Etude Actuarielle du Cyber-Risque

Mémoire d'actuariat présenté pour l'obtention du

Master professionnel Sciences de gestion, mention finances de marché
Spécialité Actuariat du CNAM

Et l'admission à **l'Institut des Actuaires**

Mémoire soutenu le 26 novembre 2014

par Florian Pons (florian.pons.2009@asso-supelec.org)

Caractère confidentiel : non

Jury :

Président : Michel Fromenteau

Membres : Anne Serra

Florence Picard

Pierre Petauton

Vincent Ruol

François Weiss

Directeur de mémoire :

Benoit Huyot

Mots clés: Cyber-risque, Internet, Vol de données, Copule, Troncature aléatoire, R, Python, NLTK, Fouille de texte

Résumé

Les entreprises et les administrations sont de plus en plus dépendantes des systèmes d'informations. Cette dépendance existe à la fois dans les processus de production, de commercialisation, de communication et dans le stockage de données. Cela se traduit par l'existence de différents risques tels que l'arrêt du processus de production ou de commercialisation, la dégradation de l'image ou le vol de données. Ces nouveaux risques incitent à la création de nouveaux contrats d'assurance pour les couvrir.

Pendant de nombreuses années un mythe sur la possibilité de sécuriser complètement un système d'information était largement répandu. Ces dernières années une prise de conscience a amené les décisionnaires à penser qu'un risque résiduel existera toujours.

C'est un domaine technique, souvent méconnu du grand public, c'est pourquoi nous allons commencer par une présentation du sujet du point de vue de l'ingénieur en informatique. Nous allons faire quelques études de cas pour montrer au lecteur l'existence de ce risque et son coût. Puis nous présenterons le risque du point de vue de l'assureur ainsi que les difficultés rencontrées pour assurer ce risque. Ce qui permettra de comprendre alors pourquoi le marché a mis longtemps à exister. Comme il s'agit d'un marché émergeant il semble important de parler des acteurs actuels puis de son évolution future.

En deuxième partie nous allons présenter les indicateurs et protocoles qui ont été créés pour aider les entreprises et les administrations à identifier leurs risques informatiques afin de mieux les contrôler. Nous finirons cette partie sur la règlementation ainsi que les modèles techniques qui permettent d'évaluer le risque passé et présent. Ces outils peuvent aider à mieux quantifier les risques souscrits par les assureurs : c'est cela qui a motivé cette présentation.

Parmi tous les risques, celui de « vol de données personnelles » est un risque majeur. Il est soumis à des obligations de déclaration aux USA. Et les études de l'Institut Ponemon sur le coût de ces vols de données donnent des informations précises depuis plusieurs années. D'autre part le développement de l'obligation de déclaration pousse les entreprises à s'assurer de plus en plus en cas de vol de données, c'est donc le marché le plus porteur. La troisième partie, consacrée à l'étude actuarielle, sera donc restreinte aux problèmes des vols de données personnelles.

Il apparait qu'il existe une relation entre la quantité d'informations volées et le coût pour l'organisation. Cette relation **nous permet de construire une prime pure prudente** qui semble réaliste, en utilisant uniquement des données publiques.

Key words: Cyber-risk, Internet, Data breach, Copula, Random truncation, R, Python, NLTK, Text Mining

Abstract

Businesses and governments increasingly rely on their IT systems. This dependency is ubiquitous, from production through data storage and from communications to sales. This translates into different hazards such as stopping production or sales processes, image degradation or data theft. These new risks encourage the creation of new insurance policies to cover them.

For many years a myth about the possibility of completely securing information system was widespread. Awareness, which took place in recent years, has led policymakers to believe that a residual risk will always exist.

This is a technical field, often unknown to the general public, so I will start with a presentation of the subject from the point of view of the IT engineer. I will do some case studies to show the reader the existence of this risk and the associated cost. I then present the risk from the point of view of the insurer and the challenges of ensuring that risk. This will help to understand the late beginning of this market. As an emerging market it seems important to talk about current players and its future evolution.

In second part I will present the indicators and protocols that have been created to help businesses and governments to identify their IT risks in order to better control them. I will end this section on regulatory and technical models that can assess the past and present risks. These tools can help to better quantify the risks underwritten by insurers.

I cannot consider all cyber risks because of the expanse of this area and therefore in the third section of my actuarial study, I will consider the problem of theft of personal data. This will look at the reporting requirements in the USA and the Ponemon Institute studies of the cost of data theft. These reports cover developed and emerging countries from USA to India or Brazil from 2006 to 2014. The development of the reporting requirement makes companies take out more insurance for data theft, so it is a potential growth market.

It appears that a relationship exists between the amount of information stolen and the cost to the organization. This relationship **allows us to build a conservative pure premium** which seems realistic, using only public data.

Remerciements

Je remercie ma famille, et surtout mes parents, sans qui rien de tout cela n'aurait été possible.

Merci aux professeurs qui m'ont formé tout au long de ma vie d'étudiant, en « prépa » puis en école d'ingénieur. J'ai une reconnaissance particulière pour mes professeurs du CNAM (Conservatoire national des arts et métiers) et à leur faculté à partager leurs connaissances pointues, leur enthousiasme et leur goût pour l'actuariat. Je remercie aussi ceux qui ont pris le temps de me faire découvrir l'actuariat. En particulier Pierre de Villeneuve qui fut le premier à me présenter le métier et Alix Bakhos qui m'a accueilli dans ses équipes pour un stage en 2008.

Je remercie tous ceux qui m'ont aidé pour l'élaboration de ce mémoire. Je pense aux équipes de Thales dans lesquelles j'ai travaillé qui m'ont apporté leur expertise en statistique et en cyber sécurité. Je suis reconnaissant envers les équipes de SCOR qui m'ont apporté les connaissances métier et leur expertise en actuariat, en particulier envers Vincent Foucart et Fabien Gandrille pour avoir coordonné les travaux.

Enfin, merci à Florence Picard qui m'a suivie tout au long de mon stage, Michel Fromenteau qui m'a suivi tout au long de ma formation en actuariat, Olivier Lopez qui m'a apporté ses conseil en actuariat et en statistique, Catherine Gouttas qui m'a accueilli dans l'équipe du CENTAI (Centre de Traitement et d'Analyse de l'Information) et Benoît Huyot qui a fait le tutorat de mon mémoire.

Sommaire

I. Présentation du cyber-risque 12
A. Le cyber-risque 12
1. Fuite d'informations 13
2. Interruption/Dégradation d'un service 13
B. Les protections techniques 13
1. Les systèmes d'informations modernes 13
2. Les risques techniques principaux 15
3. Outils de protection principaux 16
C. Etudes de cas 18
1. TJX 18
2. Système de paiement Heartland 18
3. Sony 19
4. Target 19
5. Stuxnet 19
6. Smartphone 19
D. Les risques étudiés 20
1. Dommages aux biens 20
2. Dommages aux tiers 21
3. Dommage d'image 21
4. Autres dommages 21
E. Assurabilité 22
1. Historique 22
2. Aléa moral 22
3. Asymétrie d'information 22
4. Inter corrélation 22
F. Historique de la cyber-assurance 23
G. Les acteurs du marché 24
1. L'assurance 24
2. La réassurance 27

3. Les courtiers 27
4. Les acteurs spécialisés 28
5. Les services disponibles chez Thales 28
H. Le marché potentiel 31
II. Normes, méthodes et modèles techniques 31
A. Normes et méthodes 31
1. Les Normes ISO 27000 32
2. PCI DSS 34
3. Référentiel SP800-30 35
4. Méthode MEHARI 35
5. Méthode EBIOS 36
6. Méthodes OCTAVE 36
7. COBIT 5 / Risk IT Framework 37
8. SOC 2 37
9. Cyber essentials scheme 38
B. Réglementation 38
1. Obligations et responsabilités 38
2. Obligation de notification 39
C. Modèles basés sur la topologie du réseau 40
1. Graphe d'attaque 40
2. CHASSIS 41
3. Arbre de défaillance / Arbre d'attaque 42
4. BDMP (Boolean logic Driven Markov Processes) 42
5. Comment avoir la topologie du réseau 43
D. Modèles sans topologie : Approche Multistate 44
E. Conclusion 44
III. Modélisation actuarielle 45
A. Modèles existants 45
1. Propagation Virale 45
2. Modèle économique 45
3. Queue épaisse 46
4. Indépendance entre Fréquence et Sévérité 46

5. Modèles de copule pour l'évaluation tarifaire 47
6. Choix des indicateurs 47
7. Conclusion 47
B. Les données Open Data 48
1. Base de vulnérabilités 48
2. Base de violation de SI 49
3. Nombres d'entreprises aux USA 51
4. Institut Ponemon 51
5. Tarifs 54
C. Analyse du commentaire des violations USA 55
1. Les dates 55
2. Les quantités 56
3. Les montants 56
D. Fréquence 57
1. Données USA 57
2. Reste du monde 57
E. Sévérité (en volume de données) 58
1. Institut Ponemon 58
2. Base de données des violations aux USA 60
3. Comparaison 64
F. Sévérité (en coût) 68
1. Institut Ponemon 68
G. Relation entre sévérités 71
1. Etude préliminaire 71
2. Modèle linéaire 72
3. Copule 75
H. Tarification : Application sur le calcul de la prime pure 80
I. Evolution du risque 83
J. Limites des modèles 84
IV. Conclusion 85
V. Bibliographie 86

Introduction

Nous appellerons cyber-risque l'ensemble des risques susceptibles d'apparaitre suite à l'usage d'un ou plusieurs systèmes informatiques éventuellement reliés en réseau. Ce risque émergeant amène la création de nouveaux produits d'assurance que nous allons donc étudier.

Les entreprises sont de plus en plus exposées aux cyber-risques, quelle que soit leur taille. Cela fait suite, d'une part, au développement des technologies de l'information dans tous les processus des entreprises et d'autre part, au développement d'Internet qui ouvre les systèmes d'informations des dites technologies vers l'extérieur et crée de nouvelles opportunités d'attaques.

Nous sommes de plus en plus dépendants des ordinateurs lorsque nous travaillons et de plus en plus de documents sont dématérialisés. Ainsi la protection des systèmes d'informations et de données numériques devient un enjeu majeur pour les entreprises.

Lorsqu'une entreprise choisit de sécuriser son système d'information, certaines contraintes opérationnelles ou des coûts excessifs la poussent à conserver un risque résiduel. Ce dernier nécessite alors la présence d'une offre d'assurance sur le marché.

L'objectif de ce mémoire est donc de faire une synthèse des connaissances actuelles sur la modélisation des risques résiduels, c'est-à-dire, étudier avec une vision assurancielle les modélisations possibles de l'évaluation financière du risque d'attaque, qu'il s'agisse d'attaques externes ou internes, visant à une récupération d'informations.

La première partie permet au lecteur de se familiariser avec le sujet. Nous abordons donc le vocabulaire technique utile pour une bonne compréhension des problématiques. Puis nous faisons quelques études de cas et une présentation du risque étudié pour montrer au lecteur que ce risque a de réels impacts financiers. Enfin nous présentons le marché actuel, qui est un marché émergeant donc peu connu mais sur lequel plusieurs acteurs de l'assurance interviennent déjà.

Dans une deuxième partie nous étudions les normes existantes en cyber sécurité. Ces normes peuvent devenir des indicateurs sur le risque porté par une entreprise et leur bonne application est un levier pour inciter l'assuré à limiter son risque. Il semble donc nécessaire pour un souscripteur de connaître ces normes. Par la suite nous présentons aussi les grandes lignes de la réglementation actuelle ou à venir.

Enfin, les modèles d'évaluation développés en cyber-sécurité sont présentés. Ils servent à identifier les risques. Les résultats de ces modèles peuvent servir d'indicateur pour les souscripteurs et de paramètres discriminants pour les actuaires qui pourraient travailler sur le sujet, c'est pourquoi nous les présentons ici.

Dans une dernière partie, nous abordons le sujet d'un point de vue quantitatif. Du fait de la réglementation le sujet du vol d'informations personnelles constitue le premier besoin en assurance exprimé par les entreprises et le domaine où les données sont le plus facilement disponibles. Nous allons donc étudier cette problématique-là. Dans un premier temps nous présentons les modèles existants. Malgré un sujet encore peu exploré, il apparait quelques informations intéressantes : en particulier l'indépendance entre la fréquence des sinistres et leur sévérité qui sera une de mes hypothèses structurantes. Puis je présente les données que j'ai à ma disposition : en particulier la base de données des incidents déclarés à l'administration des USA et les résultats des enquêtes menées depuis plusieurs années dans de nombreux pays par l'Institut Ponemon.

En considérant que les systèmes informatiques sont très semblables entre tous les pays, on fera l'hypothèse qu'en termes de « risque de violation » les USA sont représentatifs du monde entier. A partir de la base des violations aux USA, qui nous renseigne sur le volume de données volées, on peut en déduire le risque quantifié en volume de données.

A l'aide des données de l'Institut Ponemon on peut en déduire une relation entre le nombre de données volées et le coût total pour l'organisation. Cette relation dépend du pays. Cependant on remarque que l'Institut Ponemon a eu tendance à exclure les plus grands et les plus petits sinistres de ses études, sans pour autant avoir un seuil d'exclusion net. C'est pourquoi on va considérer que les données de l'Institut Ponemon sont issues d'un sondage avec une troncature aléatoire sur l'univers complet (représentée par la base de violation aux USA). Cela permet d'en déduire une relation entre le volume de données volées (aussi appelé taille de la violation) et le coût dans l'univers complet. On aura ainsi trouvé la loi du coût et on en déduira une prime pure.

Présentation de l'environnement de travail

Le stage s'est déroulé dans le cadre d'un congé individuel de formation pris auprès de l'entreprise Alten SIR. Je travaillais dans les locaux de Thales avec l'équipe CENTAI et le projet a été organisé avec la participation de SCOR.

Au sein de Thales Communications & Security, l'équipe du CENTAI est un centre d'expertise d'une douzaine de personnes spécialisées dans l'analyse de l'information. Cette équipe propose des solutions innovantes et a un aspect fort de Recherche et Développement (R&D). L'équipe travaille sur plusieurs projets en parallèle allant de la cyber-sécurité à la billettique en passant par l'analyse des réseaux sociaux. Ces projets sont pour la plupart développés sur des plateformes Big Data.

L'équipe de Thales avait pour rôle d'apporter une expertise dans le traitement des données, l'analyse statistique ainsi que des connaissances en sécurité des systèmes d'informations. SCOR avait pour rôle d'apporter une expertise métier en assurance et réassurance, des conseils en actuariat ainsi que des données relatives au sujet.

La jeunesse et la spécialisation du sujet n'ont pas permis à SCOR de sélectionner des données pertinentes dans ses bases de sinistres. Ainsi, le mémoire est concentré sur les données librement disponibles. De par ma spécialisation en sécurité des systèmes d'informations acquise durant ma formation d'ingénieur, j'ai pu facilement communiquer avec les spécialistes du domaine chez Thales. J'ai donc concentré la première moitié de mon stage sur le recueil d'informations, ce qui constitue les deux premières parties du mémoire. La seconde moitié du stage était ciblée sur l'analyse actuarielle et statistique qui constitue la troisième partie du mémoire.

I. Présentation du cyber-risque

L'objectif de cette partie est de présenter le risque étudié.

On cherche à donner au lecteur à la fois le vocabulaire mais aussi l'environnement dans lequel les contrats d'assurance peuvent être créés.

A. Le cyber-risque

Quels sont les risques auxquels sont exposées les entreprises qui nous intéressent ici ?

Dans la majorité des entreprises, les incidents survenant sur un ordinateur sont majoritairement gérés par les administrateurs réseaux. On va donc présenter le sujet en se familiarisant avec l'approche de l'ingénieur en informatique. Cela permettra aussi au lecteur de se familiariser avec le vocabulaire technique.

Dans un premiers temps, voici un tableau reliant les risques techniques avec les protections possibles. Nous détaillerons par la suite le vocabulaire utilisé.

Risque technique	**Protection**
Défaillance matérielle	• Réplication • Duplication
Prise de contrôle à distance	• Identification / Gestion de droits • Chiffrement et Signature • HoneyPot
Déni de service (DoS)	• Duplication • Load Balancing • Firewall
Ecoute (Sniffing)	• Chiffrement • Identification / Gestion de droits • Firewall
Usurpation d'identité	• Signature • Identification / Gestion de droits
Intrusion	• Identification / Gestion de droits • Signature • Firewall • HoneyPot

Les deux principaux problèmes rencontrés sont soit une fuite d'informations confidentielles soit un service fourni non conforme aux attentes par le système informatique.

1. Fuite d'informations

L'information volée ou perdue peut être de deux natures :

- Information stratégique qui constitue un secret technique ou marketing, un savoir-faire, etc. Cette perte constitue la diminution d'un avantage concurrentiel. Dans le cas de secrets techniques, il peut s'agir d'une découverte pouvant être connue par un concurrent avant que le brevet la protégeant ne soit déposé. On peut citer un évènement de grande ampleur dont certaines fuites peuvent en diminuer l'impact médiatique, dans le cas d'un secret marketing.
- Information sur des clients. Il peut y avoir un risque d'image pour l'entreprise car le client ne désire pas la divulgation de certaines données. Il peut y avoir également préjudice financier, par exemple lorsque des numéros de cartes de paiement sont volés.

2. Interruption/Dégradation d'un service

On peut citer les exemples suivants :

- Interruption d'un service commercial. Par exemple, le site de vente en ligne est inaccessible ou les terminaux des magasins ne fonctionnent plus ce qui empêche la vente des produits.
- Interruption ou dégradation d'un service marketing, ce qui peut engendrer une dégradation de l'image de marque. Par exemple, un site web qui est vandalisé.
- Interruption ou dégradation d'un service interne, ce qui cause une perte de productivité des salariés. Par exemple, la messagerie interne de l'entreprise ne fonctionne plus, donc les salariés doivent se déplacer pour informer leurs collègues.
- Dégradation d'un support de données engendrant une perte d'information. Par exemple, un serveur de disque qui brûle.

Nous avons listé le vocabulaire qui sera présenté ainsi que les risques étudiés du point de vue de l'assuré.

B. Les protections techniques

Les administrateurs réseau ont développé au fil des années des outils techniques pour se protéger contre les cyber-risques.

Nous allons présenter la structure des systèmes d'informations modernes afin de donner au lecteur une meilleure compréhension des termes utilisés. Puis nous nous intéresserons aux outils de protections les plus courants.

1. Les systèmes d'informations modernes

L'IP (Internet Protocol) représente la base des systèmes d'informations modernes. Ce protocole permet à tous les ordinateurs présents sur un même réseau de communiquer entre eux et ce même s'il n'y a pas de liaison directe. Du fait de la pénurie croissante d'adresses IP, il a été nécessaire de recourir aux routeurs.

Le routeur fait office de porte-parole d'un groupe d'ordinateurs (en tant qu'IP publique du groupe) situé au sein d'un réseau bien plus grand.

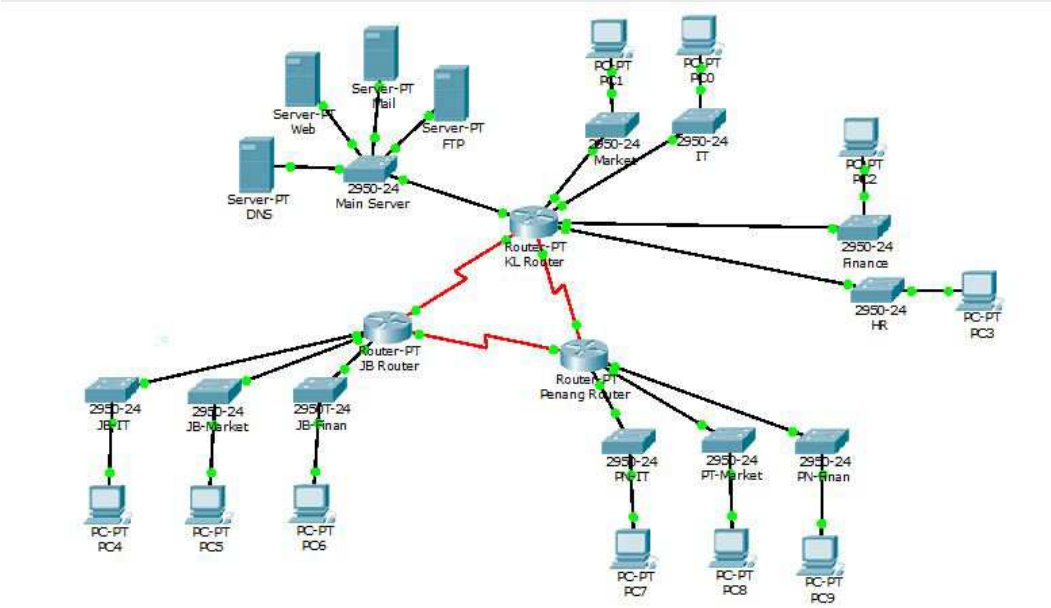


Figure 1 Rôle d'un routeur

Les ordinateurs peuvent communiquer entre eux soit par des ondes passant par des câbles soit par des ondes passant dans l'air.

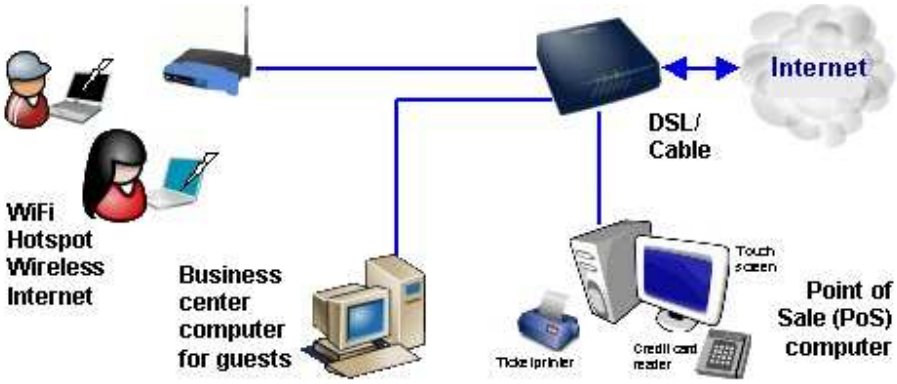


Figure 2 Réseau à topologie complexe

Les réseaux modernes évoluent selon que l'on branche ou débranche des équipements. Ces branchements peuvent être faits soit par câble, soit par Wifi. Avec l'utilisation d'objets nomades, connectés sans fil, **il est de plus en plus difficile pour un administrateur de connaitre le système dont il a la charge**.

2. Les risques techniques principaux

a) Défaillance matérielle

Ce cas recoupe tous les incidents qui font suite à une panne sur un matériel physique. Cela peut faire suite à un accident tel qu'un incendie, mais aussi suite à une usure telle que l'arrêt d'un disque dur ou un acte malveillant tel que la rupture volontaire d'un câble.

b) Prise de contrôle à distance

Une personne prend le contrôle d'un ordinateur depuis un appareil distant. Il peut alors utiliser la ressource en question comme il le désire. Cela se fait généralement à l'aide de virus informatiques. L'ordinateur infecté est alors appelé **PC zombie**.

c) Déni de service (DoS)

Cela peut être causé par une affluence record. Par exemple, suite à une publicité sur un grand média, un site internet peut se retrouver saturé. Les serveurs recevant plus de requêtes qu'ils ne peuvent en traiter, ils doivent les stocker et finissent par remplir l'espace de stockage intermédiaire. N'aillant plus de ressources disponibles, le serveur ne peut plus fonctionner et doit être redémarré.

Cela peut aussi faire suite à un acte malveillant, il existe deux cas qui peuvent être combinés.

Dans le premier cas, le serveur présente un bug et certaines requêtes vont stopper son fonctionnement normal ou alors engendrer une consommation de ressources excessives ce qui empêche le serveur de fonctionner.

Dans le second cas, l'attaquant a pris possession d'un grand nombre de PC et utilise toutes ses ressources pour saturer le serveur sur une période donnée. On parle alors d'attaque par **« déni de service distribué » ou DDOS** (Distributed denial of service).

d) Ecoute (Sniffing)

Une personne fait en sorte de récupérer les échanges faits entre deux ordinateurs pour récupérer l'information échangée. Le premier objectif est donc d'obtenir une fuite d'informations.

e) Usurpation d'identité

Le principe est de se faire passer pour une autre personne sur le réseau. Par exemple, un concurrent peut se faire passer pour un employé de l'entreprise auprès des autres employés afin de récupérer des informations stratégiques.

f) Intrusion

L'intrusion est un terme générique pour dire qu'une personne a pu mener à bien la récupération d'informations sensibles ou faire une dégradation en combinant parfois plusieurs des techniques précédentes au sein d'un réseau d'ordinateurs.

3. Outils de protection principaux

a) Chiffrement et signature

Le principe est d'encapsuler un ensemble d'informations avec une clé, ce qu'on appelle **encoder**, et cet ensemble d'informations ne peut sortir de la capsule qu'à l'aide d'une clé (qui peut être différente de la clé d'encodage). C'est ce qu'on appelle **décoder**.

Dans le cas d'un **chiffrement**, le but est de rendre l'information accessible uniquement au détenteur de la clé. On cherche ainsi à **garantir une confidentialité** des informations échangées.

Dans le cas d'une **signature**, celui qui décode le message sait que seul un autre détenteur de la clé peut avoir encodé, donc il sait que **le message vient de la bonne personne**.

Il existe deux méthodes principales, soit il y a une seule clé pour coder et décoder, soit il y a deux clés.

Dans le premier cas, il est clair qu'il faut échanger la clé de façon confidentielle avant.

Dans le second cas, on parle de « clé publique » et « clé privée ». La clé publique est connue des deux partis et la clé privée d'un seul des deux. Pour signer un message, le détenteur des deux clés signe avec la clé privée et le message peut être décodé avec la clé publique. Pour chiffrer un message, on l'encode avec la clé publique et seul le détenteur des deux clés peut alors le déchiffrer.

b) Firewall

Il se place généralement au même endroit que le routeur et sert à **protéger un sous-réseau sûr au sein d'un plus grand réseau**. Le principe du firewall est de filtrer les flux qui le traversent et de bloquer ceux qui sont considérés comme malveillants.

Le premier rôle du firewall est d'empêcher un ordinateur extérieur au réseau de déclencher une communication avec un ordinateur interne. Comme seuls les ordinateurs internes peuvent commencer les communications, ils sont protégés des attaques aléatoires qui pourraient lui être envoyées.

Il est aussi courant que le firewall bloque certaines communications initialisées depuis l'intérieur. Si un ordinateur est détourné de son usage normal et tente soit d'envoyer des informations confidentielles à l'extérieur soit de faire une action qui pourrait l'endommager, il est utile de bloquer le flux. Par exemple, un utilisateur mal informé qui va sur un site dangereux pourrait compromettre son poste.

c) Identification / Gestion de droits

On parle ici de méthodes pour **identifier une personne physique**. Il existe 2 méthodes principales qui sont l'utilisation d'un mot de passe que doit retenir la personne et qu'il est le seul à connaitre; ou l'utilisation de données biométriques (par exemple, des empreintes digitales).

Cette identification permet de décider si une personne physique a accès au système informatique ou non. Mais, pour sécuriser au mieux les données, il est important de faire une **gestion des droits par utilisateur**.

Lorsque l'identité d'une personne est usurpée sur le système d'information, si cette personne a des droits illimités, les conséquences peuvent être maximales. Pour limiter les conséquences, il est important d'encadrer les droits et donc de faire une gestion des droits. Cela devient primordial du fait du développement de la mobilité. En effet, un ordinateur portable peut constituer une porte d'entrée sur le système d'information à l'aide de technologies d'accès à distance aux ressources.

Il faut aussi prendre en compte le cas où aucun acte malveillant n'intervient, mais qu'une erreur non intentionnelle est commise. La limitation des droits permet aussi de réduire l'impact de l'erreur.

Il faut également intégrer la **gestion des droits des services ou logiciels**. En effet, lorsqu'un attaquant veut pénétrer un système d'information, il peut prendre le contrôle d'un logiciel ou d'un service à l'aide d'un code malveillant (virus par exemple). Comme pour une personne physique, il faut réduire les droits pour limiter l'impact de l'attaque.

d) Réplication

Le principe est de **recopier la donnée dans différents endroits**. La réplication permet à la fois de protéger contre des dommages matériels (incendies et inondations), les pannes et les attaques visant à détruire.

Il faut noter qu'une réplication n'est jamais instantanée, le risque de perte de donnée subsiste donc, mais la perte reste généralement mineure.

e) Duplication

Le principe est de faire **tout ou partie d'un système en plusieurs exemplaires** (identiques ou différents). Le premier but est proche de la réplication et permet de se protéger contre des dommages matériels, des pannes ou des attaques à but destructeur.

Il est plus simple et moins coûteux de dupliquer en utilisant exactement des technologies et une architecture communes à l'ensemble des duplications. Cependant, utiliser des duplications différentes les unes des autres améliore la protection. En effet, si l'une d'elles possède une erreur de conception ou une faille alors les autres ont de bonnes chances de ne pas comporter le même problème.

f) HoneyPot

Le principe ici est de faire croire à un attaquant qu'il a pénétré le système alors qu'il est dans un environnement fait pour être pénétré. Cela permet à la fois de détecter l'attaque et de ralentir l'attaquant le temps qu'il réalise qu'il a été piégé.

g) Load Balancing

C'est une technique qui permet de mieux exploiter les ressources lorsqu'on fait des duplications. Ici un serveur d'entrée répartie les requêtes sur un ensemble de serveurs en fonction des volumes de demandes. Il peut même envoyer une réponse automatique en cas de surcharge.

C'est à la fois une façon de garantir un meilleur confort d'utilisation à l'utilisateur, mais également une méthode pour se protéger contre les attaques de type DDOS.

Nous venons de voir le cadre technique dont un contrat d'assurance cyber-risque dépend. Cela pourra aider un éventuel souscripteur ou un régleur à comprendre ses interlocuteurs techniques.

C. Etudes de cas

Année	Nom	Zone géographique	Coût total	Montant assuré
2005 à 2007	TJX	USA	250 m$	
2008	Heartland	USA	140 m$	30 m$
2011	Sony	Monde	171 m$	
2013	Target	USA	61 m$	44 m$
2010	Stuxnet	Monde	Contrôle d'installations sensibles	
2009	Smartphone	Démonstration de prise de contrôle à distance		

m = million

1. TJX

TJX est une entreprise américaine fondée en 1956 spécialisée dans la vente de produits de décoration et de mode à prix réduit. Cette entreprise a été victime d'un vol de données de 2005 à 2007.

La violation a été rendue publique le 17 janvier 2007 et avait commencée mi-2005. En 18 mois, ce sont près de 46 millions de numéros de cartes bancaires qui ont été dérobés et 95 millions de numéros qui ont été exposés. Le coût total pour l'entreprise a été estimé à 250 millions de dollars (VIJAYAN, 2008).

2. Système de paiement Heartland

Heartland est une entreprise américaine fondée en 1997 de plus de 3000 employés qui fournit des services de gestion de payement électroniques. Cette entreprise a été victime d'un vol de données en 2008.

La violation a été rendue publique le 20 janvier 2009, il s'agit de la plus grosse intrusion informatique de l'histoire de par le nombre de numéros de cartes de crédit volés. Les

données volées permettaient de déclencher un paiement sans l'accord du possesseur de la carte (Wiki3). C'est près de 130 millions de numéros de cartes qui ont été impactés.

Le coût pour l'entreprise a été estimé à près de 140 millions de dollars, avec en particulier 60 millions de compensation pour Visa et 26 millions de frais légaux. L'entreprise a été indemnisée de 30 millions par les assurances (VIJAYAN, 2010).

3. Sony

Une attaque qui s'est déroulée entre le 17 Avril et le 19 Avril 2011 a permis le vol de 77 millions de comptes PlayStation Network (Wiki5). Ces informations contenaient 12 millions de numéros de cartes de crédit. La perte estimée par Sony s'élève à 171 millions de dollars.

4. Target

Rendue publique par communiqué de presse le 19 décembre 2013, c'est la seconde plus importante intrusion informatique compte tenu du nombre de codes de cartes de crédit volés (Wiki4). 40 millions de numéros de cartes de crédit ont été dérobés. Mais également 70 millions de données contenant les coordonnées des clients.

Il faut noter que la compromission provenait d'une négligence sur certains terminaux des points de vente de la marque. L'attaque est passée par l'un des prestataires en matériel de réfrigération, dont le réseau était compromis, et qui avait un accès au réseau de Target dans le cadre d'échanges commerciaux. L'accès était fourni par un serveur présentant une vulnérabilité exploitée par l'attaquant pour le transformer en point d'entrée permanent. Par la suite, la séparation entre le réseau interne des systèmes de payement et le reste du réseau de Target n'était pas suffisamment étanche. Cela a permis à l'attaquant de traverser. Enfin, l'attaquant a utilisé des postes connectés à Internet et insuffisamment protégés pour exfiltrer les informations dérobées. Le site web d'achats en ligne n'a pas été impacté.

Le coût pour l'entreprise a été estimé à 61 millions de dollars, dont 44 millions couverts par les assurances (ZDnet, 2014). Cette affaire a contribué au départ du PDG de l'entreprise le 5 mai 2014, ce dernier étant considéré comme responsable de négligence.

5. Stuxnet

C'est un ver informatique qui a ciblé en 2010 les systèmes SCADA utilisés pour le contrôle commande de procédés industriels (Wiki6).

Il faut noter que certaines centrales nucléaires ont été compromises avec ce ver.

6. Smartphone

Plusieurs failles ont été décelées sur les smartphones, sans pour autant avoir été officiellement exploitées dans l'intention de nuire.

Néanmoins, en 2009, une démonstration de prise de contrôle à distance a été faite au BlackHat (MULLINER, 2009). Depuis de nombreuses failles ont été révélées.

D. Les risques étudiés

Nous allons réaliser une classification des différents risques étudiés.

	Pertes directes	**Pertes indirectes**
Dommages aux biens	• Dégradation matérielle • Perte d'un capital immatériel • Appel à des experts externes	
Dommages aux tiers	• Dégradation matérielle • Perte d'un capital immatériel • Rupture d'activité ou dégradation du service	
Dommage d'image	• Offre pécuniaire faite aux clients • Mise en place d'un centre d'appels • Perte d'agréments	• Communications internes • Perte d'agréments • Perte de réputation
Autres dommages	• Rupture d'activité ou dégradation du service	• Enquêtes internes

1. Dommages aux biens

Les biens immatériels sont les plus exposés au cyber-risque. Néanmoins, il est possible que des biens matériels soient exposés au cyber-risque.

1) **Pertes directes**
   a) Dégradation matérielle

      Par exemple, le virus stuxnet qui avait pour but de déclencher une dégradation des installations industrielles.
   b) Perte d'un capital immatériel

      Il peut s'agir d'un secret industriel, d'une stratégie marketing ou du contenu d'un brevet avant qu'il ne soit déposé.
   c) Appel à des experts externes

      Pour réparer le problème au plus vite, par exemple arrêter la fuite de données et combler la brèche.

Il faut noter que la reconstruction d'un système informatique après une intrusion de grande envergure fait partie des coûts majeurs. En effet, lorsqu'un système a été pénétré trop longtemps il n'est plus possible d'en assurer la sécurité, il faut donc le refaire entièrement.

2. Dommages aux tiers

1) **Pertes directes**
   a) Dégradation matérielle
      Dégradation d'un bien matériel confié par un tiers.
   b) Perte d'un capital immatériel
      Perte sur un bien immatériel confié par un tiers.
   c) Rupture d'activité ou dégradation du service
      Rupture d'un service utile à un tiers qui peut demander des indemnités.

3. Dommage d'image

Une attaque peut dégrader l'image d'une organisation et donc entraver sa capacité à commercialiser ses produits.

1) **Pertes directes**
   a) Offre pécuniaire faites aux clients
      Pour compenser le préjudice subi.
   b) Mise en place d'un centre d'appels
      Pour répondre aux inquiétudes des clients et les informer.
   c) Perte d'agréments
      En particulier dans le secteur de l'armement, un vol de données peut engendrer une perte d'agréments sur les sujets classés et donc rompre de nombreux contrats en cours.
2) **Pertes indirectes**
   a) Communications internes
   b) Perte d'agréments
      Ici on peut envisager les diminutions de contrats futurs dû à une perte partielle ou totale d'agréments, même si la perte d'agréments n'est que temporaire.
   c) Perte de réputation
      En particulier, on peut citer la diminution du volume de clients ou une chute dans la croissance du nombre de clients

4. Autres dommages

1) **Pertes directes**
   a) Rupture d'activité ou dégradation du service
      Ce qui engendre une perte de productivité ou une perte commerciale dans le cas d'un site de vente en ligne.
2) **Pertes indirectes**
   a) Enquêtes internes

E. Assurabilité

Le cyber risque rencontre actuellement plusieurs problèmes d'assurabilité.

1. Historique

C'est un risque récent avec peu d'historique, qui a un peu plus de 20 ans d'âge. De plus, c'est un risque qui a changé rapidement avec le développement de nouvelles utilisations des outils informatiques. Le développement d'internet a fondamentalement changé la dimension et le type de risques en permettant une action mal intentionnée depuis n'importe quel endroit dans le monde. Internet a aussi augmenté le nombre d'interconnexions entre les réseaux d'entreprises et ainsi augmenté le risque.

D'autre part, le risque continue de changer avec l'arrivée des smartphones. Le risque est de moins en moins localisé d'un point de vue géographique. De plus, la sécurisation du matériel est particulièrement compliquée pour un appareil qui peut être volé très facilement.

Pour préserver leur image, les entreprises cachent les incidents informatiques qu'elles ont subis. Il est donc difficile d'avoir une information pour quantifier le risque. L'obligation de déclaration de vol d'information aux USA peut constituer une première base de travail, même s'il doit exister des différences de risques avec le reste du monde. Certaines entreprises commencent à avoir des obligations d'information en Europe. On peut citer le règlement européen du 24 juin 2013 dit « data breach ». Cette évolution réglementaire peut représenter une formidable évolution dans la capacité qu'auront les assureurs à évaluer précisément ce risque, mais ces données ne sont pas forcément publiques.

2. Aléa moral

Se sachant couvert, l'assuré peut réduire sa vigilance. Pour contrer ce problème, il existe deux méthodes complémentaires, l'application d'une franchise comme dans l'assurance automobile et imposer des audits du système d'information.

3. Asymétrie d'information

Ce risque comporte un biais important d'anti-sélection. En effet, il est difficile pour un assureur de bien connaitre la structure du système d'information ainsi que la formation que les équipes ont reçue sur la protection des données. L'audit du système d'information est un bon moyen pour corriger ce problème, ainsi que l'application des normes.

4. Inter corrélation

Un même incident de sécurité peut impacter un grand nombre d'entreprises. En effet, de nos jours, nombreuses sont les entreprises qui utilisent les mêmes logiciels qui sont en quelque sorte standards. On peut citer Windows avec MS Office. Lorsqu'une vulnérabilité existe sur un logiciel très répandu, le risque qu'une action malveillante soit menée à bien augmente sur toutes les entreprises utilisant ce logiciel. Les cyber-risques de toutes ces entreprises sont alors corrélés.

De plus, une compromission peut avoir des conséquences sur des entreprises indépendantes. Par exemple, de nombreux utilisateurs utilisent très peu de combinaisons *login/mot de passe* sur internet. On retrouve donc la même combinaison sur de nombreux services pour un seul utilisateur. Le hacker qui a compromis un site peut alors usurper l'identité de nombreux utilisateurs sur d'autres sites. Il peut aussi utiliser ces informations pour faire du « fishing » ciblé, qui aura alors plus de crédibilité auprès de l'utilisateur qui se fera alors hameçonner et donnera des informations sensibles au hacker.

De nombreuses entreprises mutualisent les coûts en utilisant les mêmes plateformes. Il y a donc une forte corrélation entre les risques. Par conséquent les réassureurs ne veulent pas assurer ce type de risque (BÖHME, et al., 2010).

F. Historique de la cyber-assurance

La cyber-assurance trouve ses **origines dans les années 1980** (BÖHME, et al., 2010). L'offre de cyber-assurance a **commencé dans les années 1990** dans le secteur bancaire (ANDERSON, 1994). Les premières offres sont arrivées sur le marché dans les pays anglo-saxons, mais se sont propagées depuis dans tous les pays développés et ont atteint de nombreux secteurs d'activité.

Il existe depuis longtemps des contrats d'assurance contre le vol, la détérioration des biens par un tiers et l'entrave au bon fonctionnement des services de l'entreprise. Cependant avec l'informatisation, ces risques se sont transformés. La connaissance est devenue un atout stratégique. De plus, le bon fonctionnement du système d'information est indispensable à l'entreprise. Du fait de la complexité d'un système d'information, et de la vitesse avec laquelle une attaque peut être menée en toute discrétion, le cyber-risque doit être pris en charge via des contrats spécifiques.

Malgré des débuts prometteurs, l'offre de cyber-assurance peine à trouver son public et reste aujourd'hui un marché limité. Malgré une explosion de l'industrie Internet pendant les années 2000, le marché de la cyber-assurance ne s'est pas développé dans les mêmes proportions (BÖHME, et al., 2010). De nombreuses entreprises préfèrent s'auto-assurer et les polices disponibles sur le marché ne couvrent pas l'ensemble des risques.

Une des difficultés rencontrées lors du développement du marché de la cyber-assurance est l'incapacité pour l'assuré de comparer les contrats (TOREGAS, et al., 2014). Il y a donc un besoin de standardisation pour pouvoir comparer les différentes offres disponibles. L'inter-corrélation des risques des entreprises conjuguée au manque de données lié à la jeunesse de ce secteur incite les assureurs à la plus grande prudence en termes d'offre ce qui limite l'émergence du marché.

Cependant, on observe une accélération de l'offre ces dernières années. C'est en particulier lié au fait que des acteurs majeurs tels qu'Allianz s'intéressent désormais à ce marché. De plus, une offre de réassurance se met progressivement en place, grâce à des industriels du secteur tels que la SCOR. D'autre part, de plus en plus d'entreprises envisagent de s'assurer contre le cyber-risque (TOREGAS, et al., 2014). On peut donc s'attendre à une croissance du marché dans les années à venir.

G. Les acteurs du marché

1. L'assurance

Assureur	**Capacité (montant max. pour l'assureur)**
Zurich France	25 millions d'euros
XL France	15 millions d'euros
Hiscox	
CNA	
Beazley	15 millions d'euros
ACE	
Axa	
Allianz	50 millions d'euros
AIG	

a) Zurich Assurance

Zurich France, avec une capacité de 25 millions d'euros, propose divers contrats en France.

En particulier on peut citer une solution pour les petites et moyennes entreprises. Cette formule propose une couverture mondiale des risques. De plus, elle fournit des services de spécialistes pour auditer avant intrusion et pour aider à la gestion de crise.

https://www.zurich.ch/fr/clientele-entreprises/dommage-au-patrimoine/cyber-security-and-privacy#en-detail

b) XL Assurance

XL France a une capacité de 15 millions d'euros.

L'assureur propose un panel de produits permettant d'avoir une couverture sur une grande partie de la planète.

http://xlgroup.com/insurance/insurance-coverage/professional-insurance/eo--information-technology-liability

c) Hiscox

Hiscox est un assureur anglais spécialisé dans les risques spéciaux tels que ceux concernant les objets d'art, les lieux d'exception et les assurances professionnelles. Il fait partie des assureurs bénéficiant d'une forte expérience dans le domaine du cyber-risque du fait de sa présence sur ce marché depuis 25 ans.

Le contrat « Data Risks by Hiscox » est une offre de services de prévention des risques et d'assistance techniques par des experts en cas d'attaque. Le contrat propose aussi de couvrir les pertes d'exploitation et les dommages et intérêts.

http://www.hiscox.fr/courtage/HS/Professionnels/DataRisks/tabid/982/Default.aspx

d) CNA

CNA est un assureur Américain spécialisé dans les produits à destination des professionnels et des entreprises.

L'entreprise propose un contrat appelé « NetProtect » qui offre en particulier une couverture des frais de notification, des dommages aux tiers, de la reconstruction de l'image publique et de la restauration du réseau et des données de l'assuré.

http://www.cnaeurope.com/fr-fr/Products/Technology/NetProtect/Pages/NetProtect.aspx

e) Beazley

Beazley est un assureur anglais.

Il propose une offre incluant 3 contrats sur le cyber-risque, avec une capacité maximale de 15 millions d'euros.

(1) Beazley Breach Response

Il offre des services de gestion de crise et d'expertise suite à la violation d'un système d'information. Il apporte également une prise en charge de divers frais tels que les frais de notification.

(2) Beazley Global Breach Solution

Cette offre, plus étendue que la « Beazley Breach Response », apporte en particulier une couverture des indemnités pour dommages aux tiers.

(3) Information Security & Privacy

Ce produit propose une couverture contre les accès non autorisés, le vol ou la destruction de données, les interruptions de services suite à une attaque.

https://www.beazley.com/our_business/professional_liability/tmb.html

f) ACE

ACE est un assureur suisse basé à Zurich.

Le groupe propose depuis 10 ans des produits dans l'assurance des systèmes d'informations. Le produit propose de couvrir les frais de reconstitution des données, les frais d'expertise pour l'analyse et la réparation suite à un incident, les pertes d'exploitation, les frais de restauration de l'image de marque et aussi les préjudices et pénalité liés à la divulgation de données confidentielles.

http://www.acegroup.com/fr-fr/entreprises-et-collectivites/risques-informatiques.aspx

g) Axa

Axa propose deux types de contrats, Cyber Sphere et Cyber@Risk.

Le premier, Cyber Sphere, propose aux entreprises une couverture contre les dommages personnels et les dommages aux tiers dans le cas d'une attaque informatique. Il existe une extension qui couvre aussi la fraude.

Le second contrat, Cyber@Risk, proposé en partenariat avec Cassidian, apporte aux entreprises un service de prévention et d'analyse des cyber-risques.

http://www.axa-corporatesolutions.com/Cyber-risks-wake-up-call-for.html

h) Allianz

Allianz propose en France et dans le monde « Allianz Cyber Data Protect » qui est une solution d'assurance adressant les cyber-risques construite en partenariat avec le groupe Thales.

Cette offre d'assurance propose des services d'aide à la prévention contre les menaces et des services d'intervention rapide pour répondre à une attaque en partenariat avec Thales. Le groupe Thales apporte ici son expertise en sécurité informatique.

De plus, Allianz offre des garanties couvrant les pertes financières, les dommages et les responsabilités civiles avec une capacité maximale de 50 millions d'euros.

https://www.thalesgroup.com/sites/default/files/asset/document/pr_-_allianz_teams_up_with_thales_to_offer_allianz_cyber_data_protect.pdf

i) American International Group / Chartis

AIG est un assureur américain, présent mondialement et en particulier en France. Le siège européen est basé à la Défense, en région parisienne.

AIG propose des contrats en cyber-risque couvrant la fuite et la perte de données, avec la mise à disposition d'experts du secteur. Fort de 15 ans d'expérience sur ce type de risque, l'entreprise offre un service de prévention mais également l'accompagnement des clients

lors de sinistres sur les aspects à la fois techniques, légaux mais aussi sur la gestion de la communication (pour protéger la réputation de l'entreprise).

La couverture se décompose ainsi :

- Gestion de crise
  - Atteinte à la réputation de la société
  - Atteinte à la réputation individuelle
  - Frais de notification et de surveillance
- Responsabilité Civile
- Sécurité du réseau
- Enquêtes et sanctions administratives

Avec les options :

- Garantie multimédia
- Cyber-Extorsion
- Interruption du réseau

http://www.aigassurance.fr/cyberedge_3941_558547.html

2. La réassurance

Les acteurs de la réassurance qui sont impliqués sur le sujet du cyber-risque sont :

- SCOR ;
- Swiss Re ;
- Hanover Re.

Source :http://www.argusdelassurance.com/metiers/le-cyber-risque-prend-de-l-assurance.59662

3. Les courtiers

a) Marsh

Marsh est une filiale du groupe Marsh & McLennan Companies, une structure d'envergure mondiale basée aux USA proposant des services de courtage en assurance. C'est aussi la société mère de Mercer, Oliver Wyman et Guy Carpenter.

Marsh a fait un partenariat avec Sogeti pour construire une offre d'assurance basée sur des résultats d'audits. Ce processus est détaillé dans la brochure « Maîtriser les risques cyber » disponible sur le site de Marsh.

D'autre part, ce partenariat permet d'offrir un service de suivi et de correction d'un incident informatique assuré par les experts de Sogeti dans le cadre du contrat d'assurance proposé par Marsh.

En particulier, l'offre couvre la remise en service d'une infrastructure informatique suite à une attaque. Elle couvre aussi l'analyse et la mise en place de la réponse technique en cas d'intrusion et de vol d'informations.

b) AON

Aon est une entreprise anglaise. C'est un acteur majeur du courtage en assurance présent dans le monde entier.

Au travers de « Aon Risk Solutions » l'entreprise propose un service d'audit des risques informatiques ainsi que la recherche d'un contrat d'assurance adapté aux besoins.

c) Verspieren

Verspieren est une entreprise française spécialisée dans le courtage en assurance.

Ce courtier propose une offre de couverture en cyber-risque.

4. Les acteurs spécialisés

L'audit fait partie intégrante des conditions d'assurabilité. Les sociétés spécialisées dans ce domaine sont des acteurs importants de ce marché. Thales fait partie des acteurs majeurs.

D'autre part, les offres en cyber-risque nécessitent le recours à des experts en sécurité informatique qui interviennent lorsqu'un incident survient. Du fait de l'urgence de l'intervention et des compétences de pointes nécessaires pour mener à bien ce genre de missions, des partenariats entre les assureurs et les entreprises spécialisées permettent d'améliorer l'offre d'assurance et la réactivité en cas de sinistre. Nous remarquons déjà quelques partenariats connus dans la description des offres d'assurance.

D'autre part, les fournisseurs de solutions et services en cyber-sécurité permettent de répondre aux risques par des actions préventives de protection ou de détection. Ils constituent ainsi des acteurs influençant le cyber-risque des entreprises.

5. Les services disponibles chez Thales

C'est un ensemble de services destinés à aider les entreprises à sécuriser leur système d'information et à fournir une réponse adaptée en cas d'attaque. Ces services peuvent donc être proposés au travers ou en complément d'un contrat d'assurance cyber-risque.

Nous allons donc en faire la présentation ici.

a) Thales e-Security

C'est un ensemble de services proposé par Thales dans le monde entier.

Nom	Description
Cloud Computing Security	Permet de s'assurer que les produits de *cloud computing* utilisés sont bien sécurisés.
Controlling Fraud and Protecting Intellectual Property	Aide contre les fraudes et la protection de la propriété intellectuelle.
Data Breach Notification	Aide à l'information des bonnes personnes en cas de vol de données.
Data Security and Protection Strategy	Aide à l'élaboration d'une stratégie de gestion des risques informatiques et aide aux choix technologiques.
PCI DSS Auditing and Compliance	Aide au respect des obligations de la norme PCI DSS.
Privacy Compliance	Aide au respect d'obligation de protection des données et systèmes d'informations.
Protecting Applications from Malware and APTs	Aide à se prémunir contre les attaques informatiques sur des systèmes critiques.
Trust and Identity Management	Aide à la gestion des indentifications et à la gestion des droits d'accès.

b) Audit de sécurité

Les consultants accompagnent les clients dans toutes les démarches d'évaluation du niveau de sécurité de tout ou partie de leur système d'information. L'évaluation porte à la fois sur l'organisation et les solutions techniques utilisées.

c) Test d'intrusion

Afin d'évaluer la résistance des infrastructures IT ainsi que l'efficacité des produits de sécurité implémentés chez ses clients, Thales réalise, à leur demande, des tests d'intrusion sur leurs systèmes. Ce service permet de mieux comprendre comment les vulnérabilités peuvent être exploitées par un attaquant dont l'objectif est de dérober des informations, perturber voire d'interrompre l'activité de la cible visée.

d) Conseil en sécurité

Les équipes de consultants de Thales accompagnent et conseillent les organisations du monde entier dans la définition et la mise en œuvre de la politique de sécurité de leurs systèmes d'informations en cohérence avec les missions et les contraintes propres à chaque domaine.

e) Force d'intervention rapide

Une cyber-attaque peut passer inaperçue de nombreuses semaines, si ce n'est plusieurs années. L'entreprise est alors méticuleusement pillée et ce n'est souvent qu'à l'occasion d'un événement mineur que l'ampleur du désastre apparaît. Les mesures à prendre sont alors de première importance et elles doivent permettre de reprendre le contrôle de la situation sans

risque supplémentaire et en obtenant le maximum d'informations sur les conditions de l'attaque.

f) CYBELS

CYBELS est une solution de cyber-sécurité dédiée au traitement dynamique des risques.

Nom	**Description**
CYBELS Intelligence	Cette solution surveille en continu un ensemble de blogs de hackers à des fins de prévention et d'anticipation des menaces. De plus, elle est utilisée dans le contexte de la cyber-sécurité comme cellule de veille et de prévention de la menace.
CYBELS Maps	Cette solution cartographie le système d'information et recense les vulnérabilités. Cela permet de construire et d'afficher les graphes d'attaques.
CYBELS Scan	Elle identifie les vulnérabilités qui fragilisent les données clients à protéger puis fournit des tableaux de bord et des indicateurs destinés à assurer un suivi périodique du niveau de sécurité.
CYBELS Sensor	Cette solution aide les clients à analyser le trafic et à renforcer la sécurité de leur réseau. CYBELS Sensor détecte les activités malveillantes, repère l'origine du problème et recueille des preuves à des fins d'investigation légale.
CYBELS View	Elle rassemble des informations de différents types et différentes sources, apportant des capacités dynamiques de traitement du risque, tout en donnant une vue globale de la situation de sécurité. Elle permet d'évaluer l'activité de l'entreprise et les impacts d'une cyber-attaque sur les structures organisationnelles, suggérant en temps opportun les réponses les plus adaptées.
CYBELS Practice	Ce produit permet aux administrateurs informatiques de modéliser et de déployer une simulation du système d'information qu'ils supervisent.

g) SOC

Le « Security Operations Centre » est une solution de service, permettant de centraliser toutes les traces sur les accès aux infrastructures et applications informatiques des clients, ainsi que sur les flux de données échangées, pour ensuite les corréler, en déduire les risques d'intrusion probables, analyser le parcours et l'impact des incidents de sécurité confirmés, et enfin coordonner la réponse. Il inclut généralement un service d'archivage des données à des fins statistiques, de « re-jeu » des attaques, ou comme source de preuve.

H. Le marché potentiel

D'après Daljitt Barn, Directeur en cyber sécurité chez PriceWaterhouseCoopers, le montant des primes émises pour les contrats cyber-risque aux USA cette année est de 2,2 milliards d'euros alors qu'en Europe ce montant est de 220 millions d'euros. On peut donc penser que le marché actuel de l'assurance cyber-risque en Europe est approximativement 10 fois inférieur à son potentiel.

On remarque en 2014 une augmentation de 15% sur une année du coût moyen par entreprise des fuites de données (Ponemon, 2014). Cette augmentation du risque a poussé les entreprises à une prise en compte plus rigoureuse en poussant le suivi de ce sujet directement au niveau du comité de direction.

II. Normes, méthodes et modèles techniques

En première partie nous avons présenté le cadre technique du risque étudié puis le marché de ce type de contrats.

Nous allons apporter ici des outils pour l'évaluation du risque par le **souscripteur**. Nous allons commencer par lister les normes existantes dans ce secteur qui peuvent servir de paramètres aux modèles actuariels. Puis nous nous intéresserons au cadre règlementaire. Enfin, nous finirons par les modèles techniques qui permettent une évaluation du risque passé pour permettre de mieux connaître le risque futur.

A. Normes et méthodes

L'objectif ici est de recenser les normes et les référentiels utilisés par l'industrie qui peuvent constituer pour l'actuaire et le souscripteur des indicateurs sur le risque apporté par une entreprise.

Nom	**Organisme**
Les Normes ISO 27000	ISO
PCI DSS	grandes entreprises de cartes de débit et de crédit
Méthode MEHARI	CLUSIF (Français)
Méthode EBIOS	ANSSI (Français)
Méthodes OCTAVE	Software Engineering Institute (USA)
Référentiel SP800-30	National Institute of Standards and Technology (USA)
COBIT 5 / Risk IT Framework	Information Systems Audit and Control Association (USA)
SOC 2	American Institute of Certified Public Accountants
Cyber essentials scheme	Gouvernement du Royaume-Uni

Notons que toutes ces normes et référentiels proposent ou imposent de faire une évaluation du risque informatique, mais ne proposent pas de méthode pour en faire l'évaluation. Leur premier objectif est de définir une organisation permettant de maitriser ce risque.

1. Les Normes ISO 27000

C'est un ensemble de normes dont seule la sous norme ISO 27001 permet une certification.

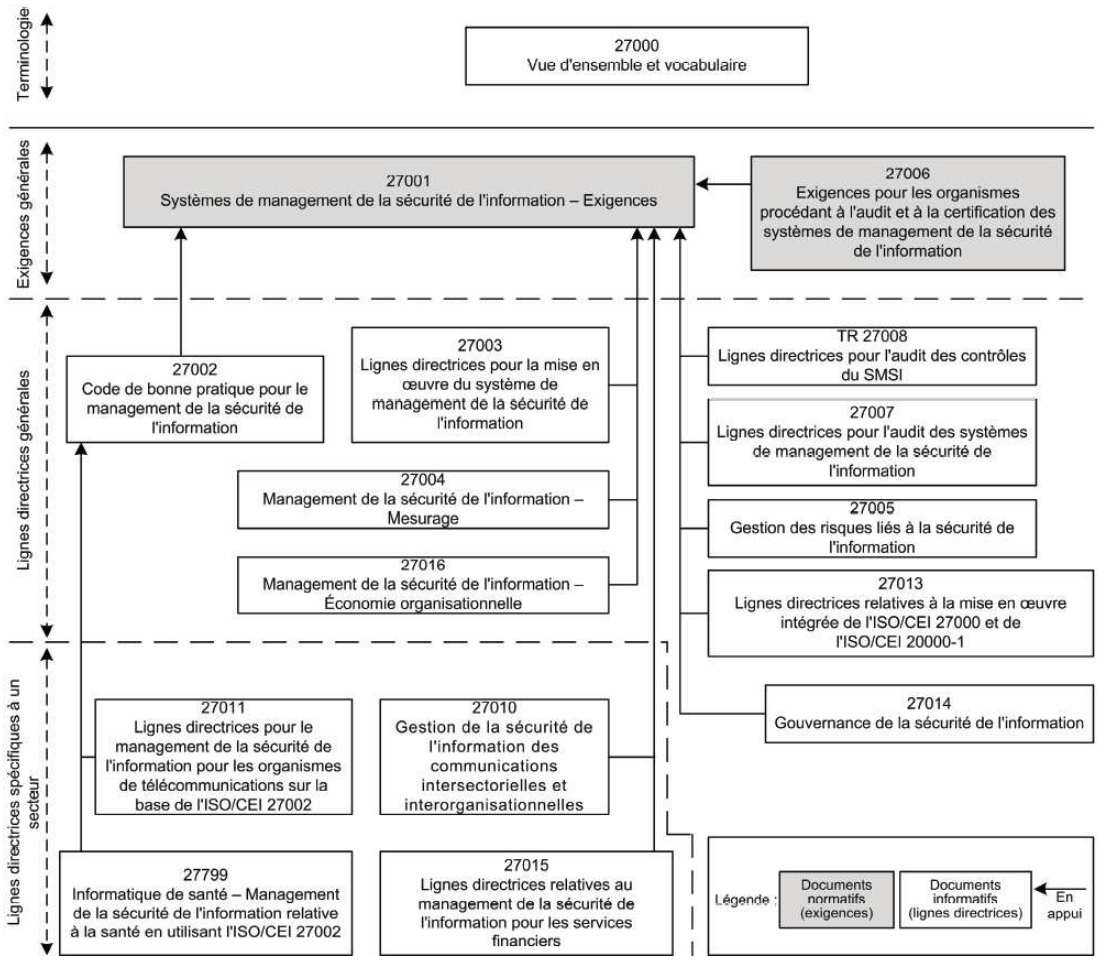


Figure 3 Organisation des normes ISO 27000

a) ISO 27001

Norme de certification des systèmes de management de la sécurité de l'information.

Cette norme prévoit une évaluation des risques à l'étape 2, cependant la méthode n'est pas précisée dans la norme.

On constate une nette augmentation du nombre de certifications au cours des dernières années.

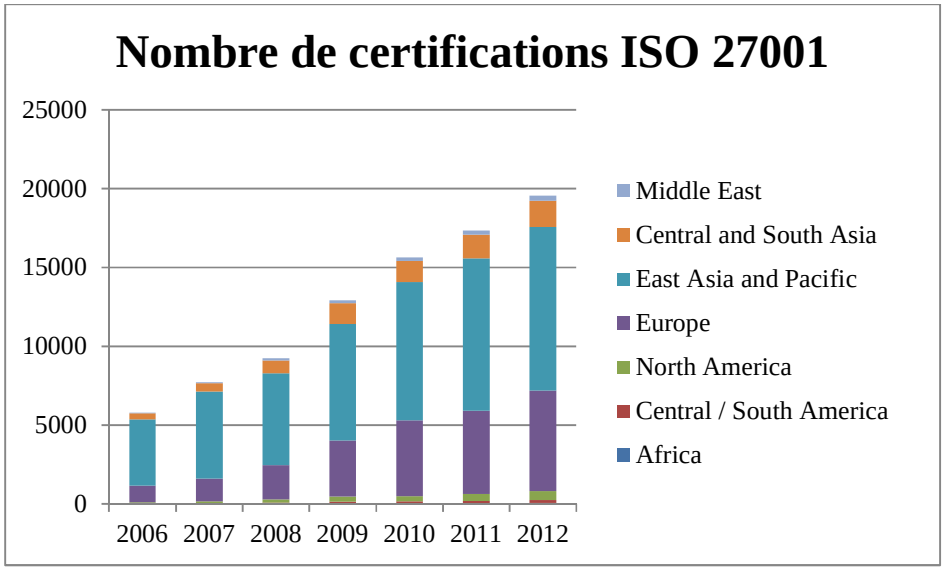


Figure 4 Pénétration de la norme ISO 27001

b) ISO 27002

C'est un code de bonne pratique pour le management de la sécurité de l'information, permettant de préparer la certification ISO27001.

La dernière mise à jour de cette norme a été faite en 2013, elle est maintenant composée de 15 chapitres (numérotés de 4 à 18) et détaille 114 points de contrôle (la version 2005 avait 133 points de contrôle) découpés en 35 catégories de sécurités.

Voici la liste des chapitres :

- Structure de la norme
- Politiques de sécurité de l'information
- Organisation de la sécurité de l'information
- La sécurité des ressources humaines
- Gestion des actifs
- Contrôle d'accès
- Cryptographie
- Sécurité physique et environnementale

- Sécurité liée à l'exploitation
- Sécurité des communications
- Acquisition, développement et maintenance des systèmes d'informations
- Relations avec les fournisseurs
- Gestion des incidents liés à la sécurité de l'information
- Aspects de la sécurité de l'information dans la gestion de la continuité de l'activité
- Conformité

c) ISO 27004

Cette norme a pour but de préciser comment mesurer et rapporter l'efficacité du système de management de la sécurité de l'information.

d) ISO 27005

Cette norme a pour but de préciser comment gérer les risques en sécurité des systèmes d'informations.

Elle est composée des chapitres suivants :

- Vue générale du processus de gestion des risques en sécurité de l'information
- Description du contexte
- Appréciation des risques
- Traitement des risques en sécurité de l'information
    - Modification du risque
    - Rétention du risque
    - Evitement du risque
    - Partage du risque
- Acceptation du risque en sécurité de l'information
- Consultation et communication sur le risque
- Surveillance et réexamen du risque

Notons que le chapitre 8.3.2 traite de l'évaluation des conséquences d'un incident ; le chapitre 8.3.3 traite de la probabilité de survenue d'un incident ; le chapitre 8.3.4 traite du calcul du niveau de risque.

2. PCI DSS

La norme « *Payment Card Industry Data Security Standard* » ou PCI DSS est un standard de sécurité des données imposé par les plus grandes entreprises de cartes de débit et de crédit aux entreprises de leur réseau. On peut compter dans le consortium Visa, MasterCard et American Express. Cette norme est régulièrement mise à jour, la dernière publication de la norme est la version 3.0 datant de novembre 2013.

Le standard est décomposé en 12 exigences :

1. Installer et gérer une configuration de pare-feu pour protéger les données du titulaire.
2. Ne pas utiliser les mots de passe système et autres paramètres de sécurité par défaut définis par le fournisseur.
3. Protéger les données du titulaire stockées.
4. Crypter la transmission des données du titulaire sur les réseaux publics ouverts.
5. Protéger tous les systèmes contre les logiciels malveillants et mettre à jour régulièrement les logiciels anti-virus ou programmes.
6. Développer et gérer des systèmes et des applications sécurisés.
7. Restreindre l'accès aux données du titulaire aux seuls individus qui doivent les connaître.
8. Identifier et authentifier l'accès aux composants du système.
9. Restreindre l'accès physique aux données du titulaire.
10. Effectuer le suivi et surveiller tous les accès aux ressources réseau et aux données du titulaire.
11. Tester régulièrement les processus et les systèmes de sécurité.
12. Maintenir une politique qui adresse les informations de sécurité pour l'ensemble du personnel.

3. Référentiel SP800-30

Ce référentiel est défini par le « *National Institute of Standards and Technology* » ou NIST. Il s'agit de l'organisme américain en charge de la définition des mesures et métriques.

Ce référentiel a pour but de définir une méthode d'évaluation du risque informatique et donc du cyber-risque. Il définit 9 étapes dans l'évaluation du risque :

1. Caractériser le système
2. Identifier les menaces
3. Identifier les vulnérabilités
4. Analyser les contrôles
5. Déterminer les probabilités de réalisation
6. Analyser les impacts
7. Déterminer les risques
8. Recommander des contrôles
9. Documenter les résultats

4. Méthode MEHARI

La Méthode harmonisée d'analyse des risques ou MEHARI a été développée par le Club de la Sécurité de l'Information Français ou CLUSIF. Cette méthode a pour but d'aider les Responsables de la Sécurité des Systèmes d'Informations ou RSSI (dit aussi CISO en anglais) dans leur tâche de gestion et de pilotage de la sécurité des systèmes d'informations.

L'objectif principal de la méthode est l'analyse et la gestion de risques. Le CLUSIF fournit, en même temps que la méthode, un outil et une base de connaissances au format Excel.

a) Base de connaissance et Outil

La feuille « Score ISO » de l'outil donne un score pour chacun des 133 points de contrôle de la norme ISO 27002 : 2005. De plus, elle est conforme aux préconisations des normes ISO 27005 et SP 800-30.

b) Outil Risicare

Cet outil développé par la société BUC et construit à partir de la méthode MEHARI. Il réalise une analyse de risque dans le cadre de la mise en œuvre d'un SMSI selon les normes ISO 27001 et ISO 27005.

5. Méthode EBIOS

Cette méthode a été développée par l'Agence Nationale de la Sécurité des Systèmes d'Informations. Elle permet en particulier d'apprécier les risques de Sécurité des systèmes d'informations (Wiki2).

Cependant cette méthode ne quantifie pas le risque. Elle ne répond donc pas pleinement aux besoins de l'actuaire.

6. Méthodes OCTAVE

La méthode « *Opérationally Critical Threat, Asset, and Vulnerability Evaluation* » ou OCTAVE est développée et améliorée depuis 1999. Cette méthode a été créée par le « Software Engineering Institute » ou SEI de l'Université de Carnegie Mellon pour répondre aux besoins de certification rencontré par le département de la défense américain. Le centre de recherche qui a développé cette méthode est financé par le département de la défense américain.

Historique	
Juin 1999	OCTAVE Framework 1.0
Septembre 2001	OCTAVE Method 2.0
Décembre 2001	OCTAVE Criteria 2.0
Septembre 2003	OCTAVE-S 0.9
Mars 2005	OCTAVE-S 1.0
Juin 2007	OCTAVE Allegro 1.0

La méthode OCTAVE s'adresse plutôt aux organisations de plus de 300 personnes avec des équipes de gestion du système d'information dédiées.

La méthode OCTAVE-S s'adresse plutôt aux organisations de plus de 100 personnes avec des équipes de gestion du système d'information qui ont déjà un bonne connaissance des métiers et de l'organisation de la structure.

La méthode OCTAVE Allegro a pour but de rationaliser et d'optimiser le processus. L'objectif est d'obtenir un niveau de sécurité satisfaisant tout en limitant les coûts et les délais de déploiement.

7. COBIT 5 / Risk IT Framework

Le « *Contrôle Objectives for Information and related Technology* » ou COBIT en version 5 est un référentiel pour la gestion et le pilotage technique des systèmes d'informations. Ce référentiel vise principalement à aider les manageurs à gérer les risques et les investissements.

Le référentiel Risk IT est construit en complément du COBIT 5 et définit un ensemble de bonnes pratiques pour identifier, gouverner et gérer les risques IT en entreprise.

Ces 2 référentiels sont réalisés par le « *Information Systems Audit and Control Association* ».

Le référentiel Risk IT définit 3 catégories de risques :

- Capacité à récupérer ou non des avantages de l'utilisation des technologies de l'information. Le risque pour l'entreprise et de ne pas profiter du gain de performance et donc de faire des investissements sous-optimaux.
- Livraison de programme et projet IT.
- Le fonctionnement ou non des produits et services IT conformément aux attentes.

On remarque que le seul point qui nous intéresse est le troisième.

D'autre part, le référentiel stipule que l'entreprise doit faire une évaluation des risques à la fois en fréquence et en impact, mais ne précise pas comment.

L'utilisation de ce référentiel dénote une implication de la gouvernance de l'entreprise dans la gestion du cyber-risque. Elle peut donc constituer un indicateur du risque porté par l'entreprise.

8. SOC 2

La norme « Service Organisation Controls 2 » ou SOC 2 est développée et conçue par le « American Institute of Certified Public Accountants » (AICPA). Ce standard définit un ensemble de rapports qui évaluent les contrôles faits sur l'organisation d'un service relevant de :

- La sécurité, le système est-il protégé contre les accès non autorisés ?
- La disponibilité, le système est disponible et opérationnel aux moments convenus.
- L'intégrité, les opérations sont exécutées intégralement, conformes dans le temps imparti et sur des données donc l'accès est autorisé.
- La confidentialité, les informations confidentielles sont protégées comme convenu.
- Du respect de la vie privée.

9. Cyber essentials scheme

Le gouvernement du Royaume-Uni a développé un référentiel définissant les règles de base d'organisation et de contrôle technique que les entreprises anglaises devraient suivre dans le cadre de la protection des données. Ce référentiel a été publié le 5 juin 2014 avec le support de plusieurs sociétés d'assurance (DBIS, 2014).

Le premier objectif est de définir un ensemble de pratiques et règles simples permettant d'aider à contrôler le cyber-risque. La rédaction de ces règles a été faite en gardant à l'esprit que leur application doit donner un niveau de protection suffisant tout en limitant les coûts au strict minimum. Ainsi, même si ces règles ne garantissent pas le meilleur niveau de sécurité possible, les rédacteurs ont essayé de garder un équilibre entre la protection des données, la simplicité d'application, la maitrise des coûts et le respect des organisations en place.

Le second objectif est d'apporter une certification qui puisse être utilisée dans la plupart des industries pour informer les consommateurs, les partenaires et les assurances de la bonne application du référentiel. Pour faciliter l'adoption du standard, celui-ci a été décliné en 2 versions, *Cyber Essentials* et *Cyber Essentials Plus*, qui offrent 2 niveaux de protection en fonction des besoins de l'entreprise.

Ce référentiel n'a pas été construit avec pour objectif d'être imposé aux entreprises par la voie légale, mais simplement afin qu'il devienne un standard du marché. L'utilisation par la majeure partie des entreprises anglaises permettrait de le rendre presque obligatoire pour pouvoir travailler avec des partenaires basés dans le royaume.

B. Réglementation

1. Obligations et responsabilités

a) USA

De nombreux états ont des lois encadrant l'utilisation de données personnelles, qui viennent compléter la réglementation fédérale. Cette dernière a été écrite pour différents secteurs et selon des lois séparées, mais couvrent un ensemble important d'activités.

Les données de santé sont fortement réglementées. D'autre part, les activités financières et les organismes gouvernementaux sont soumis à des contraintes spécifiques de protection des données et de notification des personnes concernées en cas de vol ou de perte de données.

b) Europe

La directive 95/46/EC du 24 Octobre 1995 définit les obligations d'information du consommateur sur la récupération de données personnelles ainsi que les règles de protection et de transfert de ces données.

Il faut noter que, dans cette réglementation, on désigne par « donnée personnelles » toute information relative à une personne identifiée ou identifiable, ce qui constitue une définition extrêmement vaste.

D'autre part, cette réglementation impose des règles strictes aux entreprises qui font sortir des données du territoire européen. Les données ne peuvent sortir que dans des pays offrant une réglementation aussi contraignante que la règlementation européenne sur la protection des données. En particulier, les USA n'ont pas une réglementation assez proche, ce qui a imposé la construction d'un accord spécifique. Du fait du programme de surveillance PRISM, l'accord doit être évalué par la cour de justice de l'Union Européenne qui pourrait l'invalider.

2. Obligation de notification

a) USA

Actuellement 46 états sur les 50 que compte le pays ont des lois imposant la notification d'un vol de données. Ces lois précisent les actions à mener lorsqu'un vol de données touche l'un de leurs résidents. La période de prise d'effet de ces lois s'étale de 2005 à 2010. Au vu du nombre d'état concernés et de l'interconnexion économique entre les états du pays, on peut considérer que presque toutes les violations de données aux Etats-Unis sont soumises à une obligation de notification.

D'autre part, le gouvernement américain étudie l'élaboration d'une loi fédérale qui rendrait l'obligation de notification et le protocole de notification identiques sur tout le territoire de l'union.

b) Europe

Les obligations de notification sont pour le moment non harmonisées au sein de l'union européenne. D'autre part peu d'états ont une obligation de notification publique pour toutes les entreprises et institutions. Il n'y a donc pas les mêmes sources de données en Europe que ce qu'on peut trouver aux USA.

Cette réglementation doit être amenée à changer au cours de l'année 2014, pour instaurer une obligation de notification dans l'ensemble de l'Europe.

C. Modèles basés sur la topologie du réseau

Ces différents modèles ont pour but d'évaluer le risque de panne d'un système complexe. Les événements conduisant à l'incident peuvent avoir des sources variées :

- défaillance matérielle ;
- erreur non intentionnelle ;
- attaque.

La typologie du réseau doit être cartographiée au préalable. Il apparait donc une limite à ces modèles. Lorsque l'organisation évolue et que la topologie du réseau change, l'évaluation du risque doit être refaite.

1. Graphe d'attaque

On représente le système comme un ensemble de nœuds qui sont soit des éléments logiques, soit des éléments physiques. Les arêtes sont les liaisons qui peuvent exister entre eux.

Une attaque est alors un des chemins du graphe entre un point d'entrée et un élément stratégique du système d'information.

On peut alors faire une évaluation quantitative des risques (MICONNET, et al., 2013). On évalue le risque d'apparition d'une vulnérabilité, ce qui correspond à une transition dans le graphe du système. On évalue le risque qu'un attaquant trouve la faille, en fonction de la difficulté d'exploiter la faille. Enfin on évalue la gravité de chaque évènement indésirable.

On peut donc évaluer dans le modèle à la fois la gravité ou le coût du sinistre et la fréquence de chaque sinistre.

Notons ici la simplification du système d'information, en ne prenant pas en compte l'aspect temporel et la dissymétrie possible des liaisons informatiques. D'autre part, l'évaluation des différentes probabilités utilisées dans le modèle n'est pas expliquée.

2. CHASSIS

La méthode CHASSIS (Combined Harm Assessment of Safety and Security for Information Systems) constitue une extension du langage UML (RASPOTNIG, et al., 2012).

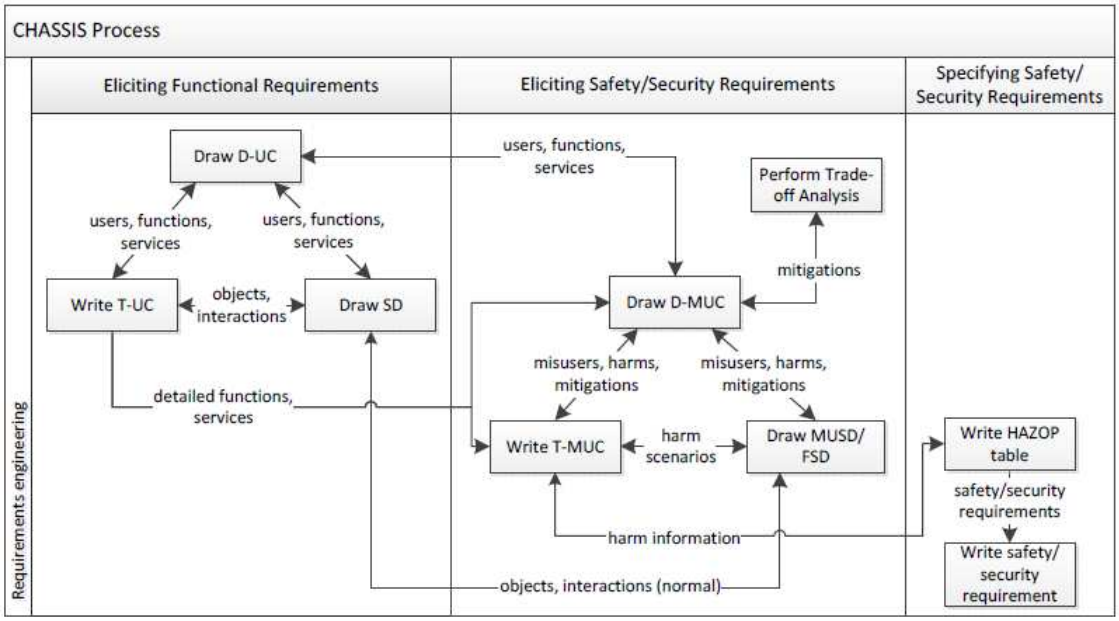


Figure 5 Méthode CHASSIS

Abréviation	Description
D-MUC	Diagrammatical misuse case
D-UC	Diagrammatical use case
FSD	Failure sequence diagram
HAZOP	Hazard and Operability study
MUSD	Misuse sequence diagram
SD	Sequence diagram
T-MUC	Textual misuse case
T-UC	Textual use case

3. Arbre de défaillance / Arbre d'attaque

On représente l'ensemble des évènements menant à une défaillance (sommet de l'arbre) en représentant les combinaisons successives qui conduisent à l'évènement indésirable (Wiki1).

On parle ici d'arbre à logique booléenne.

Ce genre de modèle permet de connaître les scénarios d'attaque ou de défaillance, mais ne permet pas d'en évaluer les probabilités.

Le modèle suivant, qui constitue une évolution du modèle des arbres de défaillance, apporte cette possibilité.

4. BDMP (Boolean logic Driven Markov Processes)

Il est inspiré du modèle précédent, mais permet en plus de supporter des scénarios dynamiques (dépendant de paramètres temporels) et propose en plus une évaluation quantitative du risque global et des scénarios distincts.

Ce modèle a été créé par les équipes de recherche d'EDF, qui fournit aussi une suite de logiciels permettant de faciliter la mise en œuvre pratique de cette méthode. On peut retrouver les outils à l'adresse : http://visualfigaro.sourceforge.net

Voici une capture d'écran de cet outil.

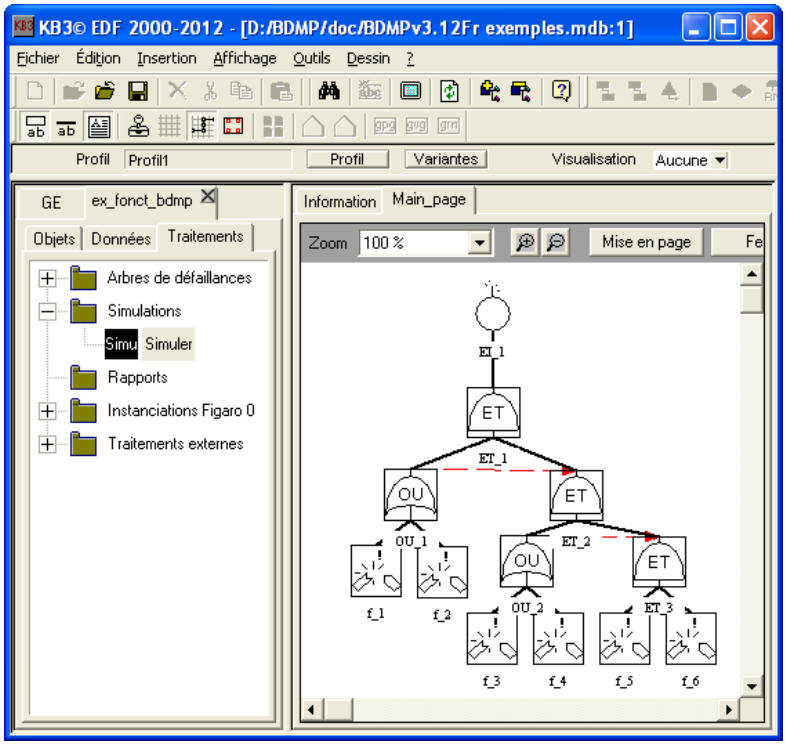


Figure 6 Logiciel K3B pour faire du BDMP

Le principe est de construire un chemin allant des feuilles jusqu'à la racine de l'arbre. Chaque feuille a une certaine fonction de distribution d'activation. Ces fonctions sont ici notées f_1, f_2, f_3 … Ces fonctions permettent de simuler le délai d'activation. Lorsque la feuille est active, elle a alors une probabilité de se désactiver en fonction du temps.

Les feuilles sont reliées à la racine par des portes logiques qui s'activent lorsque les branches nécessaires sont activées. Dans l'exemple, on trouve des portes « OU » qui sont actives lorsqu'au moins une branche est active, et les portes « ET » qui sont actives lorsque toutes les branches le sont.

Enfin, les traits rouges sont des gâchettes. Lorsque la porte OU_1 est active, la porte ET_2 est sollicitée, donc la porte OU_2 est aussi sollicitée. Par contre, la porte ET_3 étant la cible d'une gâchette, n'est pas sollicitée. Une porte ne peut devenir active que lorsqu'elle est sollicitée, ce qui représente l'idée qu'un outil qui n'est pas utilisé ne peut pas servir pour une attaque ou ne peut pas tomber en panne.

On peut modéliser une attaque précise bien connue, ou alors un ensemble d'attaques connues.

Ce modèle a été proposé en 2003 dans le cadre d'outils de la gestion des risques chez EDF (BOUISSOU, et al., 2003).

On peut retrouver une modélisation de l'attaque Stuxnet (BOUISSOU, et al.)

5. Comment avoir la topologie du réseau

Pour de grosses entreprises, les réseaux internes peuvent avoir une topologie extrêmement complexe. Il peut donc devenir très difficile d'en connaitre la structure et ce afin d'alimenter les modèles.

a) Par une cartographie manuelle

On regarde manuellement chaque machine du réseau. Cela est bien adapté aux très petites structures, les PME.

b) Par une exploration dynamique

Des logiciels tels que Nmap peuvent cartographier un réseau de façon autonome en envoyant des requêtes (paquets) bien choisis.

Avantages : simples à mettre en place, et ne nécessite pas de préparation préalable des nœuds du réseau.

Inconvénients : peut être perçu comme intrusif dans le système, ne passe pas les routeurs et ne détecte pas les machines éteintes au moment de la cartographie.

c) Par une exploration statique

Thales possède une technologie propriétaire qui permet de reconstruire la cartographie du réseau à l'aide des traces informatiques disponibles.

Avantages : non intrusif, non influencé par les routeurs et autres barrages et détecte toutes les machines qui ont agi sur le réseau.

Inconvénients : la mise en place demande des compétences spécialisées (présentes chez Thales), exige un bon paramétrage des traces avant l'étude et ne détecte pas les machines sans activité.

D. Modèles sans topologie : Approche Multistate

Ce modèle permet d'obtenir plus facilement une évaluation des risques que les précédents, mais il ne prend pas en compte certains facteurs de risque. Il peut être plus adapté aux entreprises de taille modeste.

Ce modèle s'inspire de ce qui est fait en assurance santé (BARRACCHINI, et al., 2013).

Le principe est de référencer tous les états du système, avec en particulier les états avec un dommage. Ensuite, on cherche les probabilités de transition entre états. Il faut noter qu'il y a 3 grands groupes d'états possibles :

- aucun incident ;
- incident partiellement réparable ;
- incident non réparable.

Par la suite on attribue un âge à chaque élément du système pour calculer les probabilités de transitions. En partant de l'hypothèse : « un programme a moins de failles lorsqu'il est vieux que lorsqu'il est jeune ». On attribue un âge aux éléments en fonction de leur date d'installation.

E. Conclusion

On a présenté plusieurs approches pour répondre à des besoins variés d'évaluation du risque connu. Les modèles basés sur la topologie du réseau qui sont plutôt adaptés à de grosses structures. Les graphes d'attaque et la méthode CHASSIS sont conseillés pour de grands systèmes assez homogènes d'un côté. Les arbres et la méthode BDMP sont conseillés pour les systèmes complexes et peu homogènes. Enfin, l'approche multistate va s'adapter aux petits systèmes simples.

Aucun de ces modèles n'est destiné à faire une évaluation actuarielle du risque, ils peuvent cependant servir d'indicateurs.

III. Modélisation actuarielle

Nous avons présenté le secteur étudié ainsi que les indicateurs utilisables par un actuaire ou un souscripteur dans les des deux premières parties.

Nous allons commencer par une recherche des travaux existants, puis nous présenterons les données disponibles. Enfin, nous réaliserons l'étude actuarielle de ces données afin de produire une prime pure.

A. Modèles existants

1. Propagation Virale

Il est uniquement question ici de virus informatiques. Le problème des virus informatiques subit un ralentissement depuis quelques années, suite au renforcement des mécanismes de sécurité sur les postes utilisateurs et chez les particuliers. Mais il constitue souvent une des portes d'entrée dans un système.

L'article sur le sujet de 2004 (SERAZZI, et al., 2004) donne un modèle et des comparaisons. On constate en particulier que la propagation est un phénomène exponentiel, qui comporte donc un seuil de rupture au-delà duquel la contamination devient rapidement générale. En effet, l'article montre que la probabilité de contamination en fonction du temps suit une courbe logistique. Cette forme de probabilité de contamination correspond à une diffusion exponentielle sur une population finie qui se protège.

Tout comme en biologie où le vaccin a le même rôle que l'antivirus en informatique, la protection d'une partie de la population peut protéger l'ensemble de la population. En effet, la part de la population protégée permet de conserver le taux de contamination en dessous du seuil de rupture. Cela montre l'importance de l'incitation à la mise en place d'une protection. En effet, comme pour le vaccin, l'individu pris isolément trouve un intérêt faible à se faire vacciner. Alors que la communauté a un intérêt fort à ce qu'une grosse partie de la population soit vaccinée.

2. Modèle économique

Les modèles microéconomiques montrent une réallocation des ressources des agresseurs lorsqu'on augmente le niveau de sécurité. Ils ont alors de bonnes chances de se tourner vers d'autres activités. En effet, une sécurisation augmente le coût de la ressource pour l'attaquant sans changer sa fonction d'utilité. On peut donc en conclure qu'une augmentation globale de la sécurité informatique va diminuer le risque. De plus, la réallocation de ressources devrait également favoriser l'expansion d'activités économiques socialement plus appréciées.

Les approches de modélisation sont nombreuses pour comprendre les choix faits par les acheteurs d'une couverture du risque cyber. L'article définissant un Framework pour la cyber-assurance de 2010 (BÖHME, et al., 2010) fournit une classification de ces études. Ces

études concluent globalement sur l'existence d'un marché, et au besoin de développement d'une offre adaptée.

Ces études montrent les points suivants :

- L'assurance en cyber-risque comporte les problèmes classiques d'aléa moral et d'anti-sélection du fait de l'asymétrie d'information. Il est donc nécessaire de mettre en place des audits réguliers et des franchises sur les contrats.
- Les risques sont interdépendants du fait de l'interconnexion entre les systèmes d'informations. Les assureurs ont intérêt à favoriser les investissements en protection contre les cyber-attaques.
- Le client peut faire un arbitrage entre une augmentation de la protection et souscrire une assurance. Il est donc intéressant de combiner les deux offres par le biais de partenariats.

3. Queue épaisse

Le problème des queues épaisses a été abordé en 2009 sur le volume de données volées (MAILLART, et al., 2009). Les données utilisées proviennent de l'Open Security Foundation, qui ne fournit pas librement ces données. Ils ont étudié 956 évènements entre 2000 et 2008 dans le monde entier même si la plupart est localisé aux Etats-Unis. 956 évènements qui se sont produits entre 2000 et 2008 dans le monde entier dont la grande majorité aux Etats-Unis ont été étudiés.

La loi vérifie une forme $1 - F(x) = \left(\frac{u}{x}\right)^b$ lorsque $x \geq u$. Ici $u = 7.10^4$ et $b = 0.7 \pm 0.1$.

Les auteurs expliquent le seuil u par l'absence de déclaration lors de petits incidents. Cela signifie que les données ne sont représentatives que pour les gros incidents.

L'article fait remarquer que l'épaisseur de la queue à droite est d'autant plus importante que le nombre d'employés de l'entreprise est grand. Autrement dit, les risques extrêmes croissent plus vite que le nombre d'employés.

4. Indépendance entre Fréquence et Sévérité

Le thème du volume de données volées a également permis de soulever le problème de l'indépendance entre la fréquence et la sévérité des sinistres (MAILLART, et al., 2009).

Les auteurs signalent une fréquence de sinistre très faible avant 2005, suivie d'une explosion du nombre en 2005 et 2006, puis une stabilisation de la fréquence.

Les auteurs remarquent une stabilité sur la fonction de répartition du nombre de données volées, malgré un changement de fréquence dans le temps. Cette remarque met en avant une indépendance entre la fréquence et la sévérité, sans pour autant apporter une preuve formelle.

5. Modèles de copule pour l'évaluation tarifaire

On peut citer, comme premier modèle, un article de 2006 portant sur l'étude du risque global et externe aux entreprises à l'aide de t-copules (BÖHME, et al., 2006).

Un autre article de 2006 s'intéresse à la corrélation entre le cyber-risque et le volume de transactions pouvant exposer des données privées (MUKHOPADHYAY, et al., 2006). Le modèle est basé sur une copule normale multivariée construite à l'aide d'un réseau bayésien.

Enfin, un article de 2007 utilise des copules archimédiennes pour évaluer le cyber-risque avec un point de vue actuariel (HERATH, et al., 2007). L'article étudie la corrélation entre le nombre de postes infectés par un virus et le coût pour l'entreprise. Remarquons que l'article affirme que les pertes assurées suivent une distribution de Weibull ce qui est discutable compte tenu du volume et du type de données utilisées.

6. Choix des indicateurs

A l'aide d'un modèle bayésien basé sur les dires d'experts, l'article (INNERHOFER-OBERPERFLER, et al., 2010) fournit un classement des indicateurs les plus pertinents pour discriminer les risques. L'étude se concentre sur les risques propres à l'entreprise et exclut les dommages aux tiers. En particulier, l'article se concentre sur l'ensemble des cyber-risques plutôt que sur le risque de vol ou de perte de données personnelles.

On remarque que l'importance du système d'information dans le fonctionnement de l'entreprise ainsi que le type de données manipulées sont les indicateurs les plus importants. Le secteur d'activité constitue un indicateur moins pertinent.

Nous remarquons que cela vient partiellement en contradiction avec les études faites par l'Institut Ponemon. Cela peut s'expliquer par un biais d'évaluation du risque commun à tous les experts.

7. Conclusion

Nous venons de voir un historique d'études faites sur le sujet. Cependant, aucun de ces travaux ne fournit une évaluation actuarielle du risque : nous ne pouvons pas construire de tarif avec.

De nombreux articles disent souffrir du manque de données mais certaines pistes sont présentées pour palier ce manque d'information. Les conclusions peuvent rentrer en contradictions entre les articles. Donc de nombreuses analyses restent à faire et les risques sont encore très mal quantifiés.

Malgré ces faiblesses, les articles présentent les problématiques existantes ainsi que certaines pistes pour les résoudre. En particulier, ce risque présente des phénomènes de corrélation entre les acteurs assurés et certaines similitudes avec la santé. Nous avons remarqué que le coût est dépendant de certains indicateurs qui peuvent servir à affiner les études lorsque l'information sur le coût est absente. Certains de ces travaux ont ainsi

constitué une base à la construction du modèle présenté dans ce mémoire, en particulier l'article de MAILLART.

B. Les données Open Data

1. Base de vulnérabilités

La base de vulnérabilités est fournie par l'administration américaine sur le site http://nvd.nist.gov/

Elle est constituée à partir des informations publiées par les développeurs de logiciels et les chercheurs en sécurité informatique. Elle recense les vulnérabilités connues, les dates de découverte et les versions des logiciels touchées. Cette base très détaillée est largement utilisée par les experts en sécurité informatique pour détecter les possibilités d'intrusion.

Voici un graphique donnant le nombre de vulnérabilités dans cette base en fonction de leur date d'ajout dans la base.

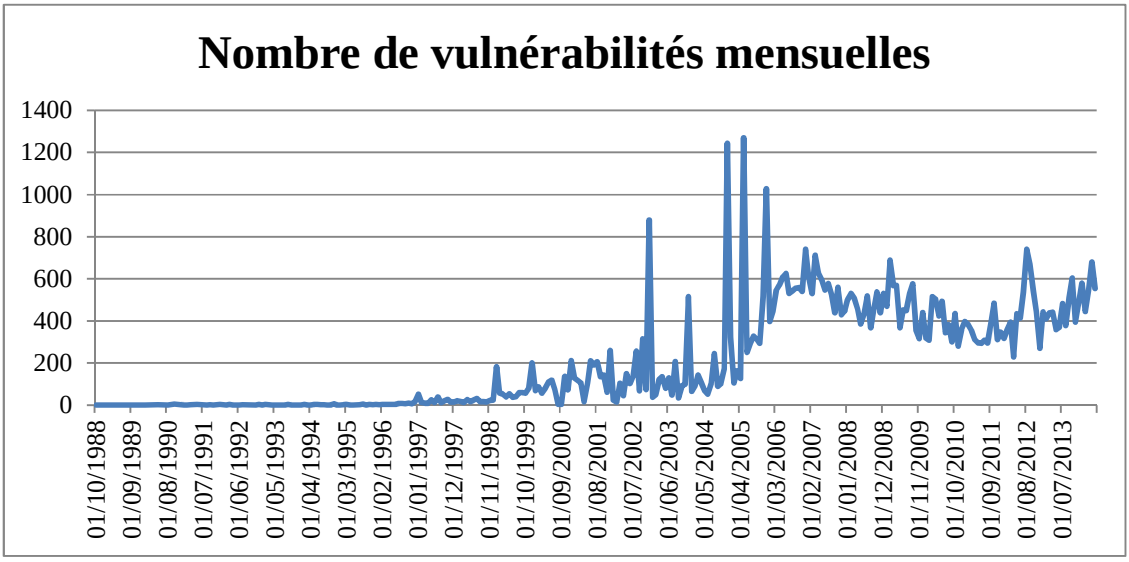


Figure 7 Evolution des vulnérabilités

On remarque une courbe en S, ce qui permet de supposer une stabilisation du nombre de nouvelles vulnérabilités.

Notons aussi une certaine similitude avec la figure 8, en particulier le seuil de 2005-2006 s'observe sur les deux courbes. On peut donc espérer une forme de corrélation entre la fréquence d'apparition des vulnérabilités et la fréquence des vols de données.

Même si nous n'exploiterons pas ce paramètre dans le modèle présenté dans ce mémoire, celui-ci peut faire l'objet de développements ultérieurs par exemple avec les équations de de Lotka-Volterra (modèle proie-prédateur).

2. Base de violation de SI

La source des données est accessible via le site suivant: https://www.privacyrights.org/data-breach.

Elle ne concerne que les incidents survenus aux USA. Il faut noter qu'avec l'obligation de déclaration des vols de données privées, cette base de données est représentative du territoire américain. Du fait de l'évolution des réglementations, on considérera comme fiables et représentatives les données à partir de 2010.

Les données disponibles sont la date de communication publique de l'incident, le nom de l'organisation, des informations de localisation géographique (souvent le siège de l'organisation), le volume de données impactées, un commentaire et 2 colonnes qui sont le type d'organisation et le type de vulnérabilité exploitée.

Les valeurs possibles sont pour le type d'organisation :

- BSO : Entreprise Autre
- BSF : Entreprise Financière ou Assurance
- BSR : Entreprise de Vente ou Marchande
- EDU : Institution du domaine de l'éducation
- GOV : Gouvernemental ou Militaire
- MED : Du domaine de la Santé
- NGO : Organisation à but non lucratif

Les valeurs possibles pour le type de vulnérabilité :

- DISC : Divulgué involontairement. Des informations sensibles laissées en accès libre sur un site web, envoyées à la mauvaise adresse mail…
- HACK : Hacking ou malware. Pénétrer, dans le système d'information d'une organisation, à l'aide d'un logiciel malveillant.
- CARD : Fraude par carte de paiement. Fraude utilisant une carte de paiement sans utilisation de méthode de hacking.
- INSD : Interne. Une personne avec des droits d'accès légitimes qui a volontairement divulgué des informations sensibles.
- PHYS : Enregistrement non électronique perdu, jeté ou volé.
- PORT : Ordinateur portable, PDA, smartphone, CD, clés USB, disque dur… perdu, jeté ou volé.
- STAT : Ordinateur non portable ou serveur perdu, jeté ou volé.
- UNKN : Non connu

Nous n'exploiterons pas les données géographiques car elles ne sont pas représentatives du lieu de l'incident ni de la réglementation en vigueur.

Le graphique suivant présente nombre de violations par mois recensées dans la base de données. Un lissage a été réalisé en moyennant sur l'année.

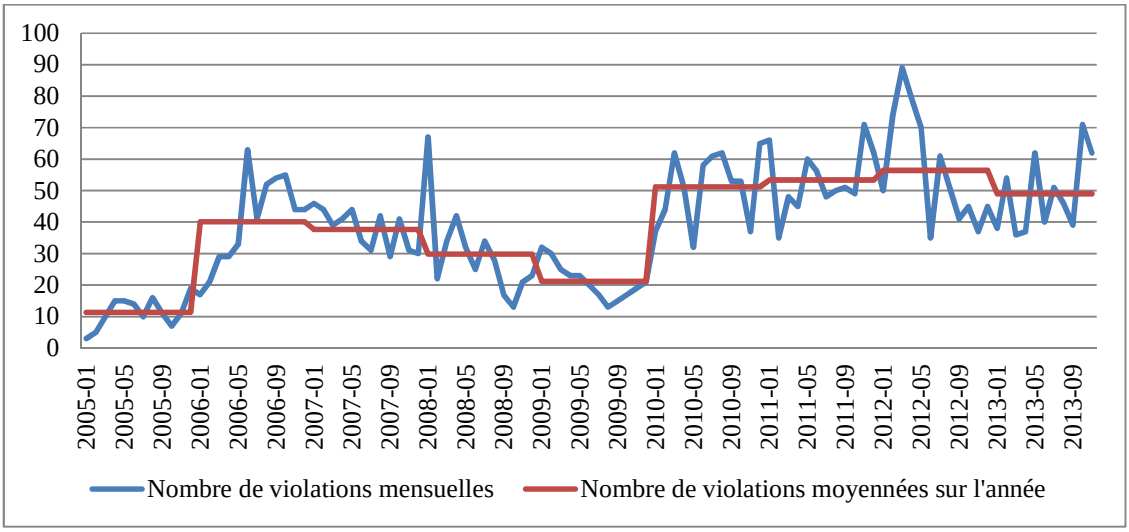


Figure 8 Evolution des nombres de violations

On remarque bien une stabilisation des volumes à partir de 2010, ce qui se justifie par la mise en place des réglementations. Le graphique ci-dessous représente le nombre de mises en application de nouvelles réglementations aux USA par année. Il peut y avoir plusieurs mises en application par état.

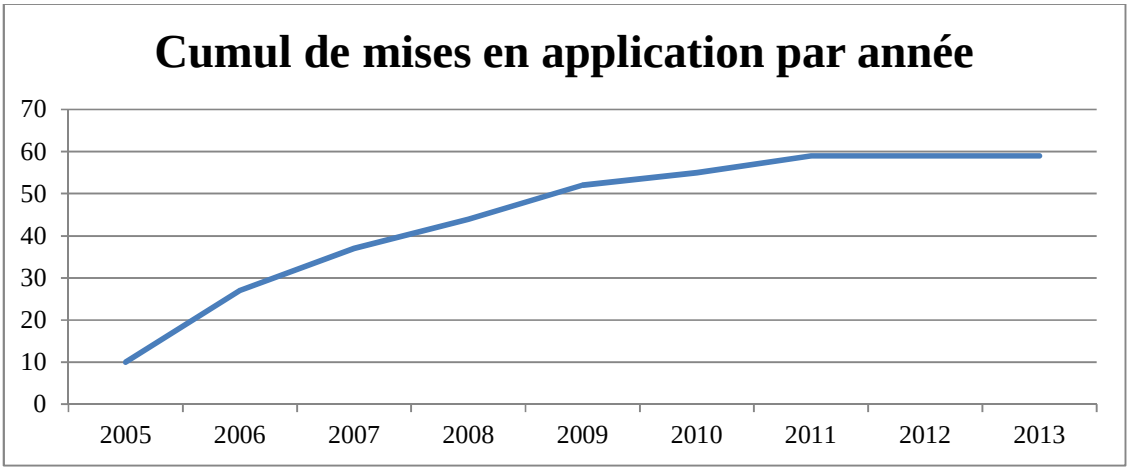


Figure 9 Evolution de la réglementation aux USA

On peut espérer une disponibilité de ce genre de données en Europe dans les années à venir du fait du règlement européen du 24 juin 2013 dit « data breach ».

3. Nombres d'entreprises aux USA

Les statistiques sur le nombre d'entreprises enregistrées aux USA par secteur d'activité sont accessibles sur le site https://www.census.gov/econ/esp/.

Les données concernent uniquement l'année 2011. Ces statistiques sont publiées tous les 5 ans, nous ne pouvons donc pas avoir d'informations plus précises. On considérera que ces nombres sont stables sur les années étudiées.

4. Institut Ponemon

L'institut Ponemon est un centre de recherche fondé par Larry Ponemon dédié à l'étude de la protection des données.

Ce sont des études qui visent à connaître le coût des violations de données en entreprise.

Ces études ont commencé en 2006 aux USA, puis ce sont progressivement étendues à de nombreux pays. A partir de 2012, les études fournissent les données du sondage par entreprise. Les données fournies sont les suivantes :

1. Nombre d'enregistrements de la violation
2. Coût de la détection et de la diffusion de l'information
3. Coût de la notification (obligation réglementaire)
4. Coût indirect (après l'agression)
5. Perte de chiffre d'affaires

On trouvera en annexe A un tableau qui récapitule les données majeures des différentes études au cours du temps.

Nous utiliserons pour les modèles statistiques les données détaillées fournies à partir de 2012, ce qui constituera un échantillon de **800 observations**.

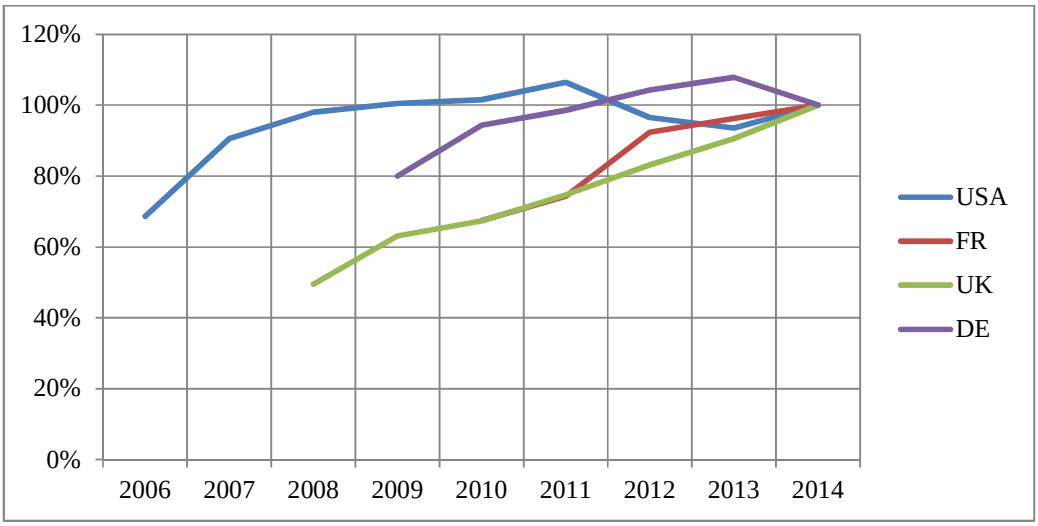


Figure 10 Coût moyen par enregistrement (comparé à 2014)

On peut constater sur tous les pays une inflation globale. Cependant, il semble que les USA observent une forme de stabilité des coûts par donnée volée. Cela peut s'expliquer par une meilleure maturité de la réglementation, alors que dans les pays européens la réglementation est encore en cours de construction.

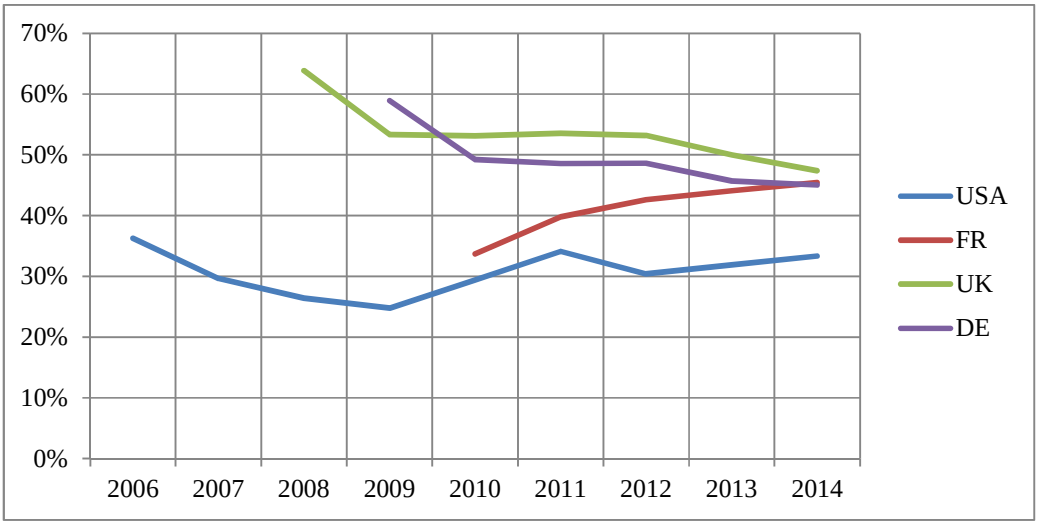


Figure 11 Proportion du coût direct

Remarquons ici que la répartition entre coût direct et indirect semble converger vers le même point pour les 3 pays européens par contre les USA se distinguent. Il faut donc en conclure que, d'un point de vue financier, il existe bien une différence entre les pays, qui vient à la fois d'une différence de culture et d'une différence de réglementation.

D'après ces études, les paramètres les plus influents sur la **fréquence de sinistre** sont :

1. Le type de données utilisées (en particulier les données bancaires et les données de santé)
2. Le secteur d'activité
3. L'existence d'une politique de gestion de la confidentialité et de la protection des données.
4. Le nombre de salariés
5. La présence d'un RSSI dans l'entreprise

Les paramètres les plus influents sur le **coût moyen par enregistrement** sont :

1. Le pays d'activité
2. Le type de données utilisées (en particulier les données bancaires et les données de santé)
3. Le secteur d'activité
4. L'existence d'une politique de gestion de la confidentialité et de la protection des données.
5. Le nombre de salariés
6. Le nombre de pays d'implantation
7. La présence d'un RSSI dans l'entreprise

Remarquons que l'étude **ne permet pas de connaître** les paramètres influençant **le volume de données impactées**.

On peut donc en déduire que :

- Le coût moyen par enregistrement est un indicateur qui a du sens.
- Il est intéressant de faire une étude en fonction de la taille de la violation, et le considérer à défaut d'avoir une donnée plus précise comme un proxy du coût.

Cependant il n'y a pas proportionnalité, donc l'utilisation de la taille de la violation comme proxy est intéressante mais insuffisante.

Notons que dans le dernier rapport de l'institut Ponemon la probabilité d'avoir un vol de données pour les entreprises localisées aux USA sur les 2 prochaines années est de 19%. Le rapport souligne une forte disparité de cette probabilité entre les industries.

5. Tarifs

Les données sont issues du site http://databreachinsurancequote.com/cyber-insurance/cyber-insurance-data-breach-insurance-premiums/

Nous y trouvons des exemples de tarifs qui sont résumés dans le tableau suivant.

Nom	Chiffre d'affaires	Prime	Limite
Doctor's Office	$ 700 000,00	$ 649,00	$ 500 000,00
Online Retailer	$ 500 000,00	$ 1 100,00	$ 1 000 000,00
Professional Consulting Services	$ 400 000,00	$ 1 200,00	$ 1 000 000,00
Psychologist's Office	$ 1 000 000,00	$ 1 600,00	$ 1 000 000,00
Doctor's Office	$ 1 700 000,00	$ 1 800,00	$ 1 000 000,00
Data Hosting Provider (startup)	$ 200 000,00	$ 2 750,00	$ 1 000 000,00
Healthcare IT Consultant	$ 150 000,00	$ 3 298,00	$ 1 000 000,00
e-Waste Company	$ 1 500 000,00	$ 3 564,00	$ 2 000 000,00
Clinical Data Analysis Research Software (startup)	$ 20 000,00	$ 4 900,00	$ 2 000 000,00
SaaS Provider	$ 3 000 000,00	$ 6 000,00	$ 2 000 000,00
Electronic Health Records (EHR) Provider	$ 5 000 000,00	$ 8 010,00	$ 1 000 000,00
Fast Food	$ 15 000 000,00	$ 9 000,00	$ 1 000 000,00
Healthcare SaaS Provider (startup)	$ 1 500 000,00	$ 30 420,00	$ 5 000 000,00
Hospital	$ 170 000 000,00	$ 42 000,00	$ 5 000 000,00
Data Storage Center	$ 15 000 000,00	$ 120 000,00	$ 20 000 000,00

Nous pouvons remarquer que la limite et le secteur d'activité ont plus d'influence sur le tarif que le chiffre d'affaires.

C. Analyse du commentaire des violations USA

Nous avons fait cette partie sur une extraction des données fin août qui comprend 4437 lignes. On trouve en moyenne 3,3 phrases par commentaire.

1. Les dates

A l'aide de techniques de fouille de texte, on peut trouver les dates présentes dans le commentaire. Cette démarche permet de se faire une idée sur la durée de résolution d'un sinistre.

Seules 1203 lignes ont des dates dans le commentaire, pour 2389 dates trouvées. Pour ces lignes, le délai moyen entre les dates extrêmes observées avoisine les 500 jours.

a) Durée de déroulement

Si on prend le délai entre la date de publication et la date maximale des commentaires, on trouve une moyenne de 110 jours. Cela souligne une durée pour résoudre un sinistre supérieur à 3 mois et qui peut s'étaler sur plusieurs années. Ce type d'information peut aider à **évaluer les PSAP** faute d'historique dans la base des sinistres d'un assureur. En effet pour ce type de sinistres émergents la méthode Chaine Ladder n'est pas adaptée.

b) Age des données

Si on prend le délai entre la date minimale des commentaires et la date de publication, on trouve une moyenne de 388 jours. Cela représente la durée entre un évènement antérieur à la déclaration de l'incident et l'incident. On remarque donc une période supérieure à un an, qui peut s'étaler sur plusieurs années.

Cela est souvent représentatif du délai entre la collecte par l'organisation et la perte des données. Parfois, cet écart de dates est représentatif du délai entre le vol des données et la détection de l'incident.

Dans les cas d'obligation d'information, ce délai peut souligner la difficulté à retrouver les personnes impactées. En effet si les données sont anciennes, les coordonnées des personnes peuvent avoir changées.

2. Les quantités

Nous avons cherché les quantités indiquées dans les commentaires. Leur écriture en nombre ou en lettre est reconnue.

Il y a 9488 quantités trouvées, nous n'avons pas pu trouver le nom quantifié pour 4560 d'entre elles. Notons que les dates sont interprétées comme des quantités mais ne quantifient pas de noms (nous avons déjà indiqué avoir trouvé 2389 dates).

Voici les noms les plus représentés :

Nom	Nb d'occurrences
email	24
september	25
april	26
week	26
august	27
hour	28
october	28
november	29
record	29
name	30
february	32
user	34
december	34
june	34

Nom	Nb d'occurrences
client	34
july	36
may	36
state	38
digit	39
day	41
man	41
march	43
member	44
january	46
credit	46
office	50
new	59
laptop	61

Nom	Nb d'occurrences
individual	61
computer	88
customer	88
theft	99
count	115
employee	125
student	127
month	147
patient	237
people	241
year	282
$	504

Remarquons que ce sont les montants en dollars qui sont le plus représentés. Ensuite, ce sont des périodes temporelles. Enfin, ce sont des catégories de personnes.

3. Les montants

Nous venons de voir que nous trouvons 504 montants dans les commentaires. Ces montants ne désignent pas forcement des coûts.

Le montant moyen est de 100 millions. L'écart type de $8{,}87.10^{8}$ souligne des montants allant de la dizaine aux milliards. Cela indique que les montants considérés sont loin d'être négligeables et peuvent représenter des enjeux importants pour les organisations touchées.

D. Fréquence

Notons que nous partons sur une approche classique d'indépendance entre fréquence et sévérité, ce qui permet d'étudier les deux problèmes séparément et de mieux exploiter l'ensemble des données disponibles.

1. Données USA

En divisant le nombre de violations par le nombre d'entreprises aux USA (en 2011), on obtient la probabilité qu'une violation se produise.

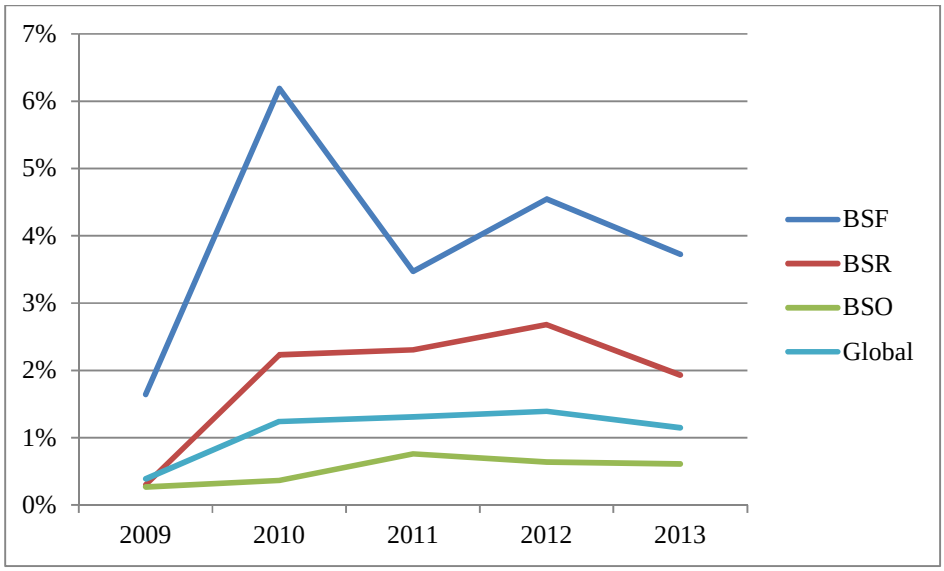


Figure 12 Taux d'entreprises touchées par une violation

On remarque une différence de fréquence entre les secteurs financiers, marchands et le reste. Par contre, on doit moyenner les fréquences sur plusieurs années pour limiter la variance car les graphiques ne sont pas constants et le volume de sinistres par années est insuffisant.

2. Reste du monde

Du fait de la mondialisation importante dans le secteur informatique on peut s'attendre à de grandes similitudes en termes de risque d'attaque entre les différentes régions du monde. D'autre part, nous disposons uniquement de données en provenance des USA. C'est pourquoi l'étude se focalisera sur les USA étant donné que les conclusions peuvent être généralisées au reste du monde.

E. Sévérité (en volume de données)

1. Institut Ponemon

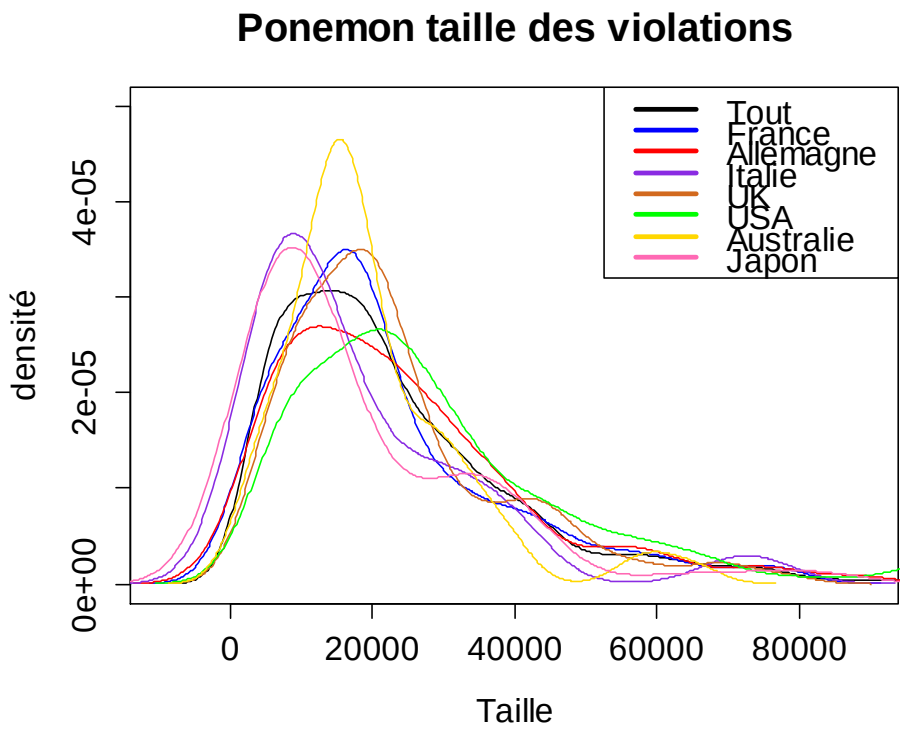


Figure 13 Institut Ponemon, densité de probabilité de la taille des violations

Remarquons ici que, malgré quelques différences entre les pays, il existe une similitude entre toutes les courbes. **On posera l'hypothèse que la sévérité de l'événement mesuré en volume de données impactées conserve la même loi sur tous les pays.** Cette hypothèse, bien que discutable, permet d'utiliser les données de violation des USA pour affiner les estimations de la prime pure dans le reste du monde.

Les densités de probabilité ont une forme proche de celle d'une loi log-normale. Pour cette raison et pour des raisons pratiques que nous verrons plus tard, nous allons nous concentrer sur l'étude du **logarithme des valeurs**.

Nous allons donc chercher le modèle le plus adapté parmi les lois normale, gamma et logistique. En effet, le logarithme peut être négatif ou positif et ces 3 lois comptent parmi les lois à support réel les plus simples d'utilisation.

Nous travaillerons donc sur un échantillon de **800 observations**.

Voici les résultats sous R avec la librairie fitdist et la fonction gofstat et ks.test pour les indicateurs:

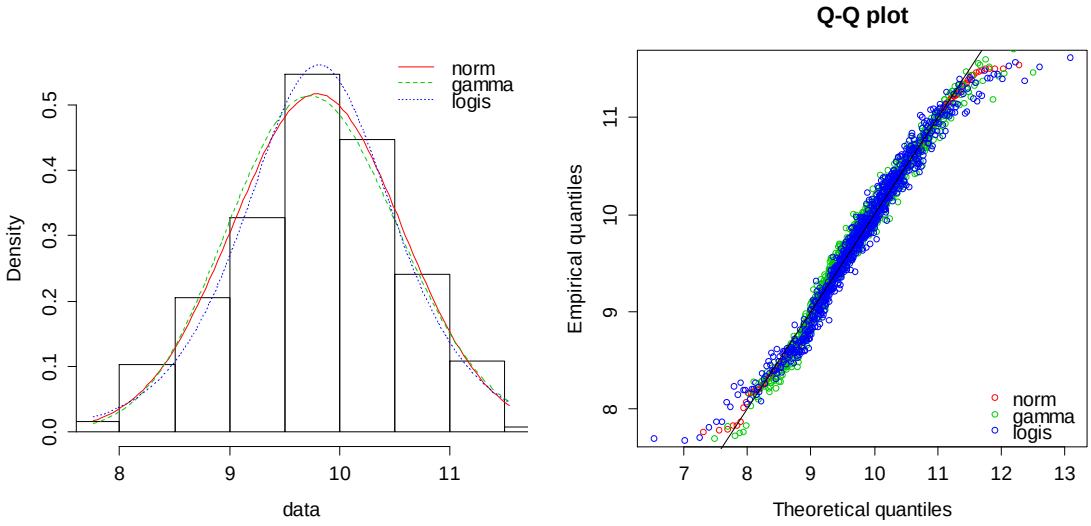


Figure 14 Adéquation des lois sur le logarithme des tailles

Kolmogorov-Smirnov	normale	gamma	logistique
D	0,03616174	0,04688828	0,03973111
p-value	0,2463	0,05935	0,1599

On note H_0 l'hypothèse que la distribution empirique et la distribution théorique suivent la même loi. On appelle le risque de **première espèce** le risque de rejeter H_0 l'hypothèse nulle alors qu'elle est vraie. La p-valeur est la probabilité d'avoir une distance plus grande sous l'hypothèse nulle. Au plus la p-valeur est petite, au plus on peut rejeter l'hypothèse nulle. On considère que **le seuil d'acceptation fort est de 5%** et **le seuil d'acceptation faible est de 1%**. Au-dessus du seuil on ne rejette pas l'hypothèse nulle.

Toutes les lois semblent acceptables au seuil de 5% mais la loi normale est acceptable bien au-delà de ce seuil. Comme nous le verrons plus tard, les données de Ponemon semblent filtrer les grandes et les petites valeurs. Nous allons donc privilégier la loi normale.

Les paramètres du modèle estimé par le maximum de vraisemblance sont:

	Estimé	Erreur type
Moyenne	9,7931064	0,02728279
Ecart-type	0,7716738	0,0192917

2. Base de données des violations aux USA

Remarquons qu'il y a 4338 observations sur les 10 années 2005 à 2014, mais **seulement 2901 sont exploitables ici**, car les autres n'ont pas le nombre d'enregistrements concernés par la violation renseignée.

Il est donc envisageable que les données comportent un biais.

a) Densité de probabilité en coupant la queue

La queue de la distribution est ici très épaisse. J'ai donc tronqué la distribution pour ne pas rendre le graphique illisible.

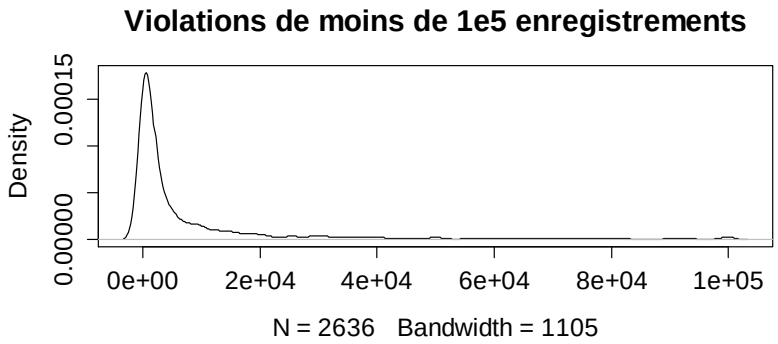


Figure 15 Distribution des tailles de violation USA

b) Densité de probabilité du logarithme

Cette fois, nous passons au logarithme du nombre d'enregistrements impacté par la violation.

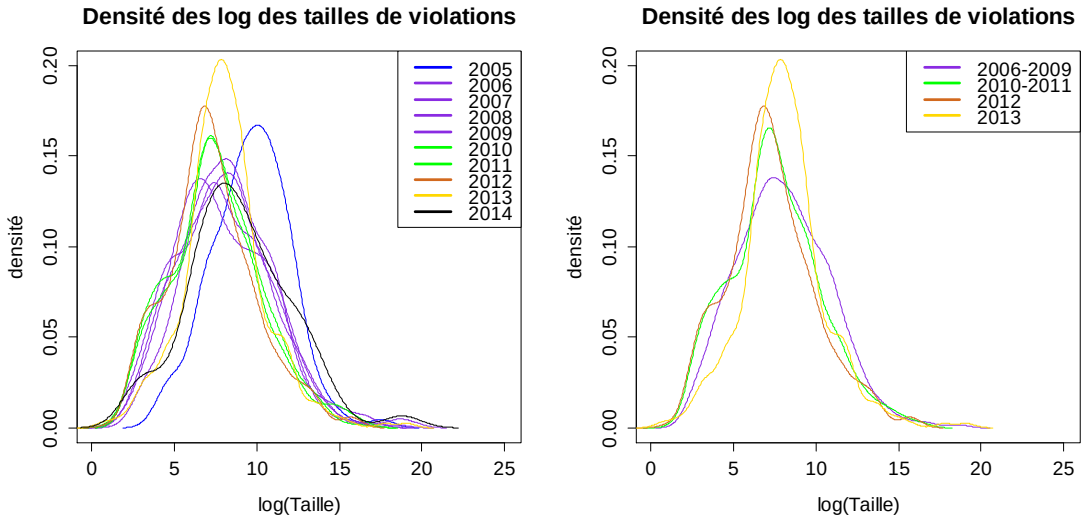


Figure 16 Distribution des logarithmes des tailles de violation USA

Voici un résumé des paramètres principaux des courbes :

Années	Nombres d'observations	Moyenne
2006-2009	1168	10,4=log(32248)
2010-2011	808	8,88=log(7211)
2012	429	8,43=log(4582)
2013	326	9,9=log(20000)

Les années 2005 et 2014 ne rassemblent respectivement que 116 et 54 observations. Elles sont donc considérées comme non représentatives.

On observe un décalage de la courbe vers de plus petites valeurs. Pour améliorer la sécurité d'une organisation, le cloisonnement des informations est largement utilisé. Cela permet de diminuer l'ampleur d'un éventuel incident. La mise en place d'outils de détection d'intrusion permet aussi de limiter la quantité de données impactées par un incident en réduisant sa durée. Ce décalage de la courbe vers la gauche peut s'expliquer par de meilleures politiques de sécurité.

Malgré une légère modification de la forme de la courbe, une grande similitude entre les représentations graphiques demeure. Cela confirme les informations données au paragraphe III.A.4 . On peut supposer l'indépendance entre fréquence et sévérité.

Nous allons donc chercher le modèle le plus adapté parmi la loi normale, la loi gamma et la loi logistique.

Voici le graphique sur toutes les années disponibles sous R:

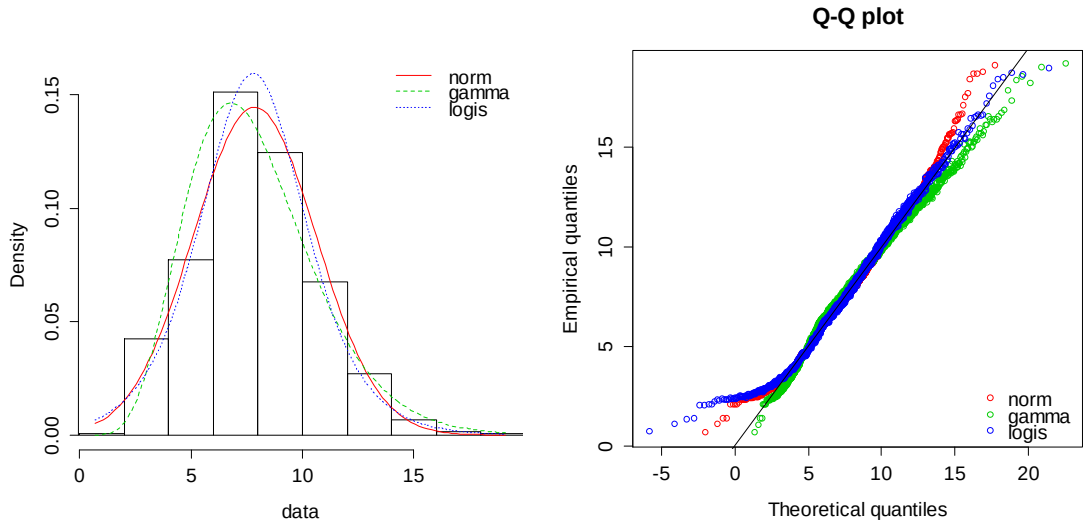


Figure 17 Adéquation des lois du logarithme des tailles

Kolmogorov-Smirnov	normale	gamma	logistique
D	0,03179978	0,05765267	0,02638612
p-value	0,005662	8,428e-9	0,03521

On remarque ici que la loi de répartition de la taille des violations se rapproche d'une log-normale mais avec une queue plus épaisse. Aucune des p-valeurs ne passe au seuil de 5%. Par contre, la loi logistique est la seule à passer le seuil acceptable des 1%. La loi logistique semble être la plus proche du logarithme de la taille des violations, en particulier sur les grandes valeurs. Nous retiendrons donc ce modèle car les autres sont rejetés.

Les paramètres du modèle estimé par le maximum de vraisemblance sont:

	Estimé	Erreur type
Moyenne (μ)	7,777499	0,05076807
Paramètre d'échelle (s)	1,567027	0,02416376

$$P(\log(X) \geq \log(x)) = \frac{1}{1 + \exp\left(\frac{\log(x)-\mu}{s}\right)}$$

Pour z grand on a donc :

$$P(X \geq x) \cong \left(\frac{e^{\mu}}{x}\right)^{1/s}$$

Où : $1/s = 0.64 \pm 0.01$

On remarque donc que l'épaisseur de queue est conforme aux résultats du paragraphe III.A.3.

Sur les graphiques suivants, nous avons fait le même exercice sur des parties filtrées du jeu de données en fonction du secteur d'activité de l'organisation et du type d'attaque. Cela permet d'observer que malgré des similitudes, toutes les fonctions de répartition ne sont pas identiques, ces indicateurs apportent donc une information sur le risque porté.

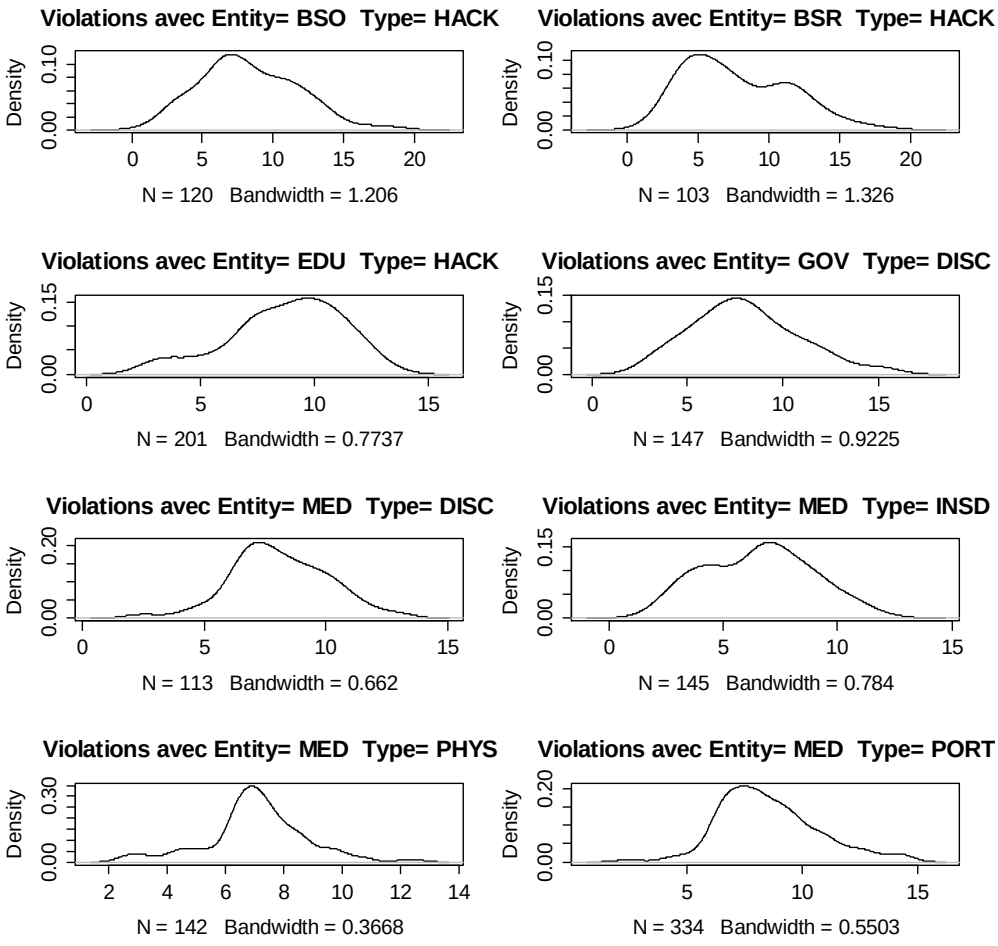


Figure 18 Distribution du logarithme des tailles par critères

3. Comparaison

On remarque que les 2 jeux de données sont différents alors qu'ils devraient représenter les mêmes fonctions de répartition.

Les études de l'institut Ponemon excluent les incidents touchant à plus de 100 000 enregistrements. Cette exclusion est justifiée par la volonté d'étudier les incidents les plus fréquents pour les entreprises. L'étude doit exclure les plus petits incidents pour des raisons de ressources et se concentrer sur les cas les plus importants pour les organisations touchées. Les données de l'institut qui sont un échantillon des sinistres, semblent résulter d'un filtre qui coupe les valeurs faibles et les valeurs fortes. Cependant, dans la base de sinistre de l'institut, on peut trouver des cas avec plus de 100 000 enregistrements. Cela nous amène à considérer **une troncature aléatoire**.

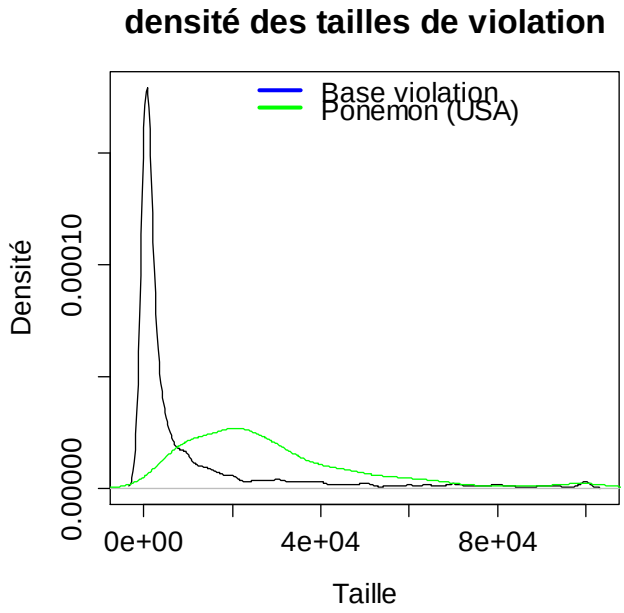


Figure 19 Comparaison des densités

On va donc supposer que la loi de l'institut Ponemon est la loi tronquée des incidents US, par 2 valeurs aléatoires T et U indépendantes.

On note X la variable aléatoire du nombre de données volées et dont la fonction de répartition est F. Après troncature, la fonction de répartition est noté H.

$$f(x) = \frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}$$

$$F(x) = \frac{1}{1 + \exp\left(-\frac{x-\mu}{s}\right)}$$

$$1 - F(x) = \frac{1}{1 + \exp\left(\frac{x-\mu}{s}\right)}$$

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-1/2 \left(\frac{x-m}{\sigma}\right)^2\right)$$

Les paramètres sont :

μ	s	m	σ
7,777499	1,567027	9,7931064	0,7716738

On cherche la loi h_1 après troncature à gauche par une variable T de loi g.

Si on fixe T=t alors la loi tronqué est $f(x|x \geq t) = \frac{f(x)}{1-F(t)}$. Donc :

$$h_1(x) = \int_{-\infty}^{+\infty} f(x|x \geq t) g(t) dt = \int_{-\infty}^{x} \frac{f(x)}{1 - F(t)} g(t) dt$$

On cherche la loi h après troncature à gauche par une variable U de loi q. De même :

$$h(x) = \int_{-\infty}^{+\infty} h_1(x|x \leq u) q(u) du = \int_{x}^{+\infty} \frac{h_1(x)}{H_1(u)} q(u) du$$

$$\frac{h(x)}{f(x)} = \int_{-\infty}^{x} \frac{g(t)}{1 - F(t)} dt \int_{x}^{+\infty} \frac{q(u)}{H_1(u)} du$$

On va poser les hypothèses :

- $\exists b, \forall t \leq b, q(t) = 0$
- $\exists a, \forall t \geq a, g(t) = 0$

Pour $x \leq a$ on a une constante K_1 qui vérifie :

$$K_1 \frac{h(x)}{f(x)} = \int_{-\infty}^{x} \left(1 + \exp\left(\frac{t-\mu}{s}\right)\right) g(t) dt$$

$$\frac{1}{K_1} = \int_{b}^{+\infty} \frac{q(u)}{H_1(u)} du$$

En dérivant l'équation on trouve donc :

$$\boldsymbol{g(t) = \frac{-k(t)}{1 + \exp\left(\frac{t-\mu}{s}\right)} K_1 \frac{1}{\sigma\sqrt{2\pi}} exp\left(-1/2\left(\frac{t-m}{\sigma}\right)^2\right)}$$

Avec :

$$\boldsymbol{k(t) = \frac{t-m}{\sigma^2} s\left(1 + \exp\left(\frac{t-\mu}{s}\right)\right)\left(1 + \exp\left(-\frac{t-\mu}{s}\right)\right) + \exp\left(\frac{t-\mu}{s}\right) - \exp\left(-\frac{t-\mu}{s}\right)}$$

Pour $x \geq b$ on a une constante K_2 qui vérifie :

$$K_2 \frac{h(x)}{f(x)} = \int_x^{+\infty} \frac{q(u)}{H_1(u)} du$$

$$\frac{1}{K_2} = \int_{-\infty}^{a} \frac{g(t)}{1 - F(t)} dt$$

De plus :

$$h_1(x) = f(x) \int_{-\infty}^{a} \frac{1}{1 - F(t)} g(t) dt$$

$$h_1(x) = \frac{f(x)}{K_2}$$

$$1 - H_1(x) = \frac{1 - F(x)}{K_2}$$

$$H_1(x) = \frac{K_2 - 1 + F(x)}{K_2}$$

En dérivant l'équation on trouve donc :

$$\boldsymbol{q(t) = k(t)\left(K_2 - 1 + F(x)\right) \frac{1}{\sigma\sqrt{2\pi}} exp\left(-1/2\left(\frac{t-m}{\sigma}\right)^2\right)}$$

Les deux expressions g et q sont nulles lorsque :

$$k(t) = 0$$

Lorsque $t \leq \mu$, k est négative et lorsque $t \geq m$, k est positive. Par continuité, il existe une valeur $a \in [\mu, m]$ de t qui annule g et tel que $\forall x \leq a, g(x) \geq 0$ et $\forall x \geq a, g(x) = 0$. On observe graphiquement qu'on ne peut trouver qu'une valeur possible pour a, qui est proche de 9,5.

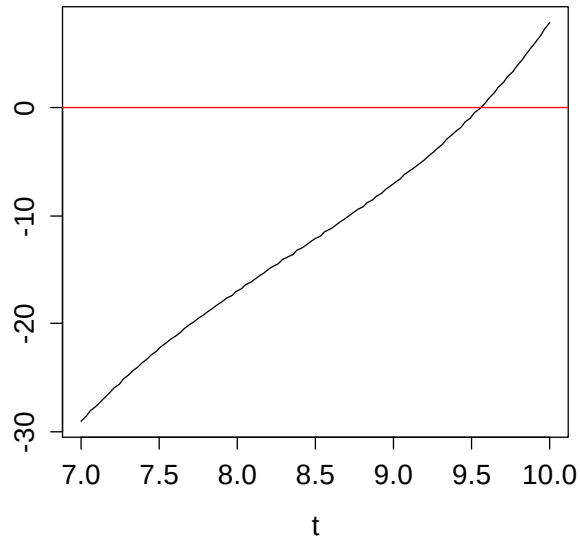


Figure 20 Recherche de a

De plus $K_2 > 0$ donc $K_2 - 1 + F(x) \geq 0$ pour x assez grand. Ce qui justifie l'existence de b avec $b \geq a$ qui annule q et tel que $\forall x \geq b, q(x) \geq 0$ et $\forall x \leq b, q(x) = 0$.

Du fait de la présence de $\exp\left(-1/2\left(\frac{x-m}{\sigma}\right)^2\right)$ dans les expressions de g et q, on peut dire que les 2 sont intégrables sur $\mathbb{R}$.

On a : $\int_{-\infty}^{+\infty} h_1(x)dt = 1$

$$\int_{-\infty}^{+\infty} h_1(x)dx = \int_{-\infty}^{+\infty} f(x) \int_{-\infty}^{x} \frac{g(t)}{1-F(t)} dt\, dx$$

$$\int_{-\infty}^{+\infty} h_1(x)dx = \left[F(x) \int_{-\infty}^{x} \frac{g(t)}{1-F(t)} dt\right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} F(x) \frac{g(x)}{1-F(x)} dx$$

$$\int_{-\infty}^{+\infty} h_1(x)dx = \int_{-\infty}^{+\infty} g(x)dx$$

De même : $\int_{-\infty}^{+\infty} h(x)dt = 1$

$$\int_{-\infty}^{+\infty} h(x)dx = \int_{-\infty}^{+\infty} h_1(x) \int_{x}^{+\infty} \frac{q(t)}{H_1(t)} dt\, dx$$

$$\int_{-\infty}^{+\infty} h(x)dx = \left[H_1(x) \int_{x}^{+\infty} \frac{q(t)}{H_1(t)} dt\right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} H_1(x) \frac{q(x)}{H_1(x)} dx$$

$$\int_{-\infty}^{+\infty} h(x)dx = \int_{-\infty}^{+\infty} q(x)dx$$

Les solutions trouvées sont donc bien conformes à toutes les hypothèses.

F. Sévérité (en coût)

1. Institut Ponemon

Voici les graphiques de la fonction de répartition du coût des violations de données évaluée dans les études l'institut Ponemon. Nous ne présentons ici que le coût total, on retrouvera en annexe B les graphiques pour tous les coûts séparés.

Tous les coûts sont exprimés en dollars US, nous utilisons le tableau de change suivant, qui est la moyenne des changes sur les années 2012 à 2014.

USA	AU	BR	AR	FR	IT	DE	IN	JA	UK
$US	$AU	Real	Riyal	Euro	Euro	Euro	Roupie	Yen	Livre
1	1.00740	0.52179	0.27225	1.33577	1.33577	1.33577	0.01901	0.01170	1.58591

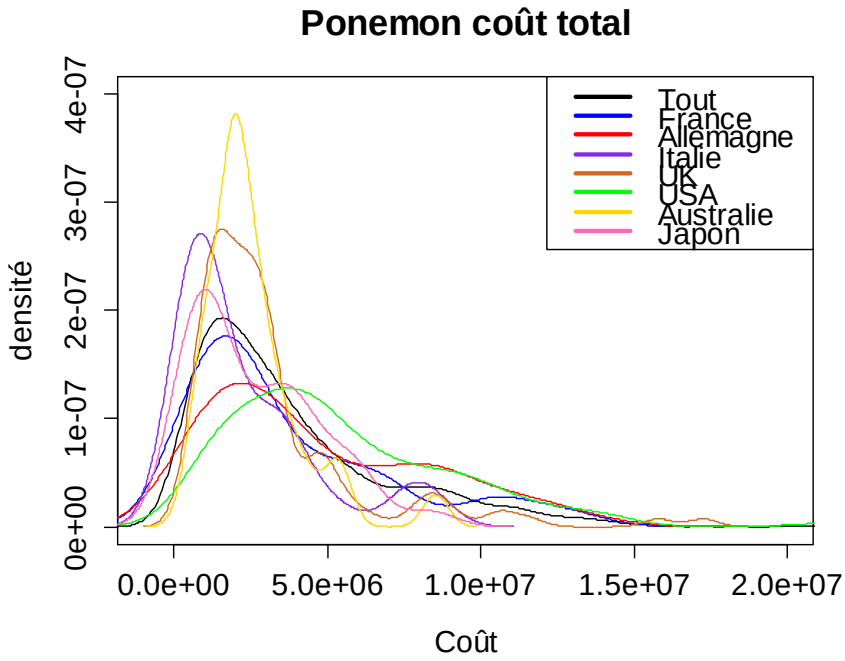


Figure 21 Densité du coût par pays

On peut remarquer des similitudes entre les courbes. Les courbes des différents pays occupent des places pour la plupart similaires à ce qu'on trouve sur les volumes de données.

La loi recherchée ici semble être une loi log-normale. Pour des raisons pratiques que nous verrons plus tard, nous allons nous concentrer sur l'étude du **logarithme des valeurs**.

Nous allons donc chercher le modèle le plus adapté parmi la loi normale, la loi gamma et la loi logistique.

Voici les résultats sous R :

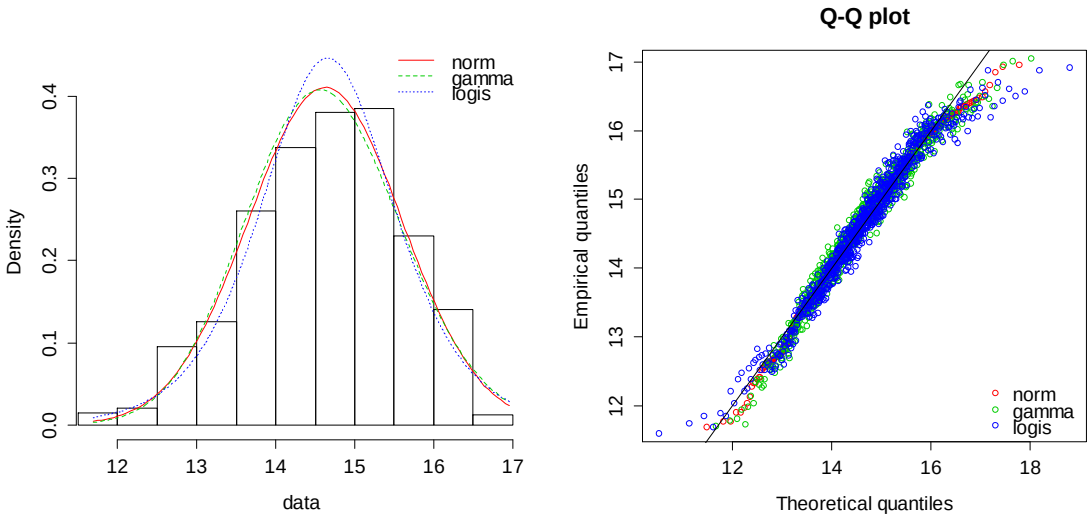


Figure 22 Adéquation des lois du logarithme des coûts

Kolmogorov-Smirnov	normale	gamma	logistique
D	0,03687764	0,04524281	0,03262722
p-valeur	0,2267	0,07561	0,362

Toutes les lois semblent convenir au seuil de 5% des p-valeur, mais les lois normale et logistique se démarquent. Les données de Ponemon semblent résulter d'un filtre sur les grandes et les petites valeurs. Nous allons donc privilégier la **loi normale** qui permet de respecter les hypothèses de la régression linéaire. Il peut néanmoins être utile de tester la loi logistique dans le modèle basé sur les copules.

Les paramètres du modèle avec la loi normale estimés par le maximum de vraisemblance sont:

	Estimé	Erreur type
Moyenne	14,6274154	0,0343503
Ecart-type	0,9715731	0,02428921

Les paramètres du modèle avec la loi logistique estimés par le maximum de vraisemblance sont:

	Estimé	Erreur type
Moyenne	14,6578265	0,03473944
Echelle	0,5602921	0.01635990

G. Relation entre sévérités

1. Etude préliminaire

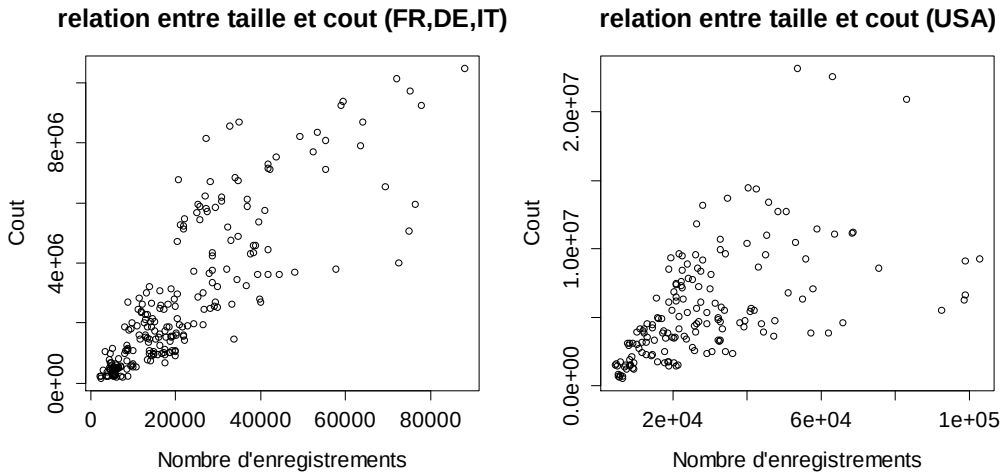


Figure 23 Relation entre taille et coût dans la monnaie d'origine

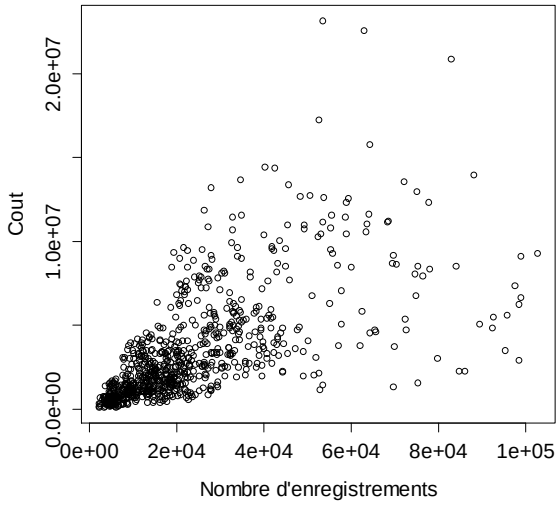


Figure 24 Relation entre taille et coût pour le monde en dollars

Remarquons ici la forme conique qui montre que les lois n'ont pas une relation homoscédastique. Les formes des courbes observées dans les paragraphes précédents sont proches de lois log-normales. De plus l'agglomération de points dans les petites valeurs incite à faire une étude du log des valeurs. C'est pourquoi nous allons maintenant mener l'étude sur les lois du logarithme des valeurs.

2. Modèle linéaire

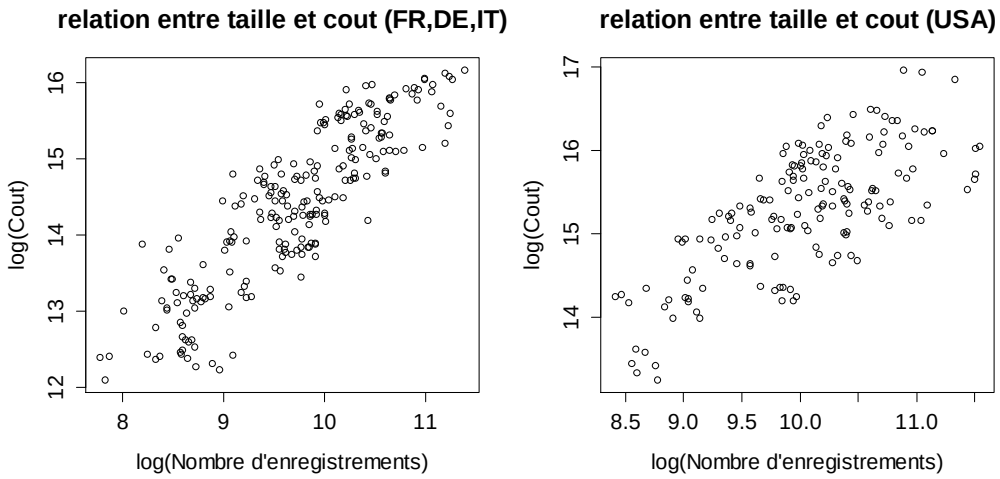


Figure 25 Relation entre taille et coût dans la monnaie d'origine

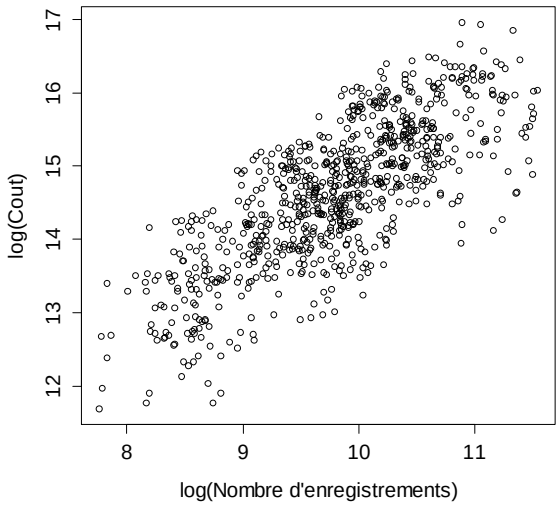


Figure 26 Relation entre taille et coût pour le monde en dollars

On observe ici une distribution des points propices à l'utilisation d'une régression linéaire.

On note y le coût, x le nombre d'enregistrements et ϵ le bruit.

Pour la zone euro :

$$\log(y) = 1{,}143 * \log(x) + 3{,}294 + \epsilon$$
$$\sigma(\epsilon) = 0{,}495$$

Pour les USA :

$$\log(y) = 0{,}772 * \log(x) + 7{,}546 + \epsilon$$
$$\sigma(\epsilon) = 0{,}516$$

Pour le monde :

$$\log(y) = 0{,}969 * \log(x) + 5{,}133 + \epsilon$$
$$\sigma(\epsilon) = 0{,}621$$

On a :

- $R^2 = 0{,}5929$
- P-value inférieur à 2.10^{-16} pour la pente et l'ordonnée à l'origine
- Erreur standard pour l'ordonnée à l'origine : 0,27936
- Erreur standard pour la pente : 0,02844

Donc le nombre de données volées est un paramètre qui n'explique pas la totalité du coût, par contre le modèle est pertinent.

Si on le fait par pays avec :

$$\log(y) = a * \log(x) + b + \epsilon$$

Pays	a	p-value	Err. Std.	b	p-value	Err. Std.	$\sigma(\epsilon)$	R^2	Nb Cas
IN	0,73	<2e-16	0,06	6,69	<2e-16	0,62	0,47	0,64	77
AU	0,76	<2e-16	0,06	7,2	<2e-16	0,59	0,32	0,72	65
USA	0,77	<2e-16	0,06	7,55	<2e-16	0,57	0,52	0,53	164
AR	0,8	5,02E-05	0,16	6,69	3,86E-04	1,6	0,55	0,53	24
UK	0,93	<2e-16	0,05	5,54	<2e-16	0,52	0,38	0,74	114
JA	1,01	<2e-16	0,08	4,87	4,59E-09	0,72	0,52	0,73	67
BR	1,05	<2e-16	0,08	3,68	6,70E-06	0,75	0,46	0,76	63
DE	1,05	<2e-16	0,06	4,67	9,28E-11	0,63	0,46	0,76	87
IT	1,12	<2e-16	0,07	3,52	2,88E-06	0,68	0,46	0,8	63
FR	1,14	<2e-16	0,07	3,62	7,07E-07	0,67	0,46	0,79	76

Pour les USA, l'Inde et la région Arabe le R^2 est petit, il faudrait peut-être un maillage plus précis qui soit adapté aux disparités de règlementations et de culture entre les différents états ou les différentes régions de ces pays. Vu la variabilité des paramètres on peut penser que le pays est un critère important. De plus les p-values sont petites donc on peut rejeter l'hypothèse d'indépendance entre taille et coût.

Remarquons une relation entre a et b :

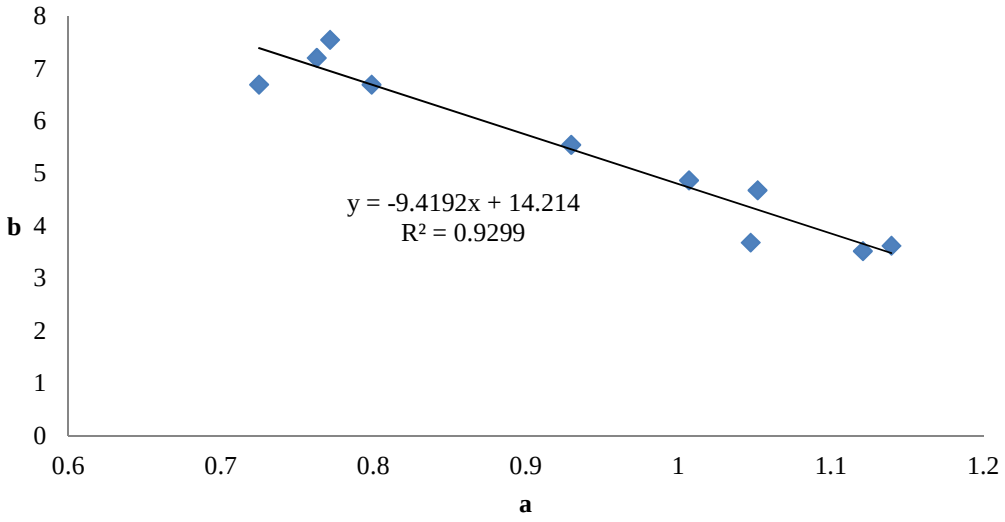


Figure 27 Relation entre les paramètres

Nous utiliserons cette relation **pour déterminer b pour un a donné** dans la suite du mémoire.

Ce qui veut dire que les courbes de chaque pays se croisent en $\log(x) = 9{,}42$ qui est proche de $E[\log(X)] = m = 9{,}79$ et $\log(y) = 14{,}2$ qui est proche de $E[\log(Y)] = 14{,}6$.

Les droites de régression passent par le barycentre. Donc cela traduit que les moyennes des logarithmes sont proches entre les pays. Par contre, les dispersions autour de la moyenne varient.

Par indépendance entre X et E, la troncature est linéaire. On peut donc supposer que dans l'univers complet on conserve la loi :

$$\log(Y) = a * \log(X) + b + E$$

Appliquons le modèle USA aux gros sinistres étudiés en première partie :

Nom	**Taille**	**Coût total**	**Montant prédit à 99%**
TJX	94 millions	250 m$	580m$-12M$
Heartland	130 millions	140 m$	730m$-16M$
Sony	77 millions	171 m$	490m$-11M$
Target	130 millions	61 m$	730m$-16M$

m = million et M = milliard

On constate que **l'extrapolation linéaire ne permet pas de prédire les coûts des plus gros sinistres,** malgré une marge d'erreur très grande.

3. Copule

On conduit ici l'étude de la copule sous les hypothèses présentées précédemment. Les deux variables de sévérité des études de l'institut Ponemon sont modélisées avec une loi log-normale pour le nombre de données impactées et log-normale ou log-logistique dans le cas du coût.

Une copule en dimension 2 est une fonction C définie sur $[0,1]^2$ et à valeurs dans $[0,1]$ vérifiant :

- $C(1,v) = v$ et $C(u,1) = u$
- $C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0$ avec $u_2 \geq u_1$ et $v_2 \geq v_1$

Théorème de Sklar : Si F est une fonction de distribution de dimension 2 dont les lois marginales F_1et F_2 sont continues, alors il existe une copule unique telle que :

$$F(u,v) = C\big(F_1(u), F_2(v)\big)$$

Remarquons que les copules sont invariantes par les transformations avec des lois 2 fois dérivables et strictement croissantes. Nous allons donc étudier le **logarithme des 2 variables qui sont donc modélisées par des lois normales ou logistiques**.

Nous aurons pour objectif de conserver un modèle simple et nous nous concentrerons sur les solutions existantes dans R. Nous allons donc étudier les copules archimédiennes présentes dans le *package copula*, donc les copules de Clayton, Frank, Gumbel, Joe et Ali-Mikhail-Haq.

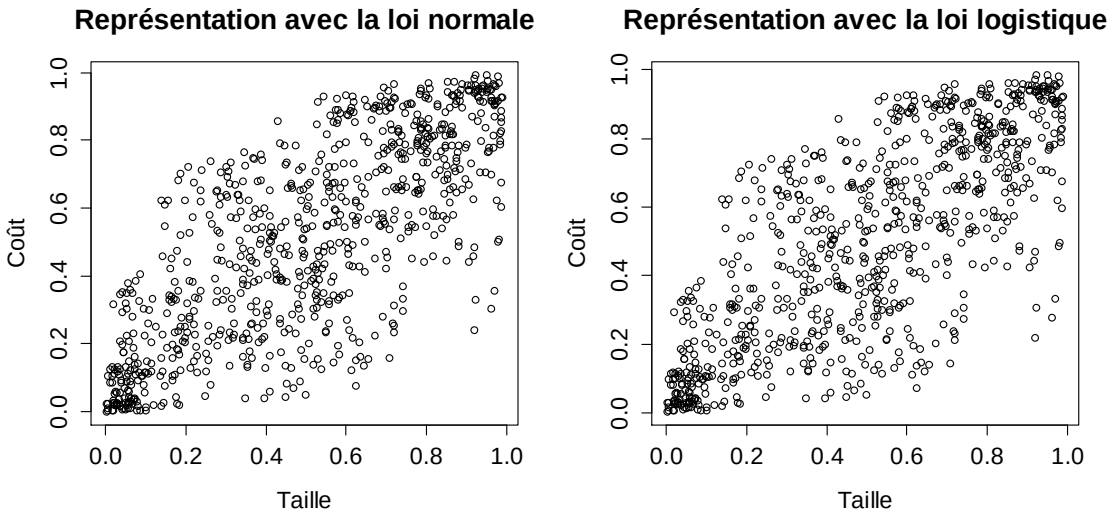


Figure 28 Représentation graphique de la copule

Graphiquement, on observe qu'une copule de Gumbel ou de Frank est acceptable comme modèle. Les copules de Clayton semblent inadaptées car on observe un rassemblement de points pour les grandes valeurs (en haut à droite) proche de ce qu'on trouve pour les petites valeurs (en bas à gauche). Enfin, les copules de Joe et Ali-Mikhail-Haq ne peuvent pas être calibrées sur les données, on les exclut donc.

On définit ces 2 copules archimédiennes ainsi :

$$C(u,v) = \varphi^{-1}\big(\varphi(u) + \varphi(v)\big)$$

Pour la copule de Gumbel :

$$\varphi(u) = (-\log(u))^{\theta}$$

$$C(u,v) = \exp\left(-\big((-\log(u))^{\theta} + (-\log(v))^{\theta}\big)^{1/\theta}\right)$$

Pour la copule de Frank :

$$\varphi(u) = -\log\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}$$

$$C(u,v) = -\frac{1}{\theta}\log\left(1 + \frac{\big(e^{-\theta u} - 1\big)\big(e^{-\theta v} - 1\big)}{e^{-\theta} - 1}\right)$$

Pour tester l'adéquation des modèles nous allons utiliser la méthode décrite dans l'article de Genest et Rivest (GENEST, et al., 1993). Le test fonctionne ainsi :

On définit W = F(X, Y), $K(w) = P(W \leq w)$ et $\lambda(v) = \frac{\varphi(v)}{\varphi\prime(v)}$.

On a la relation : $K(w) = w - \lambda(w)$

De plus on peut définir un estimateur empirique de K ainsi :

$$\widehat{K}(v) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{F}(X_i,Y_i)\leq v}$$

Où :

$$\hat{F}(x,y) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{X_j\leq x, Y_j\leq y}$$

On compare alors les courbes de λ et $\hat{\lambda}$.

Nous en déterminons les paramètres à l'aide de R avec une loi de coût normale:

	Gumbel	Frank
Log-vraisemblance négative	325,3884	338,0587
Paramètre	2,126281	6,913284
Erreur type	0,06587858	0,3714838

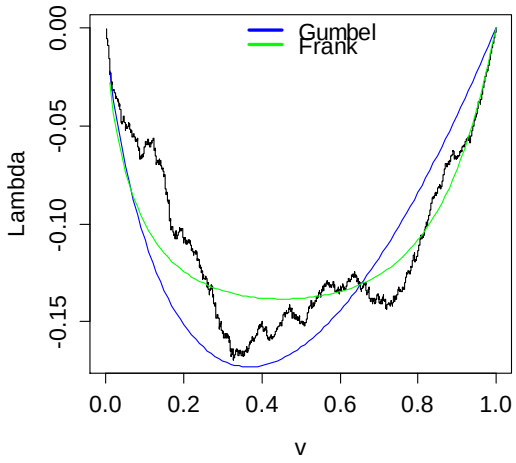


Figure 29 Adéquation des copules

Nous en déterminons les paramètres à l'aide de R avec une loi de coût logistique:

	Gumbel	Frank
Log-vraisemblance négative	327,3447	333,5393
Paramètre	2,145512	6,795609
Erreur type	0,06706458	0,3692971

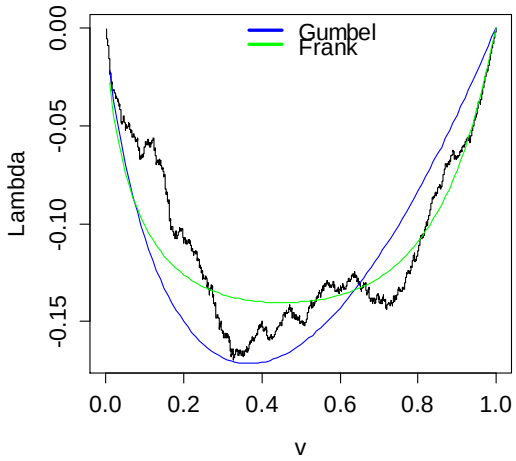


Figure 30 Adéquation des copules

Dans les deux cas la copule de Frank semble donner de meilleurs résultats sur le graphique. Cependant, on s'intéresse en priorité aux grandes valeurs de plus la copule de Gumbel présente une vraisemblance plus grande, nous allons donc privilégier la **copule de Gumbel**. Nous pouvons déjà remarquer qu'aucun des modèles ne donne une représentation graphique de même aspect que l'échantillon même si certaines ressemblances existent. Il faut donc prévoir un chargement de sécurité pour couvrir l'erreur de modèle.

On note $F(x,y)$ la fonction de répartition dans l'univers complet et $H(x,y)$ dans l'univers tronqué. On conserve les notations du paragraphe III.E.3.

$$F_X(x) = \frac{1}{1+\exp\left(-\frac{x-\mu}{s}\right)}$$

$$h_Y(y) = \frac{1}{\omega\sqrt{2\pi}}\exp\left(-1/2\left(\frac{y-\tau}{\omega}\right)^2\right)$$

$$h_X(x) = \frac{1}{\sigma\sqrt{2\pi}}exp\left(-1/2\left(\frac{x-m}{\sigma}\right)^2\right)$$

$$H(x,y) = \exp\left(-\left(\left(-\log\left(H_X(x)\right)\right)^\theta + \left(-\log(H_Y(y))\right)^\theta\right)^{1/\theta}\right)$$

Loi après troncature à gauche :

$$h_1(x,y) = \int_{-\infty}^{+\infty} f(x,y|x \geq t)g(t)dt = \int_{-\infty}^{x} \frac{f(x,y)}{1-F_X(t)}g(t)dt$$

Loi après troncature à droite :

$$h(x,y) = \int_{x}^{+\infty} h_1(x,y|x \leq u)q(u)du = \int_{x}^{+\infty} \frac{h_1(x,y)}{H_{1X}(u)}q(u)du$$

Où :

$$H_{1X}(u) = \int_{-\infty}^{u}\int_{-\infty}^{+\infty} h_1(x,y)dy\,dx$$

$$H_{1X}(u) = \int_{-\infty}^{u}\int_{-\infty}^{x} \frac{1}{1-F_X(t)}g(t)dt\int_{-\infty}^{+\infty} f(x,y)dy\,dx$$

$$H_{1X}(u) = \int_{-\infty}^{u}\int_{-\infty}^{x} \frac{F_X(x)}{1-F_X(t)}g(t)dt\,dx$$

Enfin :

$$H(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h(t,l)dl\,dt$$

$$h(x,y) = f(x,y)\int_{-\infty}^{x}\frac{1}{1-F_X(t)}g(t)dt\int_{x}^{+\infty}\frac{1}{H_{1X}(u)}q(u)du$$

En notant :

$$\delta = \frac{1}{\theta}$$

$$Z(x,y) = \log\left(-\log\left(H_X(x)\right)\right)^{\theta} + \left(-\log(H_Y(y))\right)^{\theta}$$

$$H(x,y) = \exp\left(-\left(Z(x,y)\right)^{\delta}\right)$$

$$z_X(x) = \frac{\partial Z(x,y)}{\partial x} = \theta\frac{-h_X(x)}{H_X(x)}\left(-\log(H_X(x))\right)^{\theta-1}$$

$$z_Y(y) = \frac{\partial Z(x,y)}{\partial y} = \theta\frac{-h_Y(y)}{H_Y(y)}\left(-\log(H_Y(y))\right)^{\theta-1}$$

$$h(x,y) = \frac{\partial^2 H(x,y)}{\partial x\partial y} = z_X z_Y \delta Z^{\delta-2}\left(\delta Z^{\delta} - \delta + 1\right)\exp\left(-Z^{\delta}\right)$$

On a trouvé une expression de $f(x,y)$. On remarque que pour x fixé f et h sont proportionnelles.

Dans le cas où le coût filtré suit une loi normale :

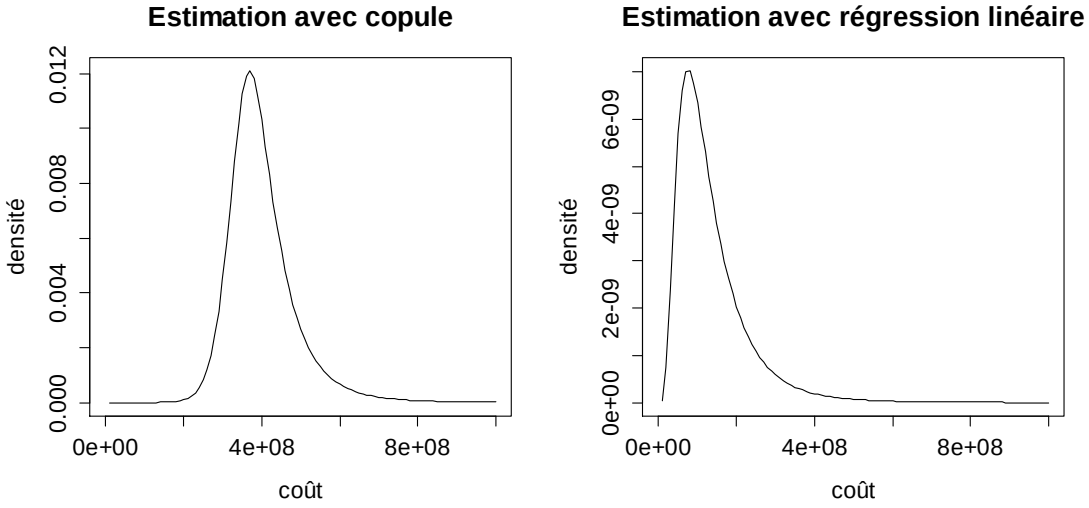


Figure 31 Estimation de la densité du coût pour un volume de données de 10^6

Dans le cas où le coût filtré suit une loi logistique :

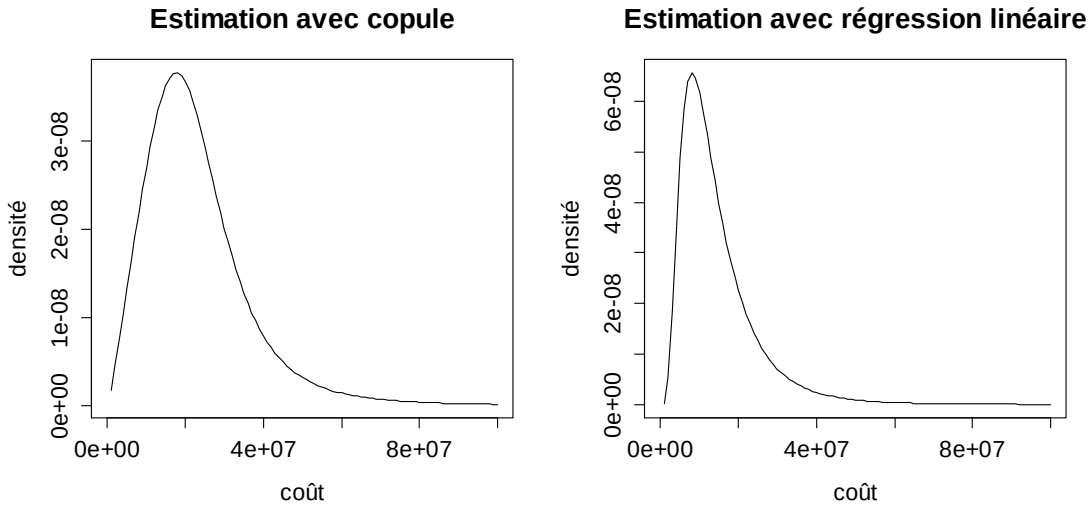


Figure 32 Estimation de la densité du coût pour un volume de données de 10^5

L'extrapolation à l'aide des copules donne une surestimation du coût des gros sinistres plus importante qu'avec la régression linéaire. On ne va donc pas les utiliser pour faire l'évaluation tarifaire. Cependant, il faut noter que l'étude théorique permet d'envisager des développements futurs. En effet, la base de données Ponemon ne comporte pas de gros sinistres. Il reste possible qu'avec une source de données différente, les copules soient une voie d'exploration utile.

H. Tarification : Application sur le calcul de la prime pure

Nous utilisons le modèle linéaire de la partie III.G.2 qui s'est avéré être aussi pertinent que le modèle basé sur des copules tout en étant plus simple.

$$\log(Y) = a * \log(X) + b + E$$

Par linéarité de la troncature, cette **écriture de Y est une solution** dans l'espace non tronqué qui vérifie la relation dans l'espace tronqué. On a démontré l'injectivité de la troncature dans les cas intégrables, donc c'est la seule solution.

On va donc utiliser la distribution déterminée sur la base de données des violations aux USA, qui comporte donc les sinistres extrêmes. Puis appliquer la relation entre la taille de la violation et le coût déterminée sur les données de l'Institut Ponemon. Ce qui nous permettra d'avoir la distribution des coûts même sur les sinistres extrêmes.

$$P(Y \leq u) = \int_{-\infty}^{+\infty} P(a * \log(X) + b \leq \log(u) - \varepsilon) f_E(\varepsilon) d\varepsilon$$

Avec :

$$f_E(\varepsilon) = \frac{1}{\sigma_E \sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2{\sigma_E}^2}}$$

Donc :

$$P(Y \leq u) = \int_{-\infty}^{+\infty} F_{\log(X)}\left(\frac{\log(u) - \varepsilon - b}{a}\right) f_E(\varepsilon) d\varepsilon$$

$$P(Y \leq u) = \int_{-\infty}^{+\infty} \frac{1}{1 + \exp\left(\frac{\varepsilon + b}{as} + \frac{\mu}{s}\right) u^{-1/as}} f_E(\varepsilon) d\varepsilon$$

De même :

$$P(Y \geq u) = \int_{-\infty}^{+\infty} \frac{1}{1 + \exp\left(-\frac{\varepsilon + b}{as} - \frac{\mu}{s}\right) u^{1/as}} f_E(\varepsilon) d\varepsilon$$

Expression de la prime pure avec une limite de couverture L :

$$E[\min(Y, L)] = \int_0^L P(Y \geq y) dy$$

On a $s = 1{,}57 \pm 0.06$ et $a \in [0{,}70; 1{,}20]$ donc $as \in [1{,}05; 1{,}95]$

On remarque que l'erreur type de b est plus grande que σ_E dans le tableau partie III.G.2, **on va donc négliger $\boldsymbol{\varepsilon}$**, et l'intégrer dans les variations de b autour de sa moyenne. L'intervalle de fluctuation de b est alors $\pm 2 * (\sigma_E + Err.Type.(b))$. Notons que les variations de μ/s sont en comparaison négligeables.

D'autre part on a une probabilité de survenance du sinistre P_s de 2,5% pour la vente et 0,7% pour les activités autres que la vente, la banque et assurance, la santé et l'éducation. En utilisant l'hypothèse d'indépendance entre fréquence et sévérité, la prime pure est alors :

$$P_s * E[\min(Y, L)]$$

On pose $k = \exp\left(\frac{b}{as} + \frac{\mu}{s}\right)$.

On teste le modèle dans le cas où $as = 1$ donc s petit, on prend donc b = 7,9. Ces paramètres nous placent dans les pays comme l'Australie ou les USA. L'intervalle de fluctuation de b est alors $[5{,}9; 9{,}9]$.

$$E[\min(Y, L)] = \int_0^L \frac{1}{1 + {}^{y}/_{k}} dy$$

$$E[\min(Y, L)] = \mathrm{k} * \log\left(1 + {}^{L}/_{k}\right)$$

Pour le secteur de la vente et en choisissant une limite à 1 million de dollars, la prime pure est de 13700 dollars. En faisant varier b, on trouve une valeur minimale de 4650 dollars et une valeur maximale de 22100 dollars.

Pour les autres activités et en conservant la limite à 1 million de dollars, la prime pure est de 3800 dollars. En faisant varier b, on trouve une valeur minimale de 1300 dollars et une valeur maximale de 6200 dollars.

On teste le modèle dans le cas où $as = 2$ donc s grand, on prend donc b = 2,65. Ces paramètres nous placent dans les pays comme la France ou l'Italie. L'intervalle de fluctuation de b est alors $[0{,}65; 4{,}65]$.

$$E[\min(Y, L)] = \int_0^L \frac{1}{1 + {}^{\sqrt{y}}/_{k}} dy$$

Avec le changement de variable $z = \sqrt{y}$:

$$E[\min(Y, L)] = \int_0^{\sqrt{L}} \frac{2z}{1 + {}^{z}/_{k}} dz$$

$$E[\min(Y, L)] = 2 * \mathrm{k}\left(\sqrt{L} - \mathrm{k} * \log\left(1 + {}^{\sqrt{L}}/_{k}\right)\right)$$

Pour la vente et en choisissant une limite à 1 million de dollars, la prime pure est de 10600 dollars. En faisant varier b, on trouve une valeur minimale de 5600 dollars et une valeur maximale de 16400 dollars.

Pour les autres activités et en conservant la limite à 1 million de dollars, la prime pure est de 3000 dollars. En faisant varier b, on trouve une valeur minimale de 1600 dollars et une valeur maximale de 4600 dollars.

Nous estimons que les primes pures sont réalistes puisque qu'elles sont proches des tarifs donnés en III.B.5. Mais elles montrent une sensibilité par rapport aux paramètres du modèle.

Il faut remarquer que le modèle et les données utilisées ne prennent pas en compte les éléments suivants :

- Il est courant d'avoir une franchise dans les contrats, ce qui n'est pas modélisé. Mais cela permet de limiter l'impact de l'erreur d'estimation faite sur les petits sinistres.
- Il peut y avoir des exclusions dans les contrats, que nous ne prenons pas en compte.

Cela justifie des primes pures estimées plutôt élevées par rapport aux tarifs du marché.

I. Evolution du risque

Les progrès techniques des dernières années ont profondément changé la pénétration de l'informatique dans les entreprises. Il en résulte donc une évolution du cyber-risque de son apparition à aujourd'hui.

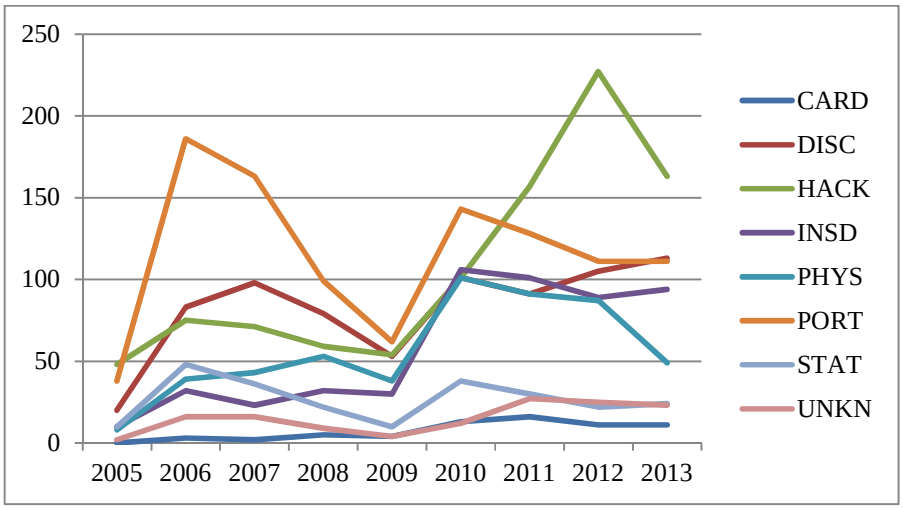


Figure 33 Nombre de violations par année et par type

La question pour l'actuaire est alors, le risque va-t-il continuer à évoluer au même rythme, ou se stabiliser ? Il apparait dans les études des attaques ces dernières années un glissement du risque d'une technologie à une autre suite à des prises de conscience ou des évolutions technologiques qui ont rendu d'anciens outils moins vulnérables. Par exemple, les outils intégrés aux systèmes d'exploitation pour limiter l'escalade des droits ont limité la fragilité des postes fixes. Combinés au déploiement systématique des anti-virus, cela a rendu bien plus difficile la propagation des virus informatiques. En conséquence ce type de menace a fortement diminué ces dernières années. En contrepartie, le développement des réseaux sociaux incite les gens à laisser de plus en plus d'informations personnelles sur internet, qui risque alors d'être utilisées pour faire du fishing.

Cependant le volume d'attaque semble se stabiliser. On peut donc penser qu'il y a une évolution des techniques d'attaque mais pas un changement en volume.

Enfin, la porosité grandissante entre les outils informatiques à but personnel et ceux étant à but professionnel pose aussi des difficultés pour la protection des systèmes d'informations modernes. De façon similaire, les réseaux informatiques entre un donneur d'ordre et un prestataire sont de plus en plus connectés entre eux. Une faille mineure chez l'un d'eux peut devenir un risque majeur pour l'autre, voire un client de l'autre. Se posent alors les problèmes de responsabilité ainsi que la limite du contrat d'assurance pour l'entreprise.

J. Limites des modèles

La population assurée peut être différente de la population assurable. En effet, le modèle est construit sur toute la population or un assureur ne s'intéresse qu'à la population qu'il assure. De plus, la population assurée peut avoir un comportement ou des caractéristiques spécifiques. Pour un actuaire qui utiliserait ce modèle il est important qu'il étudie les spécificités de sa clientèle.

Nous avons fait l'hypothèse que le risque d'attaque et la taille des violations ne dépendent pas du pays. Nous l'avons faite faute d'avoir des données plus précises mais il faudrait la vérifier et éventuellement la remettre en cause. En effet, pourvu d'avoir les données adéquates le modèle n'a pas besoin de cette hypothèse.

Nous avons été limités dans l'étude de la relation entre la taille des violations et le coût par les données disponibles. Il serait nécessaire d'approfondir cette étude, en particulier sur les gros sinistres. Il peut alors apparaître que le modèle de relation linéaire peut ne pas convenir, c'est pourquoi le développement du modèle basé sur les copules peut devenir nécessaire. Cette problématique justifie que nous ayons vu l'approche avec les copules même si par la suite elle est écartée.

Certains coûts peuvent être exclus des contrats. Nous n'étudions pas ici les influences des exclusions ou des franchises sur les primes des assurés.

L'Institut Ponemon ne fournit pas ses bases de données avec les informations de sectorisations, mêmes si ces informations sont collectées. Cependant, dans les rapports il apparait que la relation entre la taille et le coût dépend du domaine d'activité. Il semble donc nécessaire d'approfondir l'étude de la segmentation du marché avec des données plus précises telles que le secteur d'activité ou la taille de l'entreprise.

Nous avons concentré l'étude sur la prime pure, mais d'autres paramètres influencent le tarif. Par exemple, les attaques informatiques peuvent se faire depuis n'importe quel lieu dans le monde. Pour déterminer les chargements de sécurité il faudrait étudier les dépendances entre pays.

Enfin j'attire l'attention sur l'hypothèse d'indépendance entre fréquence et sévérité. C'est une hypothèse très courante dans les modèles actuariels pour des raisons pratiques. Elle est tout de même discutable.

IV. Conclusion

Nous avons, dans un premier temps, permis au lecteur de se familiariser avec les spécificités du cyber-risque. Nous avons présenté le vocabulaire technique et métier, l'environnement réglementaire, le marché, des études de cas, les normes et les modélisations techniques utiles. Cela permettra au lecteur voulant approfondir le sujet d'avoir les bases lui permettant de faire des études plus poussées.

Un souscripteur peut donc trouver dans ce mémoire le vocabulaire lui permettant de communiquer avec son client ainsi que trouver les indicateurs qui pourraient lui permettre de se faire un meilleur avis du risque assuré. L'actuaire trouvera une présentation technique lui permettant de replacer ses études dans leur contexte. Enfin, l'assureur pourra trouver des informations sur le marché actuel ainsi que les partenaires dont il pourrait avoir besoin sur ce marché.

Nous avons aussi étudié les **données en libre accès**. Nous avons montré qu'on peut, simplement avec ces données, en déduire une **évaluation prudente de la prime pure** sur toute la population assurable. Cela peut donc aider à constituer une étude faite sur les données confidentielles d'un assureur. Il reste, en particulier, à affiner l'évaluation sur les gros sinistres et faire une segmentation plus fine. Notons que nous avons montré la possibilité théorique d'utiliser les copules. Même si ce modèle ne présentait pas d'intérêt avec les données disponibles, il pourrait être utile sur une base de sinistres comportant des incidents majeurs.

D'autre part, nous avons observé que l'exploitation des commentaires à l'aide des techniques de **fouille de texte** nous apporte des informations utiles. En particulier, nous avons trouvé quelques indications sur la durée des sinistres. De plus, le modèle est construit sur l'étude de la relation entre le volume de données impactées et le coût final. Le volume de données impactées est une information généralement connue rapidement lorsque le sinistre survient. Nous remarquons donc que ces informations pourraient affiner les modèles de **provisionnement** de ce type de sinistres. Les données fournies en commentaire de la base de violation pourraient fournir plus d'informations grâce à l'utilisation de méthodes de fouilles de données plus précises.

Enfin nous avons présenté des indicateurs, en fin de partie II, qui ne sont pas utilisés faute de données. Ils constituent tout de même une source d'améliorations potentielles de la qualité de la tarification. Nous avons aussi présenté un ensemble de risques en partie I qui ne sont pas étudiés d'un point de vue quantitatif et qui nécessiteraient une étude spécifique.

V. Bibliographie

ANDERSON, Ross J. 1994. Liability and Computer Security: Nine Principles. *Computer Security.* 1994, Vol. ESORICS 94, pp. 231-245.

BARRACCHINI, Carla et ADDESSI, M. Elena. 2013. Cyber Risk and Insurance Coverage: An Actuarial Multistate Approach. *Review of Economics & Finance.* 2013, 1923-7529-2014-01-57-13.

BÖHME, Rainer et KATARIA, Gaurav. 2006. Models and Measures for Correlation in Cyber-Insurance. *WEIS.* 2006.

BÖHME, Rainer et SCHWARTZ, Galina. 2010. *Modeling Cyber-Insurance: Towards A Unifying Framework.* Harvad : Workshop on the Economics of Information Security, 2010.

BOUISSOU, Marc et BON, J-L. 2003. A new formalism that combines advantages of fault-trees and Markov models: Boolean logic driven Markov processes. *Reliability Engineering & System Safety.* 2003, pp. 149-163.

BOUISSOU, Marc, KRIAA, Siwar et PIÈTRE-CAMBACÉDÈS, Ludovic. Modeling the Stuxnet Attack with BDMP: Towards More Formal Risk Assessments.

DBIS, UK. 2014. Cyber essentials scheme: overview. *GOV.UK.* [En ligne] 2014. https://www.gov.uk/government/publications/cyber-essentials-scheme-overview.

GENEST, Christian et RIVEST, Louis-Paul. 1993. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American statistical Association.* 1993, Vol. 88, 423.

HERATH, Hemantha S. B. et HERATH, Tejaswini. 2007. Cyber-Insurance: Copula Pricing Framework and Implication for Risk Management. *WEIS.* 2007.

INNERHOFER-OBERPERFLER, Frank et BREU, Ruth. 2010. Potential rating indicators for cyberinsurance: An exploratory qualitative study. *Economics of Information Security and Privacy.* Springer US, 2010, pp. 249-278.

MAILLART, T. et SORNETTE, D. 2009. Heavy-tailed distribution of cyber-risks. [En ligne] 2009. http://arxiv.org/abs/0803.2256..

MICONNET, Emmanuel, et al. 2013. *Un exemple d'usage des graphes d'attaques pour l'évaluation dynamique des risques en cyber-sécurité.* 2013.

MUKHOPADHYAY, Arunabha, et al. 2006. e-Risk management with insurance: a framework using copula aided Bayesian belief networks. 2006.

MULLINER, Collin. 2009. [En ligne] 2009. http://www.blackhat.com/presentations/bh-usa-09/MILLER/BHUSA09-Miller-FuzzingPhone-PAPER.pdf.

Ponemon. 2014. Ponemon Institute Releases 2014 Cost of Data Breach: Global Analysis. *Ponemon Institute.* [En ligne] 05 05 2014. http://www.ponemon.org/blog/ponemon-institute-releases-2014-cost-of-data-breach-global-analysis.

RASPOTNIG, Christian, KARPATI, Peter et KATTA, Vikash. 2012. https://bora.uib.no/bitstream/handle/1956/6172/Guideline_for_applying_CHASSIS_draft_BORA.pdf. [En ligne] Bergen Open Research Archive, 09 11 2012.

SERAZZI, Giuseppe et ZANERO, Stefano. 2004. Computer virus propagation models. *Performance Tools and Applications to Networked Systems.* 2004.

TOREGAS, Costis et ZAHN, Nicolas. 2014. *Insurance for Cyber Attacks: The Issue of Setting Premiums in Context.* s.l. : George Washington University, 2014.

VIJAYAN, Jaikumar. 2010. Heartland breach expenses pegged at $140M -- so far. *computerworld.* [En ligne] 10 Mai 2010. http://www.computerworld.com/s/article/9176507/Heartland_breach_expenses_pegged_at_140M_so_far.

—. 2008. One year later: Five takeaways from the TJX breach. *computerworld.* [En ligne] 17 Janvier 2008. http://www.computerworld.com/s/article/9057758/One_year_later_Five_takeaways_from_the_TJX_breach.

Wiki1. Arbre de defaillances. *Wikipedia.* [En ligne] fr.wikipedia.org/wiki/Arbre_de_defaillances.

Wiki2. EBIOS. *Wikipedia.* [En ligne] fr.wikipedia.org/wiki/EBIOS.

Wiki3. Heartland Payment Systems. *Wikipedia.* [En ligne] http://en.wikipedia.org/wiki/Heartland_Payment_Systems.

Wiki4. Target Corporation. *Wikipedia.* [En ligne] http://en.wikipedia.org/wiki/Target_Corporation.

Wiki5. PlayStation Network outage. *Wikipedia.* [En ligne] http://en.wikipedia.org/wiki/PlayStation_Network_outage.

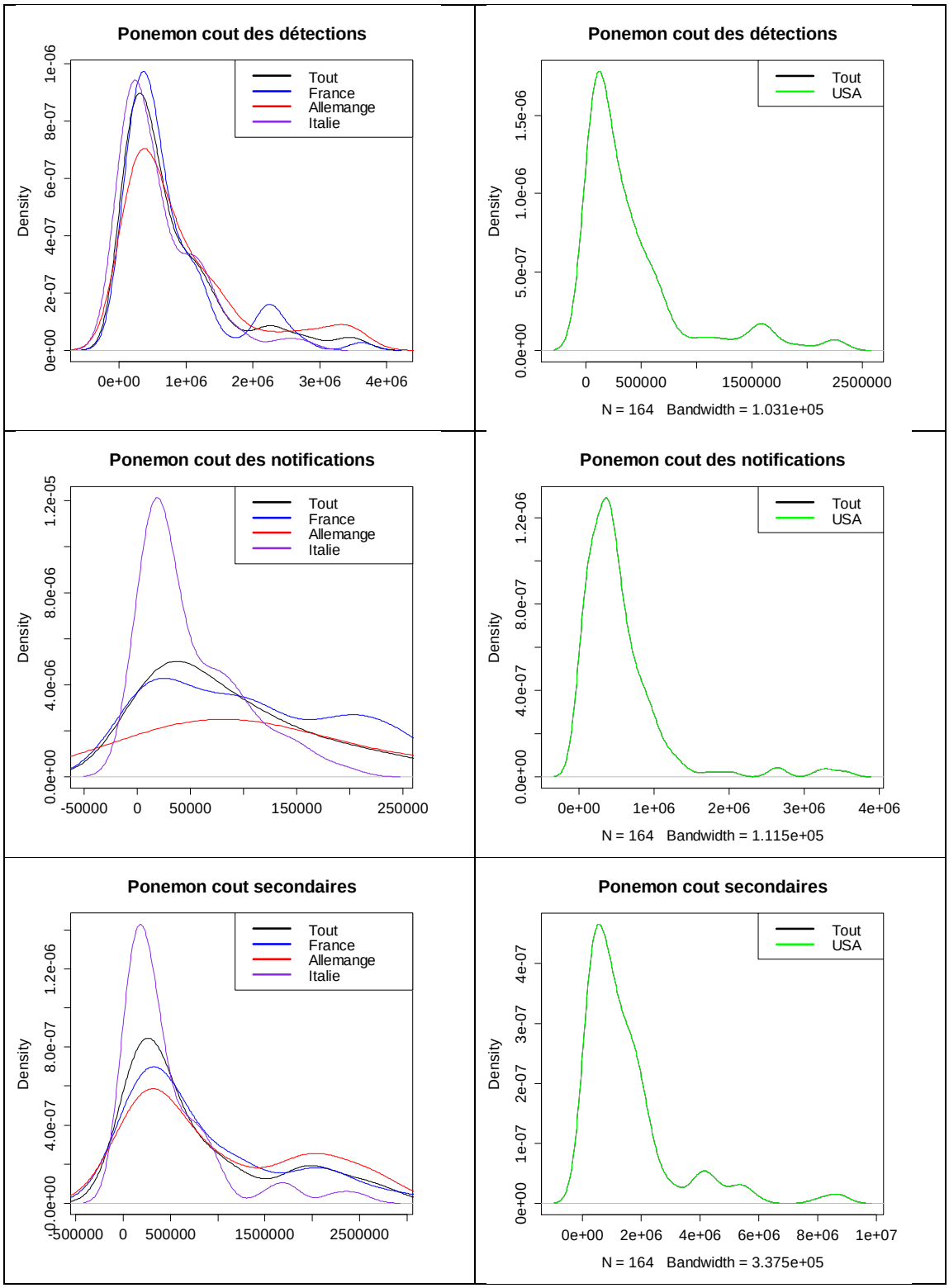
Wiki6. Stuxnet. *Wikipedia.* [En ligne] http://fr.wikipedia.org/wiki/Stuxnet.

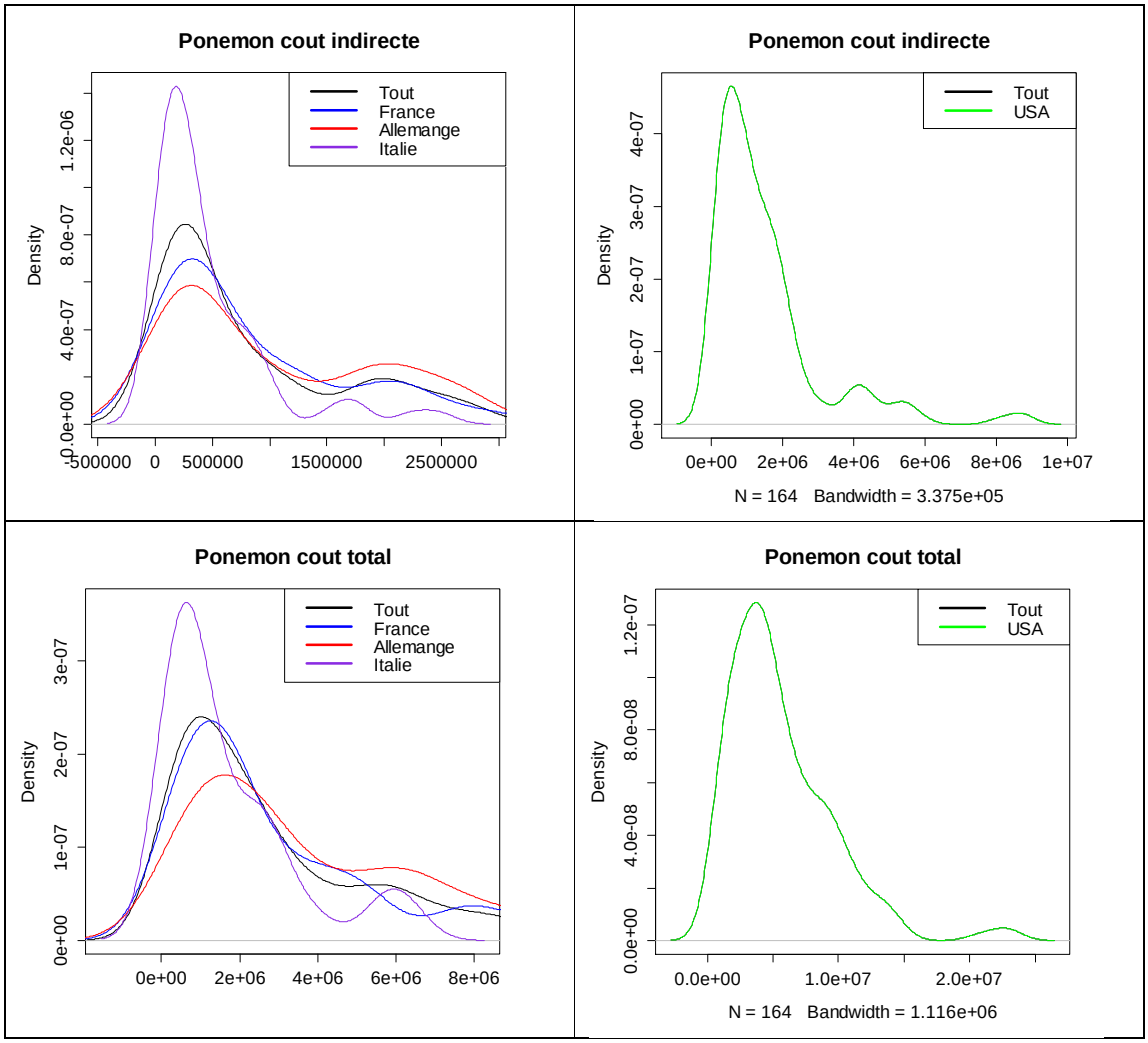
ZDnet. 2014. *ZDnet.* [En ligne] 05 05 2014. http://www.zdnet.com/target-ceo-out-after-massive-cyberattack-cfo-to-replace-7000029064/.

Annexe A Données Ponemon

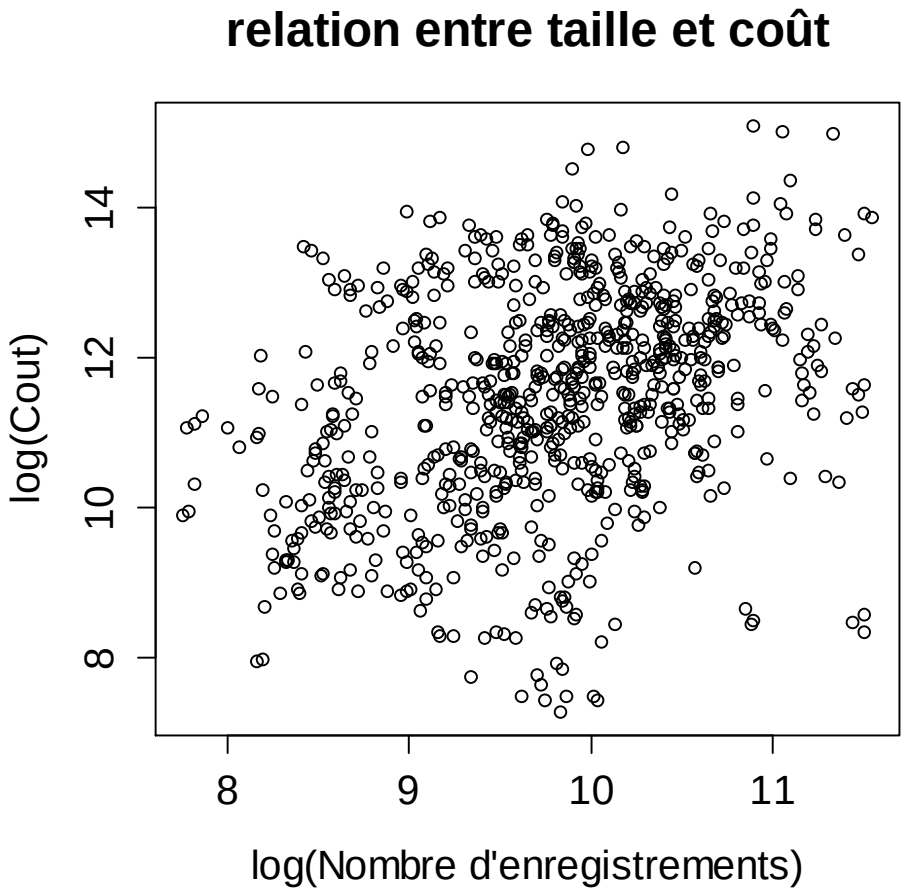
Année	Pays	Nb Entreprises	Nb Enregistrement par entreprise	Coût moyen par incident
2006	USA	14	100 000	$13 800 000
2007	USA	31	26 290	$4 784 780
2008	USA	35	31 980	$6 300 060
2009	USA	43	32 921	$6 650 042
2010	USA	45	33 088	$6 749 952
2011	USA	51	33 645	$7 200 030
2012	USA	49	28 351	$5 500 094
2013	USA	54	28 765	$5 407 820
2014	USA	61	29 087	$5 846 487
2008	UK	21	30 213	£1 420 000
2009	UK	30	28 833	£1 730 000
2010	UK	33	26 250	£1 680 000
2011	UK	38	26 761	£1 900 000
2012	UK	36	22 152	£1 750 000
2013	UK	38	23 833	£2 049 638
2014	UK	40	23 365	£2 219 675
2009	DE	19	21 518	2 410 000 €
2010	DE	22	19 545	2 580 000 €
2011	DE	25	24 493	3 380 000 €
2012	DE	26	23 288	3 400 000 €
2013	DE	31	24 280	3 666 280 €
2014	DE	30	24 371	3 411 940 €
2010	FR	17	21 348	1 900 000 €
2011	FR	21	22 449	2 200 000 €
2012	FR	23	20 902	2 550 000 €
2013	FR	26	22 462	2 852 674 €
2014	FR	27	22 869	3 018 708 €
2012	IT	18	17 792	1 387 798 €
2013	IT	22	18 285	1 737 075 €
2014	IT	23	19 034	1 941 468 €

Annexe B Graphique des coûts Ponemon



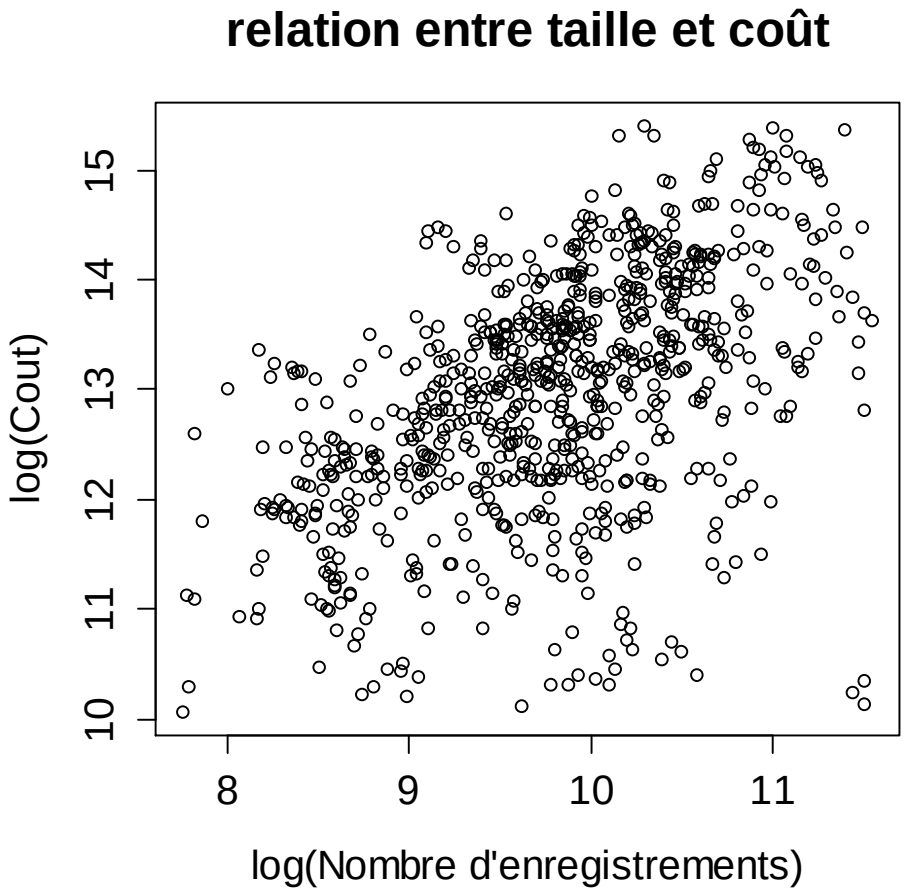

Ponemon cout indirecte
Tout
France
Allemange
Italie
Density
0.0e+00
4.0e-07
8.0e-07
1.2e-06
-500000
0
500000
1500000
2500000
Ponemon cout indirecte
Tout
USA
Density
0e+00
1e-07
2e-07
3e-07
4e-07
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
N = 164 Bandwidth = 3.375e+05
Ponemon cout total
Tout
France
Allemange
Italie
Density
0e+00
1e-07
2e-07
3e-07
0e+00
2e+06
4e+06
6e+06
8e+06
Ponemon cout total
Tout
USA
Density
0.0e+00
4.0e-08
8.0e-08
1.2e-07
0.0e+00
1.0e+07
2.0e+07
N = 164 Bandwidth = 1.116e+06

Annexe C Relation entre Taille et Coût de Notification



Pays	a	p-value	Err. Std.	b	p-value	Err. Std.	$\sigma(\epsilon)$
IN	0,6153	7,76E-05	0,1461	3,6927	0,0121	1,4321	0,9609
AU	0,1384	0,394	0,1613	9,5217	7,91E-08	1,5674	0,8486
USA	0,1975	0,137	0,1323	10,7387	1,48E-13	1,3304	1,197
AR	0,8728	4,66E-02	0,4114	2,7688	5,09E-01	4,1141	1,36
UK	-0,2653	0,0481	0,1327	14,6168	<2e-16	1,3089	0,9591
JA	1,2297	<2e-16	0,0951	-0,8554	3,47E-01	0,903	0,6563
BR	0,639	0,0015	0,1914	4,3149	2,51E-02	1,8753	1,162
DE	0,4779	0,00199	0,1498	7,4548	2,47E-06	1,4755	1,066
IT	0,86	2,74E-10	0,1141	2,4968	2,50E-02	1,0866	0,7331
FR	0,7409	0,000202	0,1894	4,061	3,12E-02	1,8493	1,28

Annexe D Relation entre Taille et Coût de détection



Pays	a	p-value	Err. Std.	b	p-value	Err. Std.	$\sigma(\epsilon)$
IN	0,79976	1,49E-12	0,09436	4,74387	2,45E-06	0,92968	0,7009
AU	0,66006	2,64E-13	0,07155	7,23338	2,64E-15	0,69544	0,3765
USA	0,3318	0,00746	0,1225	9,0088	1,12E-11	1,2314	1,108
AR	0,6359	3,30E-02	0,2795	5,9681	4,51E-02	2,809	0,9716
UK	0,99445	<2e-16	0,08016	3,31057	5,62E-05	0,79041	0,5792
JA	0,9782	3,01E-13	0,1072	3,915	2,77E-04	1,0181	0,7399
BR	0,96141	<2e-16	0,08084	3,14977	1,89E-04	0,79233	0,4923
DE	1,00748	<2e-16	0,08896	3,71642	5,63E-05	0,87622	0,6328
IT	1,0606	2,04E-13	0,1132	2,9943	7,29E-03	1,0784	0,7276
FR	1,02583	<2e-16	0,08153	3,43207	4,93E-05	0,79599	0,5508

Annexe E Notations

Notation	Définition
log(x)	Logarithme népérien de x
P(H)	Probabilité de l'assertion H
E[X]	Espérance de la variable aléatoire X