suivante est retenue : l'estimation de $Op_{Cyber_{DCPbrut}}$ à partir de la formule simplifiée est supposée issue d'un seul sinistre survenu dans l'année. Cette hypothèse simple permet d'appliquer une couverture non-proportionnelle s'appliquant par sinistre.

Pour passer de $Op_{Cyber_{DCPbrut}}$ à $Op_{Cyber_{DCP}}$, il suffit désormais d'appliquer la couverture de transfert de risque à la quantité $Op_{Cyber_{DCPbrut}}$.

Sur les entités fictives La Française et L'Allemande, la priorité et la portée exprimées en € deviennent :

Indicateur	La Française	L'Allemande
$Op_{Cyber_{DCPbrut}}$	52,80m €	89,32m €
Priorité (€)	4,46m €	4,46m €
Portée (€)	35,71m €	35,71m €
Prime de reconstitution (€)	0,89m €	0,89m €
$Op_{Cyber_{DCP}}$	17,99m €	54,50m €

TABLE 4.9 – Passage de $Op_{Cyber_{DCPbrut}}$ à $Op_{Cyber_{DCP}}$ pour les entités fictives

L'application proposée permet de se rendre compte que l'entreprise a peu de leviers d'action sur la calcul de $Op_{Cyber_{DCPbrut}}$, mais que la quantité $Op_{Cyber_{DCP}}$ dépend très fortement de la solution de transfert de risque retenue. De plus, l'application proposée permet de se rendre compte qu'à solution de transfert de risque identique, les résultats obtenus pour $Op_{Cyber_{DCP}}$ varient fortement en fonction des caractéristisques propres de l'entreprise : pays et taille.

4.3.2.4 Mise en perspective sur des cas réels

Une proposition de prise en compte du risque Cyber propre à une entreprise a été émise et mise en oeuvre sur des entités fictives. Dans cette sous-section, des cas réels vont être choisis, afin de comparer l'évaluation de la quantité Op par deux entreprises réelles avec les quantités $Op_{Cyber_{DCP}}$ évaluées dans la sous-section précédente 4.3.2.3.

Pour la mise en perspective avec des cas réels, les SFCR (*Solvency Financial Condition Report*) des deux acteurs seront étudiés. Deux acteurs ont été recherchés : un acteur de petite taille et un second de grande taille.

Le choix a finalement été porté sur la Mutuelle des Motards pour l'acteur de petite taille et Allianz Group pour l'acteur de grande taille. On notera que la Mutuelle des Motards a évalué son risque opérationnel avec la formule standard contre un modèle interne pour Allianz Group.

Indicateur	La Mutuelle des Motards	Allianz Group
Chiffre d'affaires	108,82m € (2018)	142 400,00m € (2019)
Nombre d'assurés	337k (2018)	78 000k (2019)
SCR Opérationnel (Op)	4,69m € (2018)	3 059,44m € (2019)
$Op_{Cyber_{DCPbrut}}$	37,72m €	89,32m €

TABLE 4.10 – Chiffres clés pour des cas réels

D'autres informations sont intéressantes dans les deux SFCR. La mention de risque Cyber a été recherchée. Le SFCR de la Mutuelle des Motards ne mentionne pas ce risque. Le niveau de détail de la couverture du risque opérationnel est cohérente avec la taille de cet acteur sur le marché, modeste et donc conforme avec l'esprit de la réglementation Solvabilité 2 qui impose un niveau de détail croissant avec la taille de l'entreprise. Le risque Cyber étant une simple composante du risque opérationnel, un acteur de cette taille peut choisir de ne pas le détailler dans son SFCR.

En revanche, dans le SFCR d'Allianz Group, il est mentionné dans la partie *Risk Profile - Operational Risk - Mitigation of risks*, ces explications sur la manière d'Allianz Group pour atténuer le risque Cyber : *Cyber risks are mitigated through investments in Cyber security, Cyber insurance that Allianz buys from third-party insurers, and a variety of ongoing control activities.*

En français, ces explications se trouveraient dans la partie *Profil de Risque - Risque Opérationnel - Atténuation des risques* et pourraient se traduire par *Les risques Cyber sont atténués grâce à des investissements dans la sécurité des systèmes d'informations, le recours à l'assurance Cyber souscrite auprès d'assureurs tiers et une multitude d'activités de contrôle.*

La confrontation à des cas réels permet de voir que l'impact de cette nouvelle formule serait moindre voire nulle pour les plus grands acteurs du marché. En effet ici, il a été mis en évidence que pour Allianz Group, le SCR Opérationnel (3 059,44m €) est très nettement supérieur au $Op_{Cyber_{DCPbrut}}$ (89,32m €). De plus, Allianz Group a déjà mis en place des solutions d'atténuation de ses risques Cyber tel que précisé dans le SFCR.

En revanche pour La Mutuelle des Motards, le SCR Opérationnel (4,69m €) est nettement inférieur à $Op_{Cyber_{DCPbrut}}$ (37,72m €). Dans le cadre de l'utilisation de cette formule de calcul, un petit acteur tel que La Mutuelle des Motards serait donc très fortement incité à initier des solutions d'atténuation de son risque Cyber notamment via l'achat de couvertures d'assurance Cyber auprès de tiers.

Analyse

L'estimation du risque opérationnel dans la Formule Standard, Op, étant une combinaison linéaire de quantités directement représentatives de la taille d'une entreprise (primes acquises ou provisions techniques), il est possible de s'attendre à observer une relation à peu près proportionnelle entre Op et la taille d'une entreprise. Sur la table 4.10, on constate qu'Allianz Group est 1308 fois plus grande que La Mutuelle des Motards (en chiffre d'affaires). On peut également observer que $\frac{Op_{Allianz}}{Op_{LMdM}} = 652$: le rapport de proportionnalité est à peu près respecté. On notera qu'Allianz Group estime son risque Opérationnel avec un modèle interne, ce qui permet d'être plus représentatif de son exposition propre mais cela permet également le plus souvent d'obtenir une estimation moins élevée qu'avec l'utilisation de la Formule Standard, et donc le rapport est ici légèrement favorable à Allianz Group (652 contre 1308).

Or, la composante $Op_{Cyber_{DCPbrut}}$ ne croît pas proportionnellement avec la taille de l'entreprise : il a été vu avec les chiffres de la table 4.3 qu'entre une entreprise avec 200 000 assurés contre 10 millions d'assurés, soit un rapport 50, le rapport entre les coûts moyens est légèrement inférieur à deux.

Le fait que les plus grandes entreprises ne soient pas impactées par la proposition de modification de la Formule Standard est donc un résultat vérifié sur des cas réels dans cette sous-section mais qui était prévisible.

Si d'autres risques Cyber étaient inclus dans la proposition de modification, notamment des risques susceptibles d'engendrer des pertes d'exploitation ou nuisant fortement à la réputation des entreprises, les grandes entreprises pourraient également être impactées par le risque Cyber dans l'estimation du risque opérationnel.

4.3.2.5 Synthèse de l'application au BGS de l'ORSA

Cette section a porté sur l'inclusion du risque Opérationnel Cyber dans le calcul du capital économique évalué dans le cadre du Pilier 1 de Solvabilité 2. Il a tout d'abord été rappelé de manière précise la manière dont est évalué le risque opérationnel dans la Formule Standard, sous l'hypothèse de l'ajustement proposé.

Une proposition d'ajustement a ensuite été proposée, de manière simple et conforme à la fois à la logique du Pilier 1 et au risque mesuré. Ajuster la manière dont est évalué le capital réglementaire au profil de risque de sa propre entreprise est un exercice du Besoin Global de Solvabilité exigé par l'ORSA. Plus précisément, une transcription de la mesure du risque effectuée avec un modèle stochastique a été proposée, basée sur une formule simple et déterministe et prenant en compte les variables explicatives de l'entreprise considérée : le pays et la taille de l'entreprise exprimée en nombre de clients.

En se basant sur les entités fictives utilisées tout au long du mémoire, les montants de SCR bruts obtenus ont été transformés en montants nets en appliquant les conditions de transfert de risque. Cette étape a permis de montrer le fort impact des mesures de transfert de risque sur l'inclusion du risque Cyber dans le calcul de la Formule Standard.

Enfin, ces calculs effectués sur des entités fictives ont été confrontés à la réalité des chiffres et du marché, grâce à l'utilisation de cas réels. Pour cela, les SFCR (*Solvency Financial Condition Report*) de deux entreprises ont été explorés : La Mutuelle des Motards et Allianz Group. Cette étape a permis de montrer que la proposition d'ajustement de la Formule Standard impacterait plus fortement les compagnies d'assurance et mutuelles de taille modeste que celles de taille plus importante. Il a également été précisé que l'assureur de taille importante pris en exemple, Allianz Group, considère déjà le risque Cyber comme un risque à part entière et utilise déjà des solutions de transfert de risque auprès d'assureurs tiers.

L'inclusion du risque Opérationnel Cyber dans la Formule Standard, dans le cadre du BGS par exemple, permettrait une meilleure prise de conscience du risque Cyber porté par les entreprises et inciterait fortement les compagnies d'assurance et mutuelles à rechercher des solutions de transfert de risque et plus particulièrement celles de taille modeste ou moyenne.

De plus, cette proposition d'inclusion du risque Cyber ne se base que sur le risque de perte de DCP. Si d'autres risques étaient inclus (attaque par déni de service ou altération de données causée par un rançongiciel), il est possible que les grandes entreprises se retrouvent proportionnellement autant impactées que les plus petites car ces deux risques causent d'importantes pertes d'exploitation.

4.4 Constitution de scénarios, central et stressé, pour l'ORSA

L'évaluation présentée ci-dessus est basée sur la calibration statistique d'un modèle sur un historique de sinistres.

Cette approche a des limites, il convient de rappeler qu'il est nécessaire pour une entreprise de travailler à la fois sur l'aspect prévention que sur l'aspect résilience pour limiter la sévérité des incidents.

Dans le cadre de l'ORSA, il est important d'apporter un éclairage tant quantitatif que qualitatif de l'évaluation des risques de l'entreprise. La première sous-section de cette section apportera un éclairage qualitatif des bonnes pratiques d'atténutation des risques, les autres sections se concentreront sur la quantification du risque.

4.4.1 Prévention : Mesures d'atténuation des risques

Dans cette section, des éléments d'atténuation des risques seront évoquées, qui sont basées sur un document produit par le cabinet PwC : [Horne, 2017].

D'autres documents présentant des techniques et bonnes pratiques de gouvernance existent, par exemple la revue de l'Institut Français de l'Audit et de Contrôle Interne : [IFACI, 2018].

Les bonnes pratiques d'atténuation des risques Cyber sont :

— Une bonne compréhension de son exposition au risque Cyber, avec une mise à jour continue

— La constitution de ressources humaines et matérielles suffisantes et adaptées à son exposition

— Considérer une approche intégrée, plutôt que silotée, en explicitant les inter-connexions et dépendances possibles entre les points de vulnérabilité de son exposition

— Constituer des équipes de revue, de validation et de test indépendantes pour évaluer son système de protection

— Documenter en interne tous les incidents survenus et conserver ces traces afin de constituer une base de travail et expliciter ses pistes d'amélioration

— Considérer et évaluer les risques juridiques engendrés par son exposition Cyber

— Partager et échanger avec les autres entreprises sur ses propres incidents de telle sorte à faire progresser les bonnes pratiques de marché. Certaines mesures de publication, telles qu'existantes aux États-Unis, ont par exemple permis la constitution de la base VERIS et la réalisation de cette étude actuarielle.

4.4.2 Constitution d'une matrice de sévérité Cyber de référence

Dans le cadre de l'ORSA et de l'évaluation prospective des besoins en capital, des scénarios probables de type Cyber-attaques impliquant des pertes de données peuvent être définis. Contrairement à la proposition de modification de la Formule Standard, il n'est pas nécessaire ici de se baser sur un niveau de risque à 99,5 %.

A partir du modèle de sévérité retenu au chapitre 3, une matrice de sévérité de référence faisant le lien entre le nombre de données perdues et le coût peut être estimée. Dans cette section, le premier modèle estimant un nombre de données perdues n'est pas utilisé, car le nombre de données perdues est supposé être une hypothèse qui doit être faite dans le cadre de la définition des scénarios ORSA.

Classe	Nb données	Forme	mu	sigma	Moyenne	Médiane	Écart-type
A	100	Log-normale	11,40	2,287	1 226 266	89 707	16 717 824
B	1 000	Log-normale	12,19	2,287	2 692 510	196 969	36 707 294
C	10 000	Log-normale	12,98	2,287	5 911 939	432 484	80 598 133
D	100 000	Log-normale	13,76	2,287	12 980 834	949 604	176 969 160
E	1 000 000	Log-normale	14,55	2,287	28 501 991	2 085 043	388 570 835
F	10 000 000	Log-normale	15,34	2,287	62 581 765	4 578 125	853 184 212

TABLE 4.11 – Paramètres utilisés pour estimer la matrice de référence de sévérité (en $)

	Scénarios et coûts associés selon le quantile					
Quantile	A	B	C	D	E	F
10,0%	4 786	10 508	23 072	50 660	111 234	244 236
20,0%	13 089	28 739	63 102	138 554	304 222	667 979
30,0%	27 038	59 367	130 352	286 214	628 440	1 379 865
40,0%	50 257	110 348	242 291	531 998	1 168 107	2 564 811
50,0%	89 707	196 969	432 484	949 604	2 085 043	4 578 125
60,0%	160 124	351 584	771 973	1 695 020	3 721 751	8 171 841
70,0%	297 629	653 504	1 434 897	3 150 602	6 917 770	15 189 333
75,0%	419 516	921 131	2 022 527	4 440 858	9 750 782	21 409 773
80,0%	614 822	1 349 963	2 964 111	6 508 293	14 290 246	31 377 065
85,0%	959 937	2 107 732	4 627 944	10 161 566	22 311 731	48 989 824
90,0%	1 681 520	3 692 109	8 106 757	17 799 990	39 083 402	85 815 348
95,0%	3 859 624	8 474 568	18 607 592	40 856 654	89 708 873	196 973 596
96,0%	4 916 557	10 795 274	23 703 163	52 044 987	114 275 076	250 913 557
97,0%	6 620 457	14 536 525	31 917 818	70 081 891	153 878 668	337 871 087
98,0%	9 832 696	21 589 632	47 404 312	104 085 555	228 540 446	501 805 804
99,0%	18 341 147	40 271 621	88 424 319	194 153 101	426 301 580	936 029 535
99,5%	32 450 344	71 251 157	156 446 025	343 508 226	754 240 331	1 656 084 002
99,9%	105 230 242	231 053 833	507 324 443	1 113 931 275	2 445 856 693	5 370 362 695

TABLE 4.12 – Matrice de référence de sévérité

4.4.3 Estimation d'une fréquence centrale et stressée

La fréquence d'occurrence d'incidents Cyber-attaque sur les données dépend de la durée de la projection.

En première approche et sous hypothèse de l'utilisation d'une loi de Poisson, pour un plan d'**horizon 5 ans** et une espérance de fréquence de 0,42, un organisme d'assurance peut espérer 2,10 incidents sur la période du plan.

Un scénario moyen peut être par exemple la réalisation de 2 incidents, et 3 pour un scénario pessimiste.

Nb évènements (k)	$\mathbb{P}(X = k)$	$\mathbb{P}(X \leq k)$
0	0,122	0,122
1	0,257	0,379
2	0,270	0,649
3	0,189	0,838
4	0,099	0,937
5	0,042	0,979
6	0,015	0,994
7	0,004	0,998
8	0,001	0,999

TABLE 4.13 – Probabilité d'occurrence et loi cumulée pour la loi de Poisson de paramètre $2,1$

Dans le cadre de l'ORSA, l'objectif est de proposer des scénarios réalistes et probables d'occurrence d'incidents à l'horizon du plan. Étant donné l'incertitude beaucoup plus faible sur le nombre d'incidents

sous l'hypothèse retenue que sur la loi modélisant le coût d'un incident, la suite de l'étude portera essentiellement sur l'étude de la sévérité.

4.4.4 Les scénarios retenus selon les caractéristiques des entreprises

Une approche simple est pertinente pour retenir des hypothèses concernant le nombre de données perdues en cas de sinistres.

En effet, un petit acteur (qui gère par exemple 100 000 assurés) ou un acteur majeur du marché (par exemple 10 millions d'assurés) ne choisiront pas les mêmes scénarios pour l'ORSA.

Dans le cadre de l'ORSA, un organisme pourra choisir le nombre d'assurés du produit qu'il commercialise auprès du plus grand nombre afin de sélectionner un scénario, ou bien estimer le nombre d'assurés à retenir en fonction de comment sont stockées les informations clients.

4.4.5 Exemple de scénario central et scénario en environnement stressé

4.4.5.1 Scénario central

Dans cet exemple, un assureur de taille importante sera considéré, exerçant sur le marché français. Son chiffre d'affaire mondial est de 5 milliards d'euros.

Les chiffres de montant de pertes liées aux sinistres sont exprimés en $, simplement pour être en ligne avec la matrice de sévérité.

Pour le scénario central, l'hypothèse est faite que l'assureur doit faire face à deux incidents impliquant des pertes de données au cours de l'horizon du plan, qu'il définit à cinq ans dans son ORSA.

Le premier est lié à une mauvaise manipulation par un employé d'un document contenant une liste de sinistres et des informations personnelles des assurés.
L'essentiel des coûts concerne l'identification du problème et la mise en place d'actions correctives (suppression des copies du document). Enfin, des frais sont engagés pour éviter que l'incident ne se reproduise pour garantir l'anonymisation des données dès leur extraction du système d'information contenant les sinistres. Le coût de l'incident est estimé à 110 348 $.

Le second incident est lié à un dysfonctionnement du site internet de l'assureur. Il devient possible d'accéder à l'espace personnel de n'importe quel assuré sans rentrer le mot de passe, mais simplement en remplissant un formulaire de devis, puis en rentrant l'adresse email de ce dernier. Le site connecte alors directement l'individu ayant fait la demande de devis au compte personnel de l'assuré. Le dysfonctionnement est signalé par un particulier directement au service client de l'assureur, surpris du dysfonctionnement. Près de 100 000 assurés sont théoriquement impactés par le dysfonctionnement. Par mesure de sécurité, le site internet est coupé pendant 24h par les équipes IT, et un audit du site internet est engagé.

La perte pour l'assureur est évaluée à 600 000 $ pour l'audit du site et 1 million de dollars de perte d'exploitation pour les 24h d'indisponibilité du site internet. Des frais annexes (95 020 $) sont engagés pour la mise en place d'actions correctives et le signalement de l'incident aux autorités compétentes en prenant soin de souligner l'efficacité de la prise en charge de l'incident par l'assureur.

Année de projection	1	2	3	4	5
Nombre d'incidents	0	1	0	1	0
Nombre de données		B - 1000		D - 100 000	
Quantile de perte		40%		60%	
Coût en $		110 348		1 695 020	

TABLE 4.14 – Coût du **scénario central** en $

Au total, ce scénario contient deux sinistres sur cinq ans pour un coût total de 1 805 368 $.

4.4.5.2 Scénario stressé

Dans le scénario stressé, le même premier incident est pris en compte.

L'hypothèse qu'un incident supplémentaire survient au cours de la troisième année de projection du plan est retenue.

Un projet en cours au sein de l'entreprise a pour objectif d'étudier le lien existant entre la qualité du sommeil, mesurée grâce à un bandeau connecté porté la nuit par des utilisateurs volontaires, pour mieux tarifer une offre de complémentaire santé. A la suite d'un vol physique d'un ordinateur portable, les données de sommeil d'une centaine d'utilisateurs sont récupérées par un acteur malveillant qui prend contact avec l'assureur et exige une rançon de 200 000 $ en crypto-monnaie pour faire disparaître les données compromettantes. L'assureur accepte et paie la rançon. D'autres frais, à hauteur de 151 584 $ sont engagés pour mettre en place des politiques de stockage des données sensibles à distance et sécurisées.

Enfin, l'incident de la quatrième année du scénario standard est modifié. L'hypothèse que le dysfonctionnement remarqué par le particulier n'est pas signalé directement au service client mais est partagé sur un réseau social est retenue. Le service IT ne coupe le site internet qu'au bout de 12h, et plusieurs milliers d'assurés voient leurs données personnelles récupérées par une organisation malveillante. La réputation de l'assureur est largement entachée, l'affaire étant abondamment relayée dans les médias. Un procès est mené contre l'assureur par une association de consommateurs, assortie d'une amende de 0,6 % du chiffre d'affaire mondial de l'assureur suite à une violation du RGPD.

L'assureur estime ses pertes à 10m $ de perte d'exploitation sur les deux années suivants l'incident à cause de la dégradation de son image, 500 000 $ suite au procès de l'association de consommateurs, 30m $ d'amende RPGD, et 356 654 $ de frais annexes.

Année de projection	1	2	3	4	5
Nombre d'incidents	0	1	1	1	0
Nombre de données		B - 1000	C - 100	D - 100 000	
Quantile de perte		40%	60%	95%	
Coût en $		110 348	351 584	40 856 654	

TABLE 4.15 – Coût du **scénario stressé** en $

Au total, ce scénario contient trois sinistres sur cinq ans pour un coût total de 41 318 586 $.

4.4.5.3 Analyse des scénarios central et stressé

A partir de la table de sévérité de référence et d'hypothèses simples, un scénario central probable et un scénario stressé probable ont été construits.
Le rapport entre les coûts sur cinq ans est important : le scénario stressé coûte plus de vingt fois plus cher que le scénario central, l'essentiel du coût provenant d'un seul sinistre.

Cet exemple simple et fictif illustre parfaitement la nature du risque Cyber : le risque Cyber de type perte de Données à Caractère Personnel est un risque de pointe, c'est-à-dire que certains sinistres, peu fréquents, sont susceptibles d'engendrer des coûts importants.

Pour ce type de risque, à la fois assez fréquent et avec quelques sinistres pouvant coûter très cher, la mesure de transfert la plus adaptée est la souscription d'une police avec une limite élevée mais avec une franchise assez élevée également afin d'éviter que la prime d'assurance soit trop élevée.

Dans le cas d'un assureur qui souhaite couvrir un portefeuille composé de polices Cyber, les couvertures les plus adaptées sont les couvertures non-proportionnelles de type Excédent de Sinistre (*Excess of Loss - XL*) ou Excédent de Plein (*Stop Loss - SL*).

En 2020, ni le marché français de l'assurance Responsabilité Civile Professionnelle ni le marché français de la réassurance n'est majoritairement composé de ce type de couverture Cyber, à cause du manque de connaissance de ce risque.

En particulier, le marché de la réassurance est principalement composé de couvertures proportionnelles. Il est à noter que le risque Cyber des particuliers est très inférieur au risque Cyber des entreprises car elles seules sont susceptibles de subir des pertes pécuniaires conséquentes.

Conclusion

Le risque Cyber est un des risques émergents les plus importants de la décennie qui s'ouvre, d'après la FFA, avec le risque de croissance des inégalités et de tensions sociales. Cette étude utilise deux bases publiques et collaboratives qui recensent des évènements de pertes de Données à Caractère Personnel aux États-Unis en proposant des analyses comparatives et en mettant ces bases en perspective avec des rapports déjà publiés par d'autres organisations.

Les analyses comparatives ont tout d'abord permis de montrer le manque de pertinence de l'utilisation tant de la base PRC que de la base VERIS dans le cadre d'études de la fréquence à cause des limites trop importantes sur la manière dont est collectée la donnée.

Pour l'étude de la sévérité, la mesure de l'incertitude entre le lien (nombre de données perdues - coût financier) est essentielle car elle permet de différencier la vision médiane et la vision moyenne du coût d'un incident Cyber de type perte de données.

Comme seule la base VERIS contient la variable coût financier, elle permet la construction d'un modèle stochastique de sévérité que la base PRC ne permet pas et ouvre le champ des applications actuarielles possibles avec entre autres des modéles de tarification de contrats d'assurance et de réassurance, étude du capital économique dans le cadre de l'ORSA, suivi de l'appétence au risque d'un assureur.
Le modèle stochastique confirme les observations de marché, et notamment le caractère très volatil de la sévérité, mais permet néanmoins de proposer une approche pragmatique et contrôlée pour la constitution des scénarios central et en vision stressée pour l'évalutation du Besoin Global de Solvabilité exigé par l'ORSA.

Plusieurs applications quantitatives ont été développées dans ce mémoire et appliquées à des sociétés fictives. Tout d'abord, l'estimation de quantiles au niveau de risque défini par la réglementation Solvabilité 2 : 99,5 %. Cette application a mis en évidence l'impact des variables explicatives retenues : pays et taille d'entreprise exprimée en nombre d'assurés.
Puis l'application de la tarification de couvertures non-proportionnelles a été mise en oeuvre. Cette application est possible par la modélisation stochastique de la perte économique à laquelle des couvertures sont appliquées. Cette application a montré le gain en capital économique qu'il est possible d'obtenir grâce à ce type de couverture.
Plus particulièrement, tant que la Formule Standard ne prend pas en compte ce risque opérationnel Cyber, il n'est pas utile pour les entreprises de recourir à un transfert de risque lorsque l'on considère uniquement le rapport capital économique / coût de la couverture.
C'est pour cette raison que la troisième application a porté sur une proposition de la modification de la Formule Standard afin d'inclure le risque opérationnel Cyber. Cet exercice doit également être effectué dans le cadre du BGS de l'ORSA. La mise en pratique de cette modification impacterait principalement les petits et moyens acteurs. De plus, cette modification permettrait un changement de considération

de ce risque émergent : les petits et moyens acteurs seraient incités à considérer une stratégie de prévention mais également de résilience pour faire face à une menace Cyber, qui passerait par le recours à l'assurance du risque vis-à-vis de tiers.
Le régulateur a donc un rôle actif à jouer pour promouvoir la bonne gestion du risque Cyber et pour permettre l'essor de l'activité d'assurance.
La dernière application porte sur la constitution de scénarios Cyber pour l'ORSA. L'un des objectifs de cette application est de rendre concret les travaux précédents, en se concentrant via des exemples sur les différents postes de dépenses en cas de sinistres.

La modélisation sur laquelle repose ce mémoire possède des limites telle que l'évolution extrêmement rapide des composantes du coût d'un sinistre Cyber : suite à l'entrée en vigueur du RGPD en Europe le 25 mai 2018, ce modèle sous-estime-t-il les pertes juridiques potentielles que le régulateur est capable d'infliger ?
De plus, les données utilisées pourraient être complétées d'autres études pour inclure le secteur d'activité et la taille de l'entreprise dans la modélisation stochastique.
En effet, la partie stochastique du modèle n'utilise qu'une seule variable spécifique et caractéristique de la Cyber-vulnérabilité : l'exposition maximale exprimée en nombre d'assurés. Compte tenu des très importantes limites de qualité de données, il n'a pas été possible d'inclure d'autres variables.

Enfin, il convient de rappeler l'importance des mesures d'atténuation des risques que chaque acteur économique, entreprise comme institution publique, doit mettre en œuvre pour la protection de ses intérêts et la protection de la vie privée des particuliers. Il est également nécessaire que la puissance publique, européenne ou nationale, prévoit des mesures efficaces de partage d'informations des incidents subis.

Les actuaires ont leur rôle à jouer dans l'étoffement de la connaissance du risque Cyber, en enrichissant la connaissance des experts métiers par des approches statistiques et en constituant des opportunités de transfert de risque maîtrisées et adaptées à chaque acteur économique et en incitant au partage des informations entre les assureurs à l'échelle nationale et européenne.

Bibliographie

[AllianzGroup,] AllianzGroup. Rapport sur la solvabilité et la situation financière.

[Bessy-Roland et Boumezoued, 2019] Bessy-Roland, Y. et Boumezoued, A. (2019). Modélisation stochastique individuelle de sinistres cyber.

[CESIN, 2019] CESIN (2019). Baromètre de la cyber-sécurité des entreprises.

[Delignette-Muller et Dutang, 2015] Delignette-Muller, M. L. et Dutang, C. (2015). fitdistrplus : An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34.

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

[Farkas *et al.*, 2019] Farkas, S., Lopez, O. et Thomas, M. (2019). Cyber claim analysis through generalized pareto regression trees with applications to insurance pricing and reserving.

[FFA, 2019] FFA (2019). Baromètre 2019 des risques émergents.

[Haverland, 2018] Haverland, A. (2018). Chez saint gobain, "il y a un avant et un après la cyber-attaque", `https://www.usinenouvelle.com/editorial/chez-saint-gobain-il-y-un-avant-et-un-apres-la-cyber-attaque.N651134`.

[Horne, 2017] Horne, R. (2017). Governing cyber security risk : It's time to take it seriously.

[IFACI, 2018] IFACI (2018). Cyber-risques : Enjeux, approaches et gouvernance.

[Inc., 2016] Inc., R. M. S. (2016). Managing cyber insurance accumulation risk ; report prepared in collaboration with and based on original research by the centre for risk studies, university of cambridge.

[Jacobs, 2014] Jacobs, J. (2014). Analyzing ponemon cost of a data breach.

[LaMutuelleDesMotards, 2018] LaMutuelleDesMotards (2018). Rapport annuel.

[LaMutuelleDesMotards, 2019] LaMutuelleDesMotards (2019). Rapport sur la solvabilité et la situation financière.

[Ponemon et IBM, 2017] Ponemon et IBM (2017). Global cost of a data breach report.

[Ponemon et IBM, 2018] Ponemon et IBM (2018). Global cost of a data breach report.

[Ponemon et IBM, 2019] Ponemon et IBM (2019). Global cost of a data breach report.

[R Core Team, 2019] R Core Team (2019). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Verizon, 2015] Verizon (2015). Data breach investigations report.

[Wikipedia,] Wikipedia. Morris worm, `https://en.wikipedia.org/wiki/Morris_worm`.

Table des figures

Annexe A

Exploration de la base VERIS et retraitements

A.1 Description Générale

Afin de passer d'une base de 2441 colonnes à une base plus facile à manipuler, la première idée naïve est de regrouper les énumérations, or cela est délicat car une énumération peut avoir plusieurs modalités associées. Le risque est d'altérer la donnée.

Ainsi, la première étape du retraitement sera de classer chaque champ (c'est-à-dire chaque énumération et chaque variable) selon l'utilisation potentielle et sous contrainte de ne pas altérer la donnée. Les 2441 colonnes correspondent à 191 champs.

Dans cette première étape, les champs sont classés dans les catégories suivantes : Classification, Clustering – évènements d'accumulation, Filtre et qualité des données, variable quantitative, Autres / Pas pour des statistiques.

Cette première classification (figure A.1) est effectuée de manière générale, c'est-à-dire indépendamment de la considération du risque Cyber.
On s'aperçoit que beaucoup de champs sont des champs de classification, et que très peu sont des champs contenant des variables quantitatives.

Les variables quantitatives sont essentielles. Les neuf champs sont représentés figure A.2.
Note : Le champ indiquant la devise a été classé comme quantitatif car directement et exclusivement lié à la variable contenant le montant de perte. Dès cette étape, les champs quantitatifs sont identifiés :

— La taille de l'entreprise (exprimée en nombre d'employés)

Utilisation_potentielle	# champs
Classification	101
Clustering - évènements d'accumulation	5
Filtrage - Qualité des données	2
Variable quantitative	9
Autres / Pas pour des statistiques	74
Total	191

Table A.1 – Première classification des champs de la base VERIS

keyname	format	notes_Thomas
victim.orgsize	enum	can be useful ; attention to coherence with victim.employee_count
victim.revenue.iso_currency_code	**enum**	**Information about the size of the company: revenue (=~ turnover)**
victim.revenue.amount	integer	Information about the size of the company: revenue (=~ turnover)
victim.employee_count	**free text**	**# of employees of the victim organisation**
impact.overall_amount	float	Best field to represent the financial loss
impact.loss.amount	**float**	**Not filled**
impact.iso_currency_code	enum	Currency of the reported impact figures
attribute.confidentiality.data_total	**integer**	**Attention to double counting and coherence with attribute.confidentiality.data.amount**
attribute.confidentiality.data.amount	integer	# affected records per type of data

TABLE A.2 – Les neuf champs de variables quantitatives

Utilisation_potentielle	Intérêt étude Cyber	# champs
Classification	**Trop spécifique ou non utilisable**	**72**
	Grille A4	**23**
	Champ de date	**4**
	Champ de lieu	**1**
	Champ de secteur	**1**
Clustering - évènements d'accumulation	**Clustering - évènements d'accumulation**	**2**
	Trop spécifique ou non utilisable	**2**
	Nécessite un retraitement	**1**
Filtrage - Qualité des données	**Trop spécifique ou non utilisable**	**1**
	Qualité des données	**1**
Variable quantitative	**Variable quantitative**	**7**
	Trop spécifique ou non utilisable	**2**
Autres / Pas pour des statistiques	**Trop spécifique ou non utilisable**	**73**
	Référence externe (ex: journaux)	**1**
Total		191

TABLE A.3 – Deuxième classification des champs de la base VERIS

— Le chiffre d'affaires
— Le montant de perte associé à un incident
— Le nombre de données perdues

A.2 Description spécifique à l'étude Cyber

Lors de cette étape (figure A.3), une colonne supplémentaire est ajoutée de telle sorte à indiquer l'intérêt du champ pour une étude Cyber. Là encore, aucune altération de la donnée n'est tolérée.

Lors de cette étape, une attention particulière a été prêtée aux champs permettant d'identifier des incidents comme ayant une cause commune, c'est-à-dire issue d'évènements d'accumulation. Cette notion d'évènements d'accumulation est notamment décrite dans la revue RMS Cambridge – Managing Cyber Insurance Accumulation Risk. Cependant, l'information disponible n'est pas suffisamment intéressante, et concerne tellement peu d'incidents qu'elle n'est pas exploitable.

Concernant les nombreux champs dont l'utilisation potentielle est « Classification » et dont l'intérêt pour une étude Cyber est « Trop spécifique ou inutilisable », cela correspond à un niveau plus précis de la grille A^4 : la grille A^4 contient quatre A composés de respectivement 3, 7, 7 et 3 modalités soit 441 combinaisons. Or l'information de chaque modalité de A est complétée par plusieurs autres champs, généralement des énumérations. Les champs liés à la modalité *Hacking* de *Action* sont représentés figure A.4.

Cela semble un nombre raisonnable de champs supplémentaires pour l'étude Cyber. Or tous ces

keyname	format
action.hacking.result	enum
action.hacking.variety	enum
action.hacking.vector	enum

TABLE A.4 – Exemple de champs liés à une modalité de la grille A^4

champs sont des énumérations : chacune d'entre elles peut prendre un nombre très important de valeurs tel que présenté figure A.5.

Etant donné le faible nombre d'incidents disponibles pour lesquels certaines variables quantitatives sont renseignées, ces champs de classification ont été considérés trop spécifiques pour pouvoir être utilisés.

A.3 Sélection des champs pertinents et retraitements pour l'étude Cyber

Cette étape consiste à ne conserver que les champs potentiellement pertinents pour l'étude actuarielle du risque Cyber et à retraiter lorsque c'est nécessaire l'information contenue dans certains de ces champs afin d'obtenir la base de travail.

La base de travail contient 20 champs (figure 2.5). Les champs les plus retraités sont documentés dans la suite du document.

A.4 Principaux retraitements

Les champs les plus retraités sont *Malware.name*, *Actor.external.name, Size1* et *Size 2*. De plus une attention particulière sera portée au choix de la variable originale pour constituer number_records.

A.4.1 Malware.name

Pour créer le champ *Malware.name*, l'information contenue dans les champs originaux *campaign.id* et *action.malware.name* a été utilisée. Ce champ a été créé dans la perspective de créer un flag permettant de lier plusieurs incidents au même évènement.

Campaign.id contient des IDs uniques, sauf si la personne ayant alimenté la base a spécifiquement choisi de lier l'incident reporté à un autre incident déjà existant. Il existe une table de correspondance (figure A.1) sur github : `https://github.com/vz-risk/VCDB/blob/master/campaigns.md`.

Le champ *Action.malware.name* contient des noms de malware. La répartition des modalités sur les 8198 incidents est représenté figure A.2.
Étant donné que les informations contenues dans les deux champs sont *a priori* similaires et sont censés représenter la même information, les informations ont été croisées de telle sorte à créer la variable unique *Malware.name*, synthétisant et corrige si besoin l'information contenue dans les deux champs.

Une table en deux dimensions permet d'avoir une vue synthétique de l'information. Pour des raisons pratiques, les modalités de la variable *campaign_id* présentées en figure A.7 ont été remplacées

action.hacking.result.Elevate
action.hacking.result.Exfiltrate
action.hacking.result.Infiltrate
action.hacking.variety.Abuse.of.functionality
action.hacking.variety.Brute.force
action.hacking.variety.Buffer.overflow
action.hacking.variety.Cache.poisoning
action.hacking.variety.Cryptanalysis
action.hacking.variety.CSRF
action.hacking.variety.DoS
action.hacking.variety.Exploit.misconfig
action.hacking.variety.Exploit.vuln
action.hacking.variety.Footprinting
action.hacking.variety.Forced.browsing
action.hacking.variety.Format.string.attack
action.hacking.variety.Fuzz.testing
action.hacking.variety.HTTP.request.smuggling
action.hacking.variety.HTTP.request.splitting
action.hacking.variety.HTTP.response.smuggling
action.hacking.variety.HTTP.Response.Splitting
action.hacking.variety.Integer.overflows
action.hacking.variety.LDAP.injection
action.hacking.variety.Mail.command.injection
action.hacking.variety.MitM
action.hacking.variety.Null.byte.injection
action.hacking.variety.Offline.cracking
action.hacking.variety.OS.commanding
action.hacking.variety.Other
action.hacking.variety.Pass.the.hash
action.hacking.variety.Path.traversal
action.hacking.variety.Reverse.engineering
action.hacking.variety.RFI
action.hacking.variety.Routing.detour
action.hacking.variety.Session.fixation
action.hacking.variety.Session.prediction
action.hacking.variety.Session.replay
action.hacking.variety.Soap.array.abuse
action.hacking.variety.Special.element.injection
action.hacking.variety.SQLi
action.hacking.variety.SSI.injection
action.hacking.variety.Unknown
action.hacking.variety.URL.redirector.abuse
action.hacking.variety.Use.of.backdoor.or.C2
action.hacking.variety.Use.of.stolen.creds
action.hacking.variety.Virtual.machine.escape
action.hacking.variety.XML.attribute.blowup
action.hacking.variety.XML.entity.expansion
action.hacking.variety.XML.external.entities
action.hacking.variety.XML.injection
action.hacking.variety.XPath.injection
action.hacking.variety.XQuery.injection
action.hacking.variety.XSS
action.hacking.vector.3rd.party.desktop
action.hacking.vector.Backdoor.or.C2
action.hacking.vector.Command.shell
action.hacking.vector.Desktop.sharing
action.hacking.vector.Desktop.sharing.software
action.hacking.vector.Other
action.hacking.vector.Partner
action.hacking.vector.Physical.access
action.hacking.vector.Unknown
action.hacking.vector.VPN
action.hacking.vector.Web.application

TABLE A.5 – Modalités possibles pour un "A" de la grille A^4

Champ_retraité	Champ_original	Modifications / retraitements nécessaires	Usage possible
incident_id	incident_id	Aucun	Certains enregistrements (5 en tout) sont dupliqués, ce champ permet de retraiter
plus.github	plus.github	Aucun	Pas de signification explicite
reference	reference	Aucun	Lien vers des articles de journaux en ligne
incident_year	timeline.incident.year	Aucun	Date
incident_month	timeline.incident.month	Aucun	Date
incident_day	timeline.incident.day	Aucun	Date
victim.state	victim.state	Aucun	Information géographique, pertinent pour les USA seulement
Malware.name	campaign_id & action.malware.name	Retraitement sur-mesure. **Documenté spécifiquement**	Flag liés à la notion d'évènement d'accumulation
Actor.external.name	actor.external.name	L'information n'est conservée que si elle concerne au moins deux enregistrements. **Documenté spécifiquement**	Flag liés à la notion d'évènement d'accumulation
Actor	4 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Action	8 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Asset	8 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Attribute	3 colonnes	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Grille A4
Discovery_method	discovery_method enumeration	Si il existe de l'information dans une seule colonne on reporte le nom de la colonne, sinon "NA" ou "Collusion" suivant les cas où pas d'information ou information dans plusieurs colonnes	Ressemble à la grille A4
Size1	victim.employee_count	Size1 ne peut prendre que les modalités: "Small" ; "Large" or NA ; si l'information originale est "Small" ou dans une catégorie avec moins de 1000 employés alors "Small" ; si "Large" ou une catégorie à plus de 1000 employés alors "Large", sinon NA. **Documenté spécifiquement**	Classifie selon la taille de l'entreprise (en nb employés)
Size2	victim.employee_count	Reprend la même information que l'information initiale. Si la catégorie détaillé n'est pas disponible (si l'information originale est "Small", "Large" ou NA), alors "Unknown". Par conséquent, Size2 contient plus de NA que Size1.	Classifie selon la taille de l'entreprise (en nb employés)
sector	victim.industry.name	Aucun	Classifie selon le secteur de l'entreprise
number_records	attribute.confidentiality.data_total	Aucun	Information quantitative
revenue_USD	victim.revenue.amount & victim.revenue.iso_currency_code	Le montant en monnaie originale est simplement converti en USD (utilisant du taux FY18).	Information quantitative
loss_USD	impact.overall_amount & impact.iso_currency_code	Le montant en monnaie originale est simplement converti en USD (utilisant du taux FY18).	Information quantitative

TABLE A.6 – Les 20 champs de la base de travail issue de la base VERIS

This document is used for tracking the campaign_id field in VCDB incidents. When a series of incidents are linked together use one of the open UUIDs below to fill in the campaign_id field of the incidents. You'll need to edit this document with a description to show that a particular campaign ID has been taken.

Campaign ID	Description
CBAA666B-8213-4F6B-B6AC-29896A9C1455	Red October,Campaign from 2013
0A74F3BC-4844-4D5F-A350-017A9C81B464	Miniduke, Campaign from 2013
B91B10AA-1441-4C51-920F-3BD73AF6F4F0	MongoHQ breach and,follow up attack from 2013
F94E45F1-175C-48DD-A658-5F5209D43AFA	Operation,DeputyDog; FireEye 2013
84F70E33-3D69-4180-A4CC-E8FF7A8C0EC4	Sunshop Campaign;,FireEye 2013
06C2A8F8-76D4-4BC2-8C77-CB563EE1F3D8	SSNDOB outed by Krebsonline 9/25/2013
3EC75F4E-600F-498A-B0D8-39BFDE21B9AC	GCHQ hacking of Belgacom,GreenNet,Riseup
D9A0A1DC-6201-48A2-B270-E7C6EF530D13	Operation Aurora, 2009
104874B4-3EC7-4B09-95F1-930F007487B0	Operation Poisoned Hurricane, 2014
09A354EC-4E75-48A4-94B9-2E64D9D1DF15	Operation Snowman
CF78215A-6145-4556-B2C1-FD79D43BA6CD	Operation Troy
8F544E52-B6EA-4301-975C-4D6E455CB0EE	
C45C59CD-A3AA-4DDC-8288-A7C4E3E3C688	
657A1D71-EB83-438D-9AB9-FFA065A9313A	

FIGURE A.1 – Table de correspondance pour la variable Campaing.id

```
mmary(33)
                  Aurora;Hydraq             Backoff          BitPaymer Bitpaymer ransomware         Citadel
            32               11                   1                  1                    1               1
  cryptolocker      CryptoLocker          cryptowall         Cryptowall       GrayPigeon RAT           Havex
            11                1                   1                  1                    2               1
         hikit            HiKit  JavaScript malware     Kaba;PlugX;SOGU             KeyRaider        MiniDuke
             1                1                   1                  5                    1              60
            NA          nbc.exe         Red October              Rocra               Samsam         Shamoon
             0                2                   1                 88                    1               1
Terminator RAT          unknown            WannaCry           YPVo.jar    zeus,cryptolocker         ZxShell
             1                1                   1                  1                    1               1
          NA's
          7966
```

FIGURE A.2 – Vue d'ensemble de la variable Action.malware.name

	A	B	C	D	E	F	G	H	I	J	NA	Total
Rocra	88											88
MiniDuke		60										60
Aurora;Hydraq			11									11
CryptoLocker											12	12
Kaba;PlugX;SOGU					5							5
NA	1			7	1	5	3	2	1	2		22
Backoff											1	1
BitPaymer											1	1
Bitpaymer ransomware											1	1
Citadel											1	1
Cryptowall											2	2
GrayPigeon RAT											2	2
Havex											1	1
HiKit									1		1	2
JavaScript malware											1	1
KeyRaider											1	1
nbc.exe											2	2
Red October											1	1
Samsam											1	1
Shamoon											1	1
Terminator RAT											1	1
unknown											1	1
WannaCry											1	1
YPVo.jar											1	1
zeus,cryptolocker											1	1
ZxShell										1		1
Total	**89**	**60**	**11**	**7**	**6**	**5**	**3**	**3**	**2**	**2**	**34**	**222**

TABLE A.7 – Tableau croisant l'information des deux champs similaires

par des lettres, de A à J.

Pour certains évènements, l'information est plutôt cohérente : par exemple « Rocra » correspond presque parfaitement à « 1 – Red October, Campaign from 2013 ».

Pour d'autres évènements, l'information a été reportée dans un seul des deux champs : "CryptoLocker" est uniquement présent dans le champ *Action.malware.name* tandis que l'évènement « D – GCHQ hacking of Belgacom,GreenNet,Riseup » apparaît seulement dans le champ *campaing_id*.

A partir de ce tableau et dans la perspective d'une utilisation pour identifier des évènements d'accumulation, les évènements contenant trop peu d'incidents ont été exclus des modalités possibles. **Seuls les évènements contenant au moins 5 incidents ont été retenus**. De plus, pour les cas ambigus (par exemple le 89ième incident « Red October ») le choix d'inclure ou non dans la variable synthétique finale a été fait au cas par cas, en fonction des autres informations disponibles.

La variable finale *Malware.name* lie 189 incidents à des évènements, sur un total de 8 198 incidents, soit environ 2% de la base et représenté en figure A.3.

A.4.2 Actor.external.name

Ce champ contient du texte, sans aucune contrainte lors de la saisie de l'incident. La plupart des modalités n'apparaissent qu'une seule fois dans la base. Etant donné que l'on prévoit d'utiliser ce champ spécifiquement pour l'identification d'évènements d'accumulation, une revue des différentes modalités

```
> summary(df$Malware.name)
                        Aurora;Hydraq                                CryptoLocker GCHQ hacking Belgacom;GreenNet;RiseUp
                                   11                                           12                                    7
                             MiniDuke                             Poisoned Hurricane                                Rocra
                                   60                                            6                                   88
                                Troy                                          NA's
                                    5                                          8009
```

FIGURE A.3 – Modalités de la variable Malware.name

	#
TheDarkOverlord	4
The Dark Overlord	3
TheDarkOverlord (TDO)	1
Dark Overlord	1
The DarkOverlord	1
Total	**10**

TABLE A.8 – Les modalités Dark Overlord du champ Actor.external.name

« proches » a été menée, en :

— Corrigeant et harmonisant les entrées texte si nécessaire
— Remplaçant les modalités n'apparaissant qu'une fois par NA

Par exemple, ces modalités peuvent être trouvées dans la base (figure A.8).

Sur cet exemple, toutes les modalités ont été remplacées par « The Dark Overlord ». Une fois tous les retraitements effectués, la variable *Actor.external.name* contient de l'information pour uniquement 51 incidents sur 8 198, soit moins de 1% de l'ensemble des incidents, représenté en figure A.4.

A.4.3 Size1 et Size2

L'information de la taille de l'entreprise, exprimée en nombre d'employés, est représentée par deux champs originaux de la base VCDB :

— *victim.orgsize*
— *victim.employee_ count*

Une analyse fine de la cohérence des informations présentes dans ces champs a été menée. Il en est ressorti que l'information contenue dans *victim.orgsize* est de bien moins bonne qualité que *victim.employee.count*. Dès lors, seul le champ *victim.employee_ count* a été utilisé.

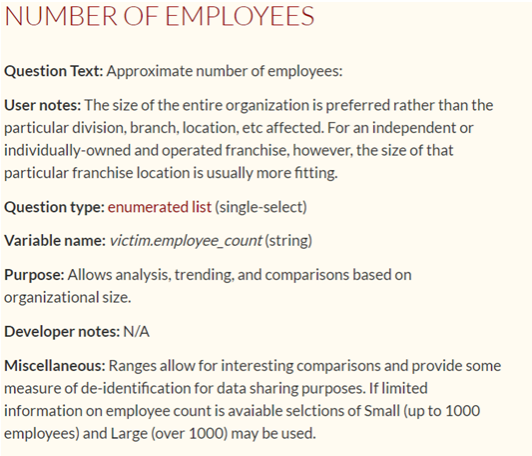
Les modalités de *victim.employee_ count* (les modalités ont été renommées mais non retraitées, par

```
> summary(df$Actor.external.name)
6A323CB8-6DC7-433C-A0F7-3D4850194425                    Anonymous                Cutting Sword of Justice
                                 11                           13                                       2
                        Guccifer 2.0                    KDMS Team                             Lauri Love
                                  2                            2                                       3
                      Pinoy Anonymouz                   Rex Mundi                Syrian Electronic Army
                                  2                            3                                       3
                    The Dark Overlord                        NA's
                                 10                         8147
```

FIGURE A.4 – Modalités de la variable Actor.external.name

```
"size.A.1.10"     "size.B.11.100"   "size.C.101.1000" "size.D.1001.10k" "size.E.10k.25k"  "size.F.25k.50k"  "size.G.50k.100k"
"size.H.100k+"    "size.Large"      "size.Small"      "size.Unknown"
```

FIGURE A.5 – Modalités de la variable originale victim.employee.count



NUMBER OF EMPLOYEES

Question Text: Approximate number of employees:

User notes: The size of the entire organization is preferred rather than the particular division, branch, location, etc affected. For an independent or individually-owned and operated franchise, however, the size of that particular franchise location is usually more fitting.

Question type: enumerated list (single-select)

Variable name: *victim.employee_count* (string)

Purpose: Allows analysis, trending, and comparisons based on organizational size.

Developer notes: N/A

Miscellaneous: Ranges allow for interesting comparisons and provide some measure of de-identification for data sharing purposes. If limited information on employee count is avaiable selctions of Small (up to 1000 employees) and Large (over 1000) may be used.

FIGURE A.6 – Information du site VERIS à propos des modalités "Small" et "Large"

exemple il n'y a pas eu de regroupement de modalités) sont représentées figure A.5.

Il y a donc deux types de modalités distincts :

— les modalités détaillées : dont le nom est du type « Size.A.[. . .] à Size.F[. . .] »
— les modalités de haut niveau : « Size.Large et Size.Small »

En recherchant la signification des modalités de haut niveau sur le site VERIS, le seuil entre *Small* et *Large* est de 1000 employés, représenté en figure A.6.

Afin d'altérer le moins possible l'information disponible mais tenant compte du fait qu'elle est sous sa forme originale impossible à utiliser directement, les deux variables Size1 et Size2 ont été créées à partir de l'information originale, tel que représenté figure A.9.

L'information est disponible pour 2602 incidents sur les 8198 présents dans la base.

Information originale	Size1	Size2	#
size.A.1.10	Small	A.1.10	260
size.B.11.100	Small	B.11.100	323
size.C.101.1000	Small	C.101.1000	454
size.D.1001.10k	Large	D.1001.10k	627
size.E.10k.25k	Large	E.10k.25k	201
size.F.25k.50k	Large	F.25k.50k	86
size.G.50k.100k	Large	G.50k.100k	49
size.H.100k+	Large	H.100k+	339
size.Large	Large	NA	108
size.Small	Small	NA	155
size.Unknown	NA	NA	excluded

TABLE A.9 – Table de correspondance entre l'information originale et les variables Size1 et Size2

FIGURE A.7 – Les colonnes de l'énumération attribute.confidentiality.data.amount

A.4.4 Number_records

La variable quantitative *number_ records* est une des variables les plus utiles de la base VCDB. Pour la constituer, deux champs possèdent de l'information utilisable, l'énumération *attribute.confidentiality.data.amount* et la variable *attribute.confidentiality.data_ total* (figure A.7).

Toutes ces colonnes contiennent des valeurs numériques. Afin d'estimer la cohérence entre les deux champs, la somme des colonnes de l'énumération a été comparée à la valeur de l'autre variable.

Tout d'abord, 4448 incidents ne possèdent aucune information concernant le nombre de données perdues.

2167 incidents sont cohérents entre les deux champs, et le reste soit 1583 sont incohérents. Étant donné que le champ total est plus fiable et que pour certains incidents, le même chiffre est reporté dans plusieurs colonnes de l'énumération ainsi que dans le champ total, le champ *attribute.confidentiality.data_ total* est retenu (figure A.8).

attribute.confidentiality.data_total	sum_enum_field
2	0
14829	0
1	0
715	1430
7	0
17	0
3	0
2000000	0
16154	0
13000	0
80	0
83000	0
61	0
106	0
14	0
55	0
1	0
1	0
199	0
1	0
1	0
115	0
1	0
16000	0
121	179
42000000	84000000

FIGURE A.8 – Quelques incohérences entre les deux champs précisant le nombre de données perdues

Annexe B

Correspondance entre les incidents des deux bases

Dans ce chapitre, la question de savoir si les deux bases reportent des enregistrements relatifs aux mêmes attques sera traitée. Les deux bases contiennent en effet des nombres d'enregistrements similaires, sont relatives à des attaques de type pertes de données ayant eu lieu pour la majorité sur le territoire des États-Unis, ou concernant des entreprises américaines.

Il existe peu de champs permettant d'identifier un incident dans la base PRC. Ces champs sont la date de publication et le nom de l'entreprise. La base VERIS contient deux types de champs de date : la date de reporting et la date de survenance.

B.1 Correspondance entre les fréquences

Les fréquences comparées entre les deux bases sont reportées en figure B.1.

Dans le cas où les deux bases avaient été alimentées par les mêmes incidents mais avec une méthodologie différente, une bien meilleure correspondance entre l'année de publication de PRC avec l'année de survenance de VERIS aurait été attendue. De plus la publication des incidents (dans la base PRC) ne semble pas pouvoir être expliquée par un décalage des données de survenance (base VERIS). La colonne année de Reporting de VERIS a été ajoutée au tableau au cas où les termes « reporting » et « publication » auraient été mal interprétés. Aucune correspondance entre les bases n'est observable.

B.2 Corrrespondance entre les entreprises

La figure B.1 est, en première approche, un moyen de comparer le nombre d'occurrences de différentes entreprises (ou institutions publiques) dans la base.

La correspondance est délicate car aucune entreprise ne se retrouve à la fois parmi les entreprises les plus présentes dans la base PRC et dans la base VERIS.

Étant donné que la saisie est libre, certaines entreprises peuvent être mal comptabilisées.

En deuxième approche, une entreprise parmi les plus présentes dans la base PRC sera recherchée dans la base VERIS et inversement. Une attention particulière est portée de telle sorte à bien capter

Base PRC	
Entreprise	Nb occurrence
Walgreen Co.	12
University of Florida	11
Private Medical Practice	10
Purdue University	10
Bank of America	9
Experian	9
Mount Sinai Medical Center	9
AT&T	8
Henry Ford Health System	8
Aflac	7
Florida Department of Health	7
Healthand hospital corporationof marion county	7
McDonald's	7
Private Dental Practice	7
University of Virginia	7
Columbia University	6
Cornell University	6
Stanford University	6
Texas A&M University	6
Yale University	6
Walgreens	6
Allina Health	5
Aetna Inc.	5
CareFirst BlueCross BlueShield	5
CVS Health	5
Children's Mercy Hospital	5
CVS Caremark	5
California Correctional Health Care Services	5
County of Los Angeles	5
Citibank	5
Franciscan Health Indianapolis	5
Eastern Illinois University	5
Kaiser Foundation Health Plan, Inc.	5
Houston Methodist Hospital	5
Indiana bureauof motor vehicles	5
North Carolina Department of Health and Human S	5
Montana State University	5
Rite Aid Corporation	5
University of Iowa	5
University of South Carolina	5
University of Toledo	5
Wells Fargo	5
Union Security Insurance Company	5
Boeing	4
Apple	4
Bank of the West	4
Brigham and Women's Hospital	4
Chapman University	4
Citigroup	4
City University of New York	4
Cook County Health & Hospitals System	4
Children's Medical Center of Dallas	4
Community Health Network	4
Central City Concern	4
Harvard University	4
Discover Financial Services	4
Health Care Service Corporation	4
Fidelity Investments	4
H&R Block	4

Base VERIS	
Entreprise	Nb occurrence
United States Department of Veterans Affairs	873
NA	342
US National Security Agency (NSA)	14
Facebook	11
Department of Veterans Affairs	11
7-Eleven	11
Internal Revenue Service	10
Experian	10
Veterans Health Administration	9
Microsoft Corp	8
Walgreen Co	7
TD Bank	7
NYU Langone Medical Center	7
Mass Event	7
Medway Maritime Hospital	7
Google Inc	7
Circle K	7
Chase Bank	7
Alberta Health Services	7
Wells Fargo & Company	6
University of Pittsburgh Medical Center	6
Stoke City Council	6
Tampa General Hospital	6
Mount Sinai Medical Center	6
Microsoft	6
Jackson Health System	6
HSBC	6
Florida Department of Juvenile Justice	6
Florida Department of Health	6
Citigroup Inc.	6
Bank of America	6
Alabama Police Department	6
UnitedHealth Group	5
Wells Fargo	5
University of California, San Francisco	5
US Army	5
Shell	5
Sunoco	5
Pfizer, Inc.	5
Orlando Health	5
Massachusetts Mutual Life Insurance Company	5
Ministry of Health	5
Midlothian Council	5
Kaiser Permanente	5
Iowa Department of Human Services	5
CVS	5
Aetna	5
BOLTON NHS FOUNDATION TRUST	5
Valve Corporation	4
Washington State Department of Social and He:	4
Verizon	4
University of Texas MD Anderson Cancer Cente	4
United States Post Office	4
US Federal Bureau of Investigation (FBI)	4
Texas Department of Health and Human Servic	4
Torrance Memorial Medical Center	4
Tesco Corp	4
Sutter Health	4
State of Alabama	4

FIGURE B.1 – Entreprises apparaissant le plus souvent dans chacune des bases

Année	PRC - Publication	VERIS - Survenance	VERIS - Reporting
2004 -	0	48	0
2005	136	21	0
2006	482	22	0
2007	456	51	0
2008	355	82	0
2009	270	93	0
2010	801	587	0
2011	793	546	0
2012	886	1270	0
2013	890	1936	2563
2014	869	1004	2241
2015	547	910	588
2016	823	819	1640
2017	853	542	740
2018	699	266	387
2019	0	1	39
Total	8860	8198	8198

TABLE B.1 – Comparaison des fréquences de publication, de survenance et de reporting pour les deux bases

Base PRC

Entreprise	Nb occurrence
Walgreen Co.	12
Walgreens	6
Crescent Health Inc., Walgreens	1
Walgreens Co.	1
Crescent Health Inc. - a Walgreens Company	1
Walgreens.com	1
Walgreens Health Initiative	1
Total général	23

Base VERIS

Entreprise	Nb occurrence
Walgreen Co	7
Walgreen Co.	2
Crescent Health Inc. - a Walgreens Company	1
Total général	10

FIGURE B.2 – Fréquence d'occurence de Walgreens dans chacune des bases

toutes les manières de saisir le nom de l'entreprise possible.

B.2.1 Exemple 1 : Walgreens

Walgreens est une chaîne de pharmacie américaine présente sur l'ensemble du territoire américain et fait partie des deux leaders du marché (avec CVS/Pharmacy). L'enseigne possède plus de 9000 emplacements physiques sur le territoire.

On remarque que pour plus de la moitié des saisies, il y a une erreur dans le nom de l'entreprise et ce quel que soit la base : en effet le nom s'orthographe bien avec un s (« Walgreens ») et non pas (« Walgreen »).

Il convient d'interpréter avec la plus grande précaution les études de fréquence fondées sur ces bases et sur ces champs, dès lors que les champs de noms d'entreprises n'ont pas été retraités.

En troisième approche, un zoom sur une année précise va permettre d'estimer si certaines corres-

Base PRC											
Entreprise Walgreens	Année de publication										
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total général
Total général	1	1	2	2	6	4	2	2	1	2	23

TABLE B.2 – Fréquence d'occurence de Walgreens dans la base PRC, par année de publication

Base VERIS							
Entreprise Walgreens	Année de survenance						
Étiquettes de lignes	2011	2012	2013	2014	2015	2016	Total général
Total général	1	2	2	2	1	2	10

TABLE B.3 – Fréquence d'occurence de Walgreens dans la base VERIS, par année de survenance

Date.Ma	Company	City	State	Type	Typ	Total.Records	Description.of.incident	Source.Ur	Year.of.Breach	Latitude	Longitude	Gather.Information.S
04/03/2016	Walgreen Co.	NA	Illinois	PHYS	MED	880	Location of breached	https://ocrp	2016	40,633125	-89,398528	US GA: Federal - HIPAA
04/03/2016	Walgreen Co.	Deerfield	Illinois	PHYS	MED	880	As reported by Health and	Human Servi	2016	42,171137	-87,844512	US GA: State

TABLE B.4 – Zoom sur l'année 2016 de PRC

pondances sont envisageables au sein d'une même année. L'année 2016 est retenue car chaque base possède deux incidents au cours de cette année (figure B.4).

Le même incident (date de publication le 4 mars 2016) a été reporté deux fois dans la base PRC : au niveau fédéral et au niveau de l'État de l'Illinois.

C'est problématique pour toutes les études basées sur la base PRC : à la fois concernant les travaux sur la fréquence que pour l'étude de la sévérité (exprimée en nombre de données perdues).

Enfin, en suivant le lien lié à l'incident : `https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf`

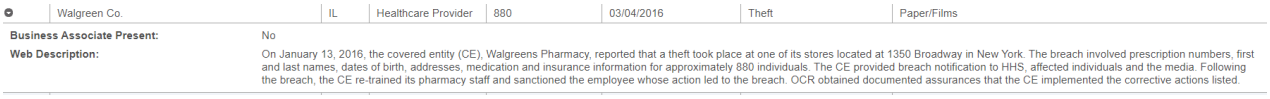
Il est possible de retrouver l'évènement en question, et seule une brève description est disponible (figure B.3).

Ainsi, un certain nombre d'informations reportées dans la base PRC est juste : le nombre de DCP, le type d'attaque (« PHYS », correspondant à un vol physique de documents), mais également une nouvelle erreur peut être remarquée : l'adresse précise de l'incident est 1350 Broadway, New York. Or, dans la base PRC l'adresse du siège de Walgreens est située à la distance de 1300 kilomètres de l'adresse new-yorkaise.

Dans la base VERIS, un même nom d'entreprise peut être associé à différentes tailles d'entreprises.

Entre les deux bases, l'incident présent en double dans la base PRC ne correspond à aucun des deux évènements de la base VERIS (il est possible de s'en rendre compte en utilisant les champs avec des liens vers des articles).

Ensuite, une même entreprise commercialisant des produits pharmaceutiques en retail, peut être classifiée en « MED » dans une des bases et en « RETAIL » dans l'autre base, ce point est problématique et reflète un problème de pertinence au sens de la qualité de données sous Solvabilité 2.



	Walgreen Co.	IL	Healthcare Provider	880	03/04/2016	Theft	Paper/Films

Business Associate Present: No

Web Description: On January 13, 2016, the covered entity (CE), Walgreens Pharmacy, reported that a theft took place at one of its stores located at 1350 Broadway in New York. The breach involved prescription numbers, first and last names, dates of birth, addresses, medication and insurance information for approximately 880 individuals. The CE provided breach notification to HHS, affected individuals and the media. Following the breach, the CE re-trained its pharmacy staff and sanctioned the employee whose action led to the breach. OCR obtained documented assurances that the CE implemented the corrective actions listed.

FIGURE B.3 – Evènement tel que visible sur ocrportal.hhs.gov

FIGURE B.4 – Le siège social de Walgreens est situé à Deerfield, près de Chicago, Illinois

incident_year	incident_month	incident_day	Size1	Size2	sector	number_records	revenue_USD	loss_USD	victim_id
2016	1	8	Large	H.100k+	Retail	NA	NA	NA	Walgreen Co
2016	7	1	Small	B.11.100	Retail	NA	NA	NA	Walgreen Co.

FIGURE B.5 – Zoom sur l'année 2016 de VERIS

B.2.2 Exemple 2 : Facebook

Facebook est l'une des entreprises les plus connues au monde, et les plus suivies par les médias notamment concernant le respect de la vie privée de ses utilisateurs.

On constate sur cet exemple que la base PRC contient moins d'incidents relatifs à Facebook que la base VERIS. De plus, la base PRC reporte 2 incidents incluant Facebook et d'autres entreprises telles que Twitter : après étude de chacun de ces incidents, il ressort qu'ils ne concernent pas Facebook directement. Ainsi, la base PRC reporte 5 incidents Facebook contre 13 pour la base VERIS (figure B.6).

En regardant les incidents de 2018 pour chacune des bases (figure B.7), la base VERIS contient des incidents correctement reportés, mis à part la donnée de la taille de l'entreprise, imprécise (Size2 est imprécise, Size1 « corrige » naturellement).

Dans la base PRC, des incohérences concernant le type d'organisation sont observables, la même entreprise étant reportée à la fois comme BSO (« Business – Other ») et BSR (« Business – Retail/Merchant – Including Online Retail »).

Entre les deux bases, il y a un incident en commun : celui concernant 50 millions d'utilisateurs. Tous les autres incidents reportés ne sont présents que dans une seule des deux bases.

C'est un point très important : les bases ont tendance à contenir les très gros incidents car ils sont plus médiatisés.

Base PRC

Entreprise Facebook	Année de publication				
	2008	2011	2013	2018	Total général
Facebook	1		2		3
Facebook, inc.				2	2
ADP, Facebook, Gmail, LinkedIn, Twitter, Yahoo, YouTube			1		1
Twitter, Facebook and PayPal		1			1
Total général	1	1	3	2	7

Base VERIS

Entreprise Facebook	Année de survenance				
Somme de count	Étiqu				
Étiquettes de lignes	2011	2012	2013	2018	Total général
Facebook	1	1	4	5	11
Facebook Inc			2		2
Total général	1	1	6	5	13

FIGURE B.6 – Fréquence d'occurence de Facebook, PRC vs VERIS

incident_year	incident_month	incident_day	reporting_year	Size1	Size2	sector	number_records	revenue_USD	loss_USD	victim_id
2018	5	18	2018	Large	F.25k.50k	Information	NA	NA	NA	Facebook
2018	**6**	**NA**	**2018**	**Large**	**F.25k.50k**	**Information**	**NA**	**NA**	**NA**	**Facebook**
2018	5	29	2018	Large	E.10k.25k	Information	NA	NA	NA	Facebook
2018	**5**	**29**	**2018**	**Large**	**F.25k.50k**	**Information**	**800 000**	**NA**	**NA**	**Facebook**
2018	NA	NA	2019	Large	D.1001.10k	Information	50 000 000	NA	NA	Facebook

FIGURE B.7 – Zoom sur l'année 2018 de VERIS

Date.Made	Company	City	State	Type.of.breach	Type.of.organization	Total.Records	Description	Information.Source	Source.URL	Year.of.Breach	Gather.Info
12/06/2018	Facebook, inc.	San Francisco	California	DISC	BSR	3 000 000	New Scientist r	Media	https://www	2018	Media
28/09/2018	**Facebook, inc.**	**NA**	**California**	**HACK**	**BSO**	**50 000 000**	**According to th**	**Media**	**https://www**	**2018**	**Media**

FIGURE B.8 – Zoom sur l'année 2018 de PRC