UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Le risque cyber

Auteur : De Cock Nathan
Promoteur : Philippe De Longueville
Lecteur : Michel Denuit
Maîtres de stage : Julie Zians et Geoffrey Feraut
Année académique 2022-2023
Master en sciences actuarielles

Avant-Propos.

J'ai réalisé ce mémoire au sein de la société Reacfin qui est une firme de consultance fondée en 2004 en tant que spin-off de l'UCLouvain. Reacfin est une entreprise spécialisée dans l'actuariat dont toutes les spécificités sont à découvrir sur son site internet https://www.reacfin.com.

Les graphiques, les fonctions, l'estimation de paramètres, les simulations et les tests ont été réalisés à l'aide de R.

Remerciements.

Je souhaite remercier toutes les personnes qui, de près ou de loin,m'ont aidé et soutenu tout au long de mon parcours académique.

Je suis très reconnaissant envers Reacfin grâce à qui j'ai eu l'opportunité de connaître une première expérience professionnelle au sein d'une équipe exclusivement composée de collaborateurs très compétents. Je profite de cette occasion pour souligner l'extrême gentillesse et le professionnalisme de Julie Zians, Geoffrey Feraut et Michael Lecuivre. Leur aide m'a été très précieuse dans l'élaboration de ce travail.

Je remercie aussi Philippe De Longueville d'avoir non seulement accepté d'être mon promoteur dans le cadre de ce mémoire mais aussi pour le temps et l'énergie qu'il y a consacré. Ses observations et remarques étaient toutes aussi pertinentes les unes que les autres.

Je n'oublie pas de citer Gordon Clark et Caroline Hillairet qui m'ont énormément apporté en me faisant partager leur connaissance dans le domaine.

Je tiens aussi à saluer la participation de SAS qui m'a fourni les données nécessaires sur le sujet choisi.

Enfin, pour terminer, je ne peux m'empêcher d'avoir une pensée pour ma famille qui m'a sans cesse soutenu tout au long de mon parcours scolaire.

Pour conclure, je dédie tout spécialement ce travail à mon grand-père qui combat une très lourde maladie et qui a toujours éprouvé une grande fierté envers mes résultats scolaires.

Résumé.

La croissance du numérique rend les entreprises de plus en plus vulnérables aux attaques cyber. Celles-ci souhaitent donc se couvrir face à ce risque de plus en plus menaçant. Cependant, les assureurs n'ont pas encore une grande maîtrise du risque cyber. Ce mémoire, dans un premier temps, définira ce risque, tentera d'identifier les caractéristiques du risque cyber et analysera ce qui est déjà proposé sur le marché. Une étude purement descriptive sera réalisée pour y dégager certaines tendances. Ensuite, le processus de Hawkes permettra de modéliser la fréquence du risque cyber. L'effet auto-excitant de ce processus donne la possibilité de ne pas négliger la dépendance entre les polices lors de la modélisation. Une application univariée et multivariée sera appliquée à des données. Enfin, le coût d'un sinistre pouvant être extrêment élevé, la théorie des valeurs extrêmes sera étudiée pour le modéliser. A l'aide d'un arbre de régression, une comparaison sera faite entre les facteurs de risque de la partie centrale de la distribution et ceux ayant une influence sur la queue de la distribution.

Summary.

The growth of digital technology is making businesses increasingly vulnerable to cyber attacks. As a result, they are looking to insure themselves against this increasingly threatening risk. However, insurers do not yet have a firm grasp of cyber risk. This brief will begin by first, define this risk, attempt to identify the characteristics of cyber risk and analyse what is already being already available on the market. A purely descriptive study will be carried out to identify certain trends. Next, the Hawkes process will be used to model the frequency of the cyber risk. The self-exciting effect of this process means that the dependency between policies dependency between policies when modelling. A univariate and multivariate application will be applied to the data. Finally, as the cost of a claim can be extremely high, extreme value theory will be used to model it. Using a regression tree, a comparison will be made between the risk factors in the central part of the distribution and those the tail of the distribution.

Table des matières

Introduction

Selon le rapport annuel d'Allianz [1], le principal risque auquel les entreprises dans le monde entier sont confrontées est le risque cyber. C'est aussi la plus grande menace pour les entreprises belges. Il devance même des risques tels que les catastrophes naturelles ou l'interruption de leur activité. De plus en plus de sociétés sont conscientes de cette nouvelle réalité et se rendent compte des dégâts que celle-ci peut leur amener. En effet, les atteintes peuvent être de plusieurs formes : une attaque cyber peut interrompre l'activité d'une entreprise, être extrêmement coûteuse ou endommager son image. Selon une étude réalisée par NetDiligence [2], en 2020, le coût moyen d'un sinistre pour une PME s'élève à 286 000$. De plus, avec les années les entreprises sont de plus en plus sujettes aux attaques cyber. Le rapport annuel d'Hiscox 2022 [3] rapporte que 43% des entreprises ont subi une attaque de type informatique en 2020. Ce chiffre est passé à 48% en 2021. La récente pandémie du coronavirus n'a pas aidé les entreprises, la plupart des travailleurs ayant dû faire du télétravail. Premièrement, les réseaux internet à domicile ne sont pas autant sécurisés que ceux en place dans les sociétés. Deuxièment, les entreprises furent encore plus interconnectées qu'elles ne l'étaient déjà. Conscientes de l'impact que cela peut avoir sur leur business, les entreprises souhaitent se protéger du risque cyber. Malheureusement, le marché de l'assurance pour le risque cyber n'est pas encore très étendu. Diverses raisons peuvent expliquer ce manque de développement. Dans un premier temps, étant donné que c'est un risque relativement nouveau, le manque de données est le principal problème auquel les assureurs sont confrontés. Sans données, il est en effet difficile pour un assureur de tarifer un produit. De plus, les attaques ne sont pas souvent rendues publiques par les entreprises qui en sont victimes car elles n'ont aucune obligation de signaler ce genre de méfaits. De surcroît, rendre cela public serait néfaste à leur image. Une difficulté supplémentaire pour le risque cyber est que l'indemnisation d'un sinistre ne se limite pas à une intervention financière mais se couple à une assistance afin de remettre les systèmes informatiques et les données en ordre. Pour illustrer cette remarque, prenons l'exemple de Manutan [4]. Cette entreprise fut victime d'une attaque cyber qui a eu comme conséquence

1. Allianz Risk Barometer 2022 : https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/press/document/Allianz_Risk_Barometer_2022_FINAL.pdf
2. NetDiligence Cyber Claims study 2021 : https://netdiligence.com/cyber-claims-study-2021-report/
3. Hiscox Cyber Readiness Report 2022 : https://www.hiscoxgroup.com/sites/group/files/documents/2022-05/22054%20-%20Hiscox%20Cyber%20Readiness%20Report%202022-EN_0.pdf
4. Manutan est une entreprise française spécialisée dans la distribution de fournitures et d'équipements industriels, de bureau et de manutention.

que l'entreprise a perdu des données, ne recevait plus les paiements et son système informatique était hors service. L'assureur, en plus de couvrir la perte financière liée à cette attaque, a fourni l'assistance pour changer le système informatique et pour récupérer les données. C'est une étape nécessaire sans quoi les personnes malveillantes pourraient reproduire l'attaque puisqu'ils connaissent la faille du système informatique. Cela mène à une autre problématique qui est qu'une tarification basée sur des variables explicatives peut ne pas être pertinente pour certaines entreprises étant donné que la nature du risque cyber peut changer une fois que l'entreprise est plus importante. Le système informatique par exemple est vraiment propre à chaque entreprise.

Ce mémoire sera divisé en trois grandes parties. Dans un premier temps, une révision de la notion du risque cyber sera réalisée, notamment, les caractéristiques de ce risque qui font de lui un danger à part. Ensuite, une étude des couvertures existantes sera effectuée afin de mieux connaître ce qui est déjà proposé sur le marché. Evaluer le coût d'une telle couverture est une chose ardue, notamment car le coût total d'une attaque peut s'étaler sur une longue période. De plus, l'assurance du risque cyber ne propose pas uniquement un dédommagement financier mais aussi des services en cas de sinistre afin d'aider l'entreprise victime. Suite aux entretiens avec Caroline Hillairet et Gordon Clark, deux personnes ayant de l'expérience sur le sujet, un compte rendu de nos échanges sera réalisé.

Le chapitre qui suit sera consacré à la modélisation de la fréquence sinistre du risque cyber. Pour commencer, une étude générale sera effectuée en utilisant des données collectées sur le site de l'assureur d'Hiscox. Cette première étude permettra de connaître de manière assez générale les premières tendances du risque cyber. Ces résultats seront également commentés par une personne travaillant en tarification du produit cyber. Le modèle a pour objectif de prendre en compte la dépendance entre les polices. La dépendance est à priori une caractéristique assez importante du risque cyber. Le modèle consacré à la fréquence sera celui de Hawkes. C'est le modèle qui sera étudié le plus en profondeur. Il s'agit d'un processus de Poisson avec une intensité qui n'est pas déterministe. L'intensité évoluera avec le temps en fonction des sinistres ayant déjà eu lieu afin de prendre en considération l'effet auto-excitant, le risque d'accumulation, du risque cyber. Une présentation détaillée du processus sera réalisée afin de l'appliquer ensuite à des données. Les limites de cette modélisation seront également évoquées.

Le dernier chapitre discutera de la partie sévérité du risque cyber et plus particulièrement des grands sinistres. A partir de données, il sera montré via un arbre de régression qu'il est important pour un assureur de bien distinguer la partie centrale de la distribution de la queue de la distribution. Ces deux parties doivent avoir une modélisation distincte pour avoir un modèle qualitatif et parce que les facteurs de risque ne sont pas les mêmes pour la partie centrale que ceux pour la queue de distribution.

Chapitre 1

Définitions et propriétés du risque cyber

Le risque cyber étant encore un phénomène relativement récent dans le secteur des assurances, il n'existe pas encore de définition claire pour ce dernier. Lors de ce chapitre, une définition la plus complète possible sera proposée et les différentes propriétés du risque cyber seront analysées. Il est primordial de veiller à comprendre les caractéristiques d'un risque avant d'entamer sa modélisation. Pour conclure ce chapitre, une étude d'une police d'assurance qui couvre le risque cyber sera faite.

1.1 Définition

Les différents acteurs du monde des assurances ne sont à ce jour pas encore parvenus à s'accorder sur une définition précise du risque cyber. D'ailleurs, celle-ci évolue avec le temps de par la nature du risque. En 2006, BÖHME, Rainer et KATARIA, Gaurav [1] définissent le risque cyber comme une défaillance du système informatique. Des définitions plus précises vont néanmoins apparaître au fil du temps et c'est ainsi qu'en 2010, CEBULA, James L. et YOUNG, Lisa R [2] définissent le risque cyber comme un risque opérationnel auquel une entreprise est sujette, un risque opérationnel pour les actifs de nature informatique et technologique qui affectent la confidentialité, l'intégrité et la disponibilité d'informations ou/et de systèmes d'informations. En 2014, l'Institut de Risk Management [1] définit le risque cyber comme étant un risque de perte financière, d'interruption de l'activité ou un dommage à la réputation suite à une défaillance du système informatique de la société. D'autre part, l'Institut de Risk Management attire l'attention sur le fait que cette attaque peut être provoquée de manière intentionnelle par une personne malveillante mais également de manière non-intentionnelle, voire accidentelle. À l'heure actuelle, les différents acteurs semblent s'accorder sur la dernière définition. A titre d'exemple, la MatMut définit le risque cyber comme une atteinte aux sytèmes informatiques et/ou aux données informatisées qui sont la conséquence d'un acte accidentel ou mal intentionné. Au fil du temps, les définitions

1. Cyber Risk. Resources for Practitioners. - The Institute of Risk Management : https://www.theirm.org/media/7237/irm-cyber-risk-resources-for-practitioners.pdf

insistent sur l'aspect accidentel ou non. En 2016, ELING, M. et SCHNELL, W. [4] définissent le risque cyber comme suit : "Le risque cyber comprend tous les risques liés aux technologies que cela soit lié à la communication ou à l'information, qui compromettent l'intégrité, la disponibilité ou la confidentialité de données ou de systèmes. La défaillance de la technologie provoquera un arrêt de l'activité, une panne ou des dommages aux humains ou aux biens. Le risque cyber est soit provoqué par des catastrophes naturelles, soit par des interventions humaines (cybercriminalité, attaques terroristes, de simples erreurs humaines, ou des guerres cyber). De plus, il est caractérisé par une grande dépendance, des évènements qui peuvent être extrêmes et des risques de changement.". En comparaison aux définitions citées précédemment, celle-ci précise la source du risque (technologies informatives et communicatives), l'objet du risque (dépendance, évènements extrêmes, risque évolutif) et l'impact du risque (défaillance, dégats corporels ou matériels). La définition utilisée sera celle établie par l'Association des Professionnels de la Réassurance de France [2] .

Définition 1.1.1. Le risque cyber désigne toute atteinte à :

- des systèmes électroniques et/ou informatiques (de production, d'exploitation, de gestion d'information et de télécommunication) sous le contrôle de l'entité ou de ses prestataires et/ou
- des données informatisées (personnelles, confidentielles ou d'exploitation) appartenant à l'entité, ou sous le contrôle de celle-ci, qu'elles soient transférées ou stockées chez elle ou chez ses prestataires

Consécutive à :

- un acte malveillant ou de terrorisme
- une erreur humaine, une panne ou des problèmes techniques
- un évènement naturel ou accidentel

Ayant pour conséquences :

- des dommages corporels, matériels, et/ou immatériels (frais ou pertes financières), subis par l'entité et/ou ses employés
- une mobilisation de ressources internes ou externes
- des dommages corporels, matériels, et/ou immatériels, frais ou pertes financières causés par l'entité à des tiers (y compris chaines logistiques / sous-traitants)
- une atteinte à la marque et/ou à la réputation de l'entité

De par cette définition, le risque cyber englobe beaucoup de notions. C'est un terme assez large et il est donc important pour un assureur de bien définir sa police d'assurance. En effet, le cyber risque peut s'étendre de l'employé qui endommage un ordinateur jusqu'à l'interruption totale de l'activité d'une entreprise pendant plusieurs jours. Dans les deux cas, l'assureur pourrait

2. Etude sur les cyber risques et leur Réassurabilité - Association des professionels de la réassurance en France : https://www.apref.org/wp-content/uploads/2020/10/note_apref_cyber_risque.pdf#:~:text=L%E2%0%99APREF%20souhaite%20par%20cette%20note%20contribuer%20%C3%A0%20une,mieux%20encadrer%20sa%20d%C3%A9finition%20et%20favorise%20sa%20compr%C3%A9hension.

intervenir mais le fera-t-il lors de chaque incident commis par un employé où il couvrirait les pertes équivalentes à des montants astronomiques ? La partie 2.4 sera consacrée aux polices existantes relatives au risque cyber.

1.2 Types de risques

Dans cette partie, les différents types de risques seront parcourus. Le monde de la technologie est vaste et en constante évolution, il existe donc différentes méthodes de piratage. Il est important pour l'assureur de pouvoir distinguer les types de menace auxquels les assurés sont sujets. En effet, l'identification des types de risques permet une meilleure modélisation : certaines attaques sont peut-être plus fréquentes que d'autres tandis que d'autres auront des conséquences plus importantes. Ci-dessous est présentée une liste (non-exhaustive) de types d'attaque cyber qui pourraient éventuellement être couverts par un assureur.

Phishing : Technique frauduleuse visant à leurrer l'internaute dans le but d'obtenir des renseignements personnels en se faisant passer pour une société (banque, impôts,...) avec laquelle la victime est familier.

Rançongiciel : Logiciel qui empêche l'internaute d'avoir un quelconque accès à son PC ou à des fichiers personnels. Le pirate demandera ensuite à la victime une rançon pour débloquer l'accès.

Piratage de compte : Il s'agit d'aborder un outil informatique de manière illégale dans le but d'en dérober son contenu.

Faux ordre de virement : Escroquerie ayant pour but d'inciter la victime (par usurpation d'identité) à effectuer un virement de fonds sur un compte frauduleux.

Exfiltration de données confidentielles : Pratique consistant à extraire des informations à haute valeur d'une entreprise à des fins malveillantes (rançon ou vente). Ces attaques peuvent être menées par un employé ayant accès aux systèmes de la société ou par des acteurs malveillants externes.

Attaque par déni de service distribué : Attaque informatique visant à perturber, voire paralyser totalement, le fonctionnement d'un serveur informatique en le bombardant à outrance de requêtes erronées ou en exploitant une faille de sécurité afin de provoquer une panne ou un fonctionnement dégradé. La plupart des ces attaques se font à l'heure actuelle à partir de plusieurs sources.

En analysant un peu plus en profondeur ces différentes attaques, il y a trois conséquences pour une entreprise lorsque l'une d'elles se produit. La première est la perte et/ou le vol de données. Cette perte compromet la confidentialité d'une entreprise. La deuxième est l'interruption de l'activité qui affecte la disponibilité des systèmes et des informations. La dernière est la fraude qui est une

atteinte à l'intégrité de la société. L'intégrité, la disponibilité et la confidentialité sont des termes qui se trouvent dans les définitions du risque cyber énoncées lors de la première section. On remarque déjà ici un problème pour un assureur qui veut fournir un produit qui couvre le risque cyber : les dégâts d'une attaque sont souvent des dégâts qui ne sont pas matériels. Il est dès lors compliqué de tarifer un produit cyber, les effets de tels dégâts pouvant se manifester que bien plus tard.

1.3 Propriétés du risque cyber

Le risque cyber est un risque complexe. Il comporte plusieurs caractéristiques qui seront ici passées en revue. Comprendre et analyser les propriétés de ce risque est une étape primordiale pour l'assureur avant de commencer la modélisation. Il faudra ensuite incorporer un maximum de propriétés dans le modèle afin que celui-ci soit le plus réaliste possible.

Dépendance : Prenons l'exemple d'un portefeuille automobile, l'hypothése d'indépendance entre les polices est faite lors de la modélisation mais en réalité celle-ci n'est pas respectée. La dépendance étant très faible pour des très grands portefeuilles, l'hypothèse n'a pas de conséquence dans le modèle. La dépendance entre les polices est une des données principales du risque cyber. Pour illustrer cette caractéristique, il suffit de penser au virus WannaCry. Ce virus fut découvert en mai 2017 et a touché plus de 250 000 entreprises et infectait les programmes Microsoft. Toutes les entreprises utilisant ce système informatique en furent dès lors victimes. Ce logiciel rendait innaccessibles les données de l'ordinateur et revendiquait une somme entre 300 et 600 dollars pour que l'utilisateur puisse retrouver une situation normale. À cette période, les assureurs ont dû avoir une charge sinistre importante. Il s'agit certes d'une situation extrême mais illustrant bien que l'assureur devait posséder des provisions conséquentes pour remplir ses engagements. Le cyber risque présente un réel risque d'accumulation.

Manque de données : Trouver des données n'est pas chose aisée. Plusieurs facteurs peuvent expliquer cette pénurie de données. Premièrement, le risque cyber est assez récent. Il est donc logique que trouver des données soit moins facile que pour un portefeuille automobile qui est un produit existant depuis des dizaines d'années. Deuxièmement, le fait qu'une entreprise ait été sujette à une attaque n'est pas forcément valorisant pour son image. L'entreprise va donc préférer, dans la mesure du possible, rester discrète à ce sujet et ne pas rendre cette information publique.

Valeurs extrêmes : De manière générale, les dégâts des attaques sont des petits montants mais lorsque cela se passe mal, cela peut prendre de grandes proportions. Certaines attaques peuvent se chiffrer à des millions d'euros. Ce sont des sinistres qui sont difficiles à couvrir pour un assureur. La théorie des valeurs extrêmes est donc importante pour le cyber risque. Étudier ce qu'il se passe en queue de distribution est un passage obligé pour la modélisation d'un tel modèle.

Nombre d'employés de l'entreprise victime d'une attaque	Médiane	percentile 95
+1000	24000	462000
250-999	17000	382000
50-249	10000	119000
10-49	12000	285000
1-9	8000	308000

FIGURE 1.3.1 – La médiane et le percentile 95% du coût de sinistre en dollars pour les attaques cybers en 2021 par taille d'entreprise dans le rappport annuel d'Hiscox.

En 2021, dans son rapport annuel sur le risque cyber, Hiscox [3] a publié des chiffres concernant le coût des sinistres. Hiscox sonde plus de 5000 entreprises de taille, de secteur, de pays et de maturité informatique différents sur les évènements cybers subis par celles-ci. Ce rapport indique, notamment, par nombre d'employés de l'entreprise victime d'une attaque, la médiane et le percentile 95% du coût de sinistre. Les résultats de ce sondage sont repris dans le figure 1.3.1. Peu importe la taille de l'entreprise, la médiane est relativement basse alors que le percentile 95 est très éloigné de la médiane ce qui illustre le fait que la queue de la distribution du coût de sinistre est épaisse.

Silent Cyber : Le cyber risque étant en constante évolution, de nouveaux types d'attaques surgissent. Ces attaques peuvent endommager des polices alors que ce type d'attaques n'était ni considéré ni tarifié lors de la modélisation. Le risque cyber est également caché dans d'autres types de couverture. Par exemple, une assurance inondation peut contenir partiellement du risque cyber. Si un assureur assure des maisons le long d'un cours d'eau, il est soumis au risque que le barrage de ce cours d'eau se fasse pirater et provoque ainsi des dégâts sur l'immobilier situé aux alentours. [13]

Évaluation difficile : Le coût du sinistre peut être extrêmement compliqué à évaluer. Connaître le prix d'un ordinateur endommagé par une attaque est aisé mais comme déjà mentionné dans la section précédente, les dommages d'une attaque cyber ne sont pas uniquement matériels. Les questions suivantes méritent donc d'être posées : Comment évaluer la perte de données ? Comment évaluer un impact sur la réputation d'une entreprise ?

Asymétrie de l'information : L'asymétrie de l'information est de deux types en ce qui concerne le risque cyber. La première est qu'une compagnie aura plus tendance à s'assurer si elle a déja subi une attaque ou si elle sait que son système informatique est moins performant. De plus, un assuré aura tendance à faire moins attention et à moins investir dans la sécurité informatique de son entreprise. [5] Les assureurs évaluent la qualité de la gestion du risque cyber auprès de l'assuré avant d'accepter de souscrire la police mais ce n'est pas évident de savoir pour l'assureur

3. "Hiscox Cyber Readiness Report 2021" - Hiscox : https://www.hiscoxgroup.com/sites/group/files/documents/2021-04/Hiscox%20Cyber%20Readiness%20Report%202021.pdf

à quel point une entreprise est développée technologiquement et à quel niveau les employés d'une société sont formés contre les attaques cyber. [3] Deuxièmement, les escrocs doivent trouver un unique moyen d'utiliser une faille tandis que les victimes ciblées doivent pouvoir contrer tout type d'attaque. Les armes ne sont pas égales puisqu'une partie a plus d'information que l'autre. [14]

Risque de changement : Le monde de la digitalisation est un milieu qui évolue constamment au cours du temps et qui n'est pas stable. La fréquence des attaques augmente avec le temps selon le rapport annuel d'Hiscox de 2021. L'utilisation d'anciennes données peut conduire à un mauvais modèle puisque ces données ne reflètent plus la réalité. De surcroît, un modèle correct peut très vite devenir obsolète. Idéalement, la modélisation du risque cyber devra être dynamique et prendre en considération cette évolution constante du risque.

Aléa moral : Comme beaucoup de couvertures d'assurance, le risque cyber comporte un danger d'aléa moral. Une société qui est assurée peut avoir tendance à être moins vigilante. Pour éviter ce souci, l'assureur mettra en place une franchise, effectuera des audits à intervalles réguliers ou fournira des formations. [3]

Le cyber risque possède une série de caractéristiques qui fait que ce risque est assez complexe. L'objectif de ce mémoire consiste à modéliser ce risque en prenant en compte un maximum de ces caractéristiques.

1.4 Police d'assurance

Cette partie sera consacrée à la présentation d'un exemple de police d'assurance. Bien que l'assurance qui protège contre le risque cyber soit encore peu développée, il existe malgré tout quelques assureurs qui fournissent ce produit. Une police d'assurance AXA [4] sera passée en revue afin de voir quelles sont les limites des dédommagements, les dommages couverts et les exclusions.

Pour commencer, une couverture est divisée en deux parties. L'assureur peut se limiter au first party coverage mais peut également faire ce qui s'appelle du third party coverage.

Définition 1.4.1. Une couverture *First Party* est une couverture qui couvre les dommages subis par la personne dont le nom est indiqué sur le contrat.

Définition 1.4.2. Une couverture *Third Party* est une couverture qui consiste à protéger l'assuré des dommages qu'il a lui-même causé à un tiers.

La notion de risque cyber est si large qu'il est important pour un assureur d'être le plus précis possible lorsqu'il s'agit de rédiger un contrat pour un assuré. Bien définir le périmètre de couver-

4. Assurance Cyber-Document d'information sur le produit d'assurance - AXA : https://cdn-prd-axa.azureedge.net/fr-be/-/media/axa-documents/library-public/ipid/7/2/b/7/8/4186149/Cyber%20protection%20-%20fiche%20IPID%20-%2020220613.pdf?rev=57e2cd508994459b8990f26eaa6261a7

ture est un enjeu important. Le risque cyber peut aller du simple chargeur cassé par un employé de manière non-intentionnelle jusqu'à l'arrêt total de l'activité suite à un hacking d'une personne malveillante. L'assureur ne couvrira sans doute pas les plus petits dégâts et vu la volatilité et l'hétéréogénité du risque, il y aura certaines limites dans les couvertures. Les clauses d'un contrat vont dépendre du type d'attaque que le contrat couvre et du secteur d'activité de la police. Les facteurs de risque pour les risques avec des grandes queues de distribution ne sont pas les mêmes que les facteurs de risques qui augmentent la valeur moyenne de la partie centrale de la distribution. Ceci sera développé dans le chapitre quatre de ce travail. Pour les distributions "heavy tail", il faudra définir un seuil, une limite de couverture. Sans disposer d'un seuil d'intervention, ce sont des risques qui sont en réalité innassurables car l'assureur se retrouve dans une situation où il est difficile pour lui de gérer ses engagements.

Exemple 1.4.3. Sur le site internet de l'assureur AXA, il est possible de trouver un document décrivant le produit d'assurance risque cyber. Ce document décrit ce qui est couvert, ce qui ne l'est pas ainsi que les phases d'exclusion.

Ce qui est couvert ?

- AXA interviendra lorsque l'attaque porte atteinte aux données, aux programmes et à la protection des données à la suite d'un acte malveillant, une erreur humaine, une panne ou un dysfonctionnement frappant du système informatique.
- L'assureur couvrira également la perte du chiffre d'affaires pour les ventes en ligne si ces dernières sont indisponibles. On remarque que lorsqu'il y a une interruption de l'activité, seulement une partie du chiffre d'affaire sera couverte. L'assureur ne précise cependant pas comment une interruption d'activité est évaluée. Une possibilité est de demander les détails des ventes annuelles sur internet pour ensuite faire un prorata entre le temps de l'interruption et le chiffre des ventes annuelles.
- De plus, la responsabilité civile de l'assuré suite à la perte, la divulgagtion ou l'indisponibilité de documents ou de données sera couverte si la cause est une défaillance technique des systèmes informatiques, une erreur humaine dans la gestion de ces systèmes ou un acte de malveillance.
- Si un évènement cyber porte atteinte à la e-réputation de l'entreprise, AXA couvrira le recours civil extra-contractuel et le noyage et nettoyage de l'information.

Ce qui n'est pas couvert ?

- Pour commmencer, les guerres et actions d'inspiration collective ne seront pas couvertes. C'est une pratique générale en assurance d'exclure le risque de guerre, mais cette exclusion devient plus difficile à définir pour le risque cyber.
- Bien entendu, les actes intentionnels ne font pas partie de la couverture. Si l'attaque est liée à la non mise à jour, à l'absence d'antivirus ou à une défaillance connue, il n'y aura pas de couverture. Les frais pour améliorer le système informatique ne sont pas couverts, c'est assez intuitif. L'assureur ne va pas couvrir cela car il y aurait de l'aléa moral. Une suite de sinistres identiques successifs ne sera pas dédommagée.

— Les amendes à caractère punitif et le dysfonctionnement d'un service informatique externe à l'entreprise ne font pas partie de la couverture. La partie responsabilité civile du contrat ne couvre pas la concurrence déloyale, les atteintes à des droits intellectuels ou l'inéxecution des engagements contractuels.

— Il n'y aura pas d'assistance lorsque le sinistre qui porte atteinte à la réputation concerne un support de communication différent d'internet, résulte d'une action collective qui a été constituée par l'assuré lui-même.

Exclusions

— Les dégâts inférieurs à la franchise.

— Le montant des sinistres qui est au-dessus de la limite d'indemnisation. Il n'y a cependant pas d'information concernant la limite de l'indemnisation.

— Il y a exclusion lorsque l'assuré ne respecte pas les mesures de prévention (installation des mises à jour, antivirus,...).

1.4.1 Entretien avec Caroline Hillairet sur le risque cyber

Selon une actuaire spécialisée dans le risque cyber, Caroline Hillairet, la couverture d'une assurance cyber n'est pas juste une compensation financière, cela va plus loin que cela en ce sens où un dédommagement pécunier sera fait mais l'assureur couplera cela avec un service d'assistance (exemple de Manutan dans l'introduction de ce mémoire). Il mettra ses assurés en relation avec des entreprises spécialisées en cybersécurité afin de les conseiller par rapport à ce risque. En plus du côté couverture et assistance, l'assureur aura à coeur d'avoir également un rôle de prévention car la maîtrise de ce risque passe également par là.

Chapitre 2

Modèle relatif à la fréquence de sinistre

Ce chapitre est consacré à l'étude de la fréquence de sinistre du risque cyber. Comme expliqué dans le chapitre précédent, le risque cyber est un risque à part. Le manque de données lié aux attaques informatiques rend la modélisation compliquée. Au moment d'écrire ces lignes, la seule base de données publique concernant des évènements cyber est la base de données PRC [1]. Il existe d'autres sources, mais qui sont développées pour des professionels et sont par conséquent coûteuses. C'est une base de données qui rapporte les vols de données d'entreprises aux Etats-Unis entre 2006 et 2019. Cependant, cette base de données ne semble pas propice à la modélisation de la fréquence. En effet, elle rapporte des évènements mais il n'y a aucune notion d'exposition. Ni le nombre d'entreprises potentiellement concernées ni le temps durant lequel les entreprises furent sujettes à ce risque étant connu, il n'y a pas moyen de pouvoir gérer l'exposition. Dans un premier temps, une analyse decriptive sera réalisée avec des données récoltées sur le site internet de l'assureur Hiscox. Une caractéristique importante du risque cyber est qu'il présente un risque d'accumulation. Il peut y avoir à certains moments du temps un pic de sinistres par la nature du risque cyber. Un assureur qui possède un portefeuille pour un produit de type cyber peut s'attendre, à un moment, à avoir un grand nombre de sinistres sur une période de temps assez courte. L'assureur devra tenir à ses engagements bien entendu et devra donc prendre en compte ce risque d'accumulation. Un modèle permettant de gérer ce risque d'accumulation sera étudié lors de ce chapitre : le processus de Hawkes. C'est un processus qui prend en compte l'auto-excitation du sujet étudié.

2.1 Analyse préliminaire

Lors de cette section, une étude descriptive de données récoltées sur le site d'Hiscox est faite. Cette base de données comportre 5181 lignes qui représentent chacune une entreprise. Il y a cinq variables pour chaque entreprise.

— La première ,celle qui nous intéresse le plus, indique si l'entreprise a subi une cyber attaque

1. Privacy Clearing Clearinghouse : https://privacyrights.org/data-breaches

	Petite	Moyenne	Grande	Très Grande	Enorme
Nombre d'employés	1-9	10-49	50-250	250-999	+1000

FIGURE 2.1.1 – Description de la variable taille pour les données d'Hiscox.

ou pas lors de l'année 2021. Notons que certaines entreprises ont répondu "Je ne sais pas" à cette question. Dans ce cas, il est supposé qu'elles n'en n'ont pas subis car cela signifie qu'il n'y a pas eu de gros dégâts.

— Une autre variable est le pays dans lequel l'entreprise se situe (USA, Angleterre, Belgique, Pays-Bas, Allemagne, France, Espagne et Irlande.

— La base de données comprend également le secteur d'activité de chaque entreprise. On dénombre quatorze secteurs d'activité différents.

— Hiscox fournit aussi la taille de l'entreprise, qui est répertorié en cinq catégories. Le détail de cette variable est résumé sur la figure 2.1.1.

— Enfin, la dernière variable est une variable catégorielle avec trois niveaux qui représente la maturité informatique de l'entreprise. Hiscox dispose d'une échelle pour mesurer la maturité informatique d'une entreprise. Cette échelle est un score qui varie entre zéro et cinq. Lorsque le score est plus grand que quatre, la société est qualifiée d'experte cyber. Une entreprise de niveau intermédiaire possède un score entre 2.5 et 4 alors qu'une entreprise qui est novice d'un point de vue maturité informatique obtient un score inférieur à 2.5.

Avant de commencer la modélisation du modèle, une petite analyse descriptive est réalisée afin de mieux appréhender les données. Cela permet de mettre en évidence les facteurs de risque potentiels et de dégager des premières tendances. Aucun test statistique ne sera réalisé pour souligner le caractère significatif ou pas. Dans la suite de ce mémoire, ces données ne seront plus utilisées pour les modèles développés.

Pour la variable taille, il semble y avoir un lien entre la taille de l'entreprise et la probabilité que l'entreprise soit victime d'une attaque. Plus une entreprise est grande, plus l'entreprise a de risques d'avoir au moins un sinistre endéans l'année. La figure 2.1.2 indique que 62% des entreprises énormes ont subi minimum une attaque tandis qu'on tombe à 32.6% pour les petites entreprises. La différence est significative, elle double d'un cas à l'autre.

Lorsqu'il s'agit du pays d'activité de l'entreprise, les différences sont moins fortes en comparaison avec la variable "taille". Le pays le plus exposé semble être les Pays-Bas avec 57.1% des entreprises touchées. Le meilleur élève est l'Angleterre avec un pourcentage égal à 41.9%. Les détails de cette variable "pays" sont repris sur la figure 2.1.3.

Pour la variable secteur d'activité, le secteur des services financiers semble être la cible principale des hackeurs puisqu'il a une probabilité de 58.7% d'être touché. A l'inverse, le secteur des services professionels est le moins victime d'attaques avec seulement une probabilité annuelle de 38.1%. La probabilité pour chaque secteur d'activité se trouve sur la figure 2.1.4.

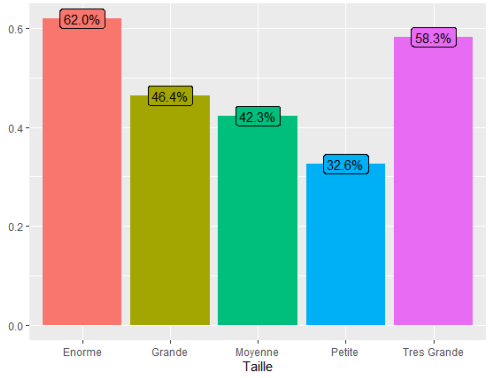

FIGURE 2.1.2 – Pourcentage des entreprises ayant subi au moins une attaque cyber lors de l'année 2021 en fonction de la taille de l'entreprise.

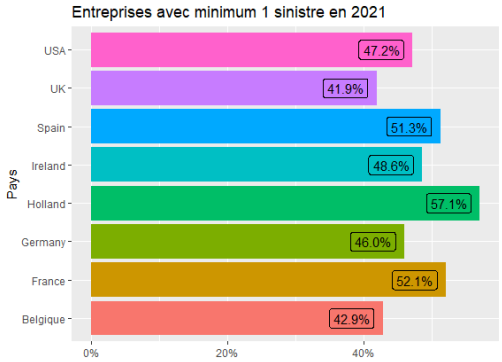

FIGURE 2.1.3 – Pourcentage des entreprises ayant subi au moins une attaque cyber lors de l'année 2021 en fonction du pays dans lequel la société se situe.

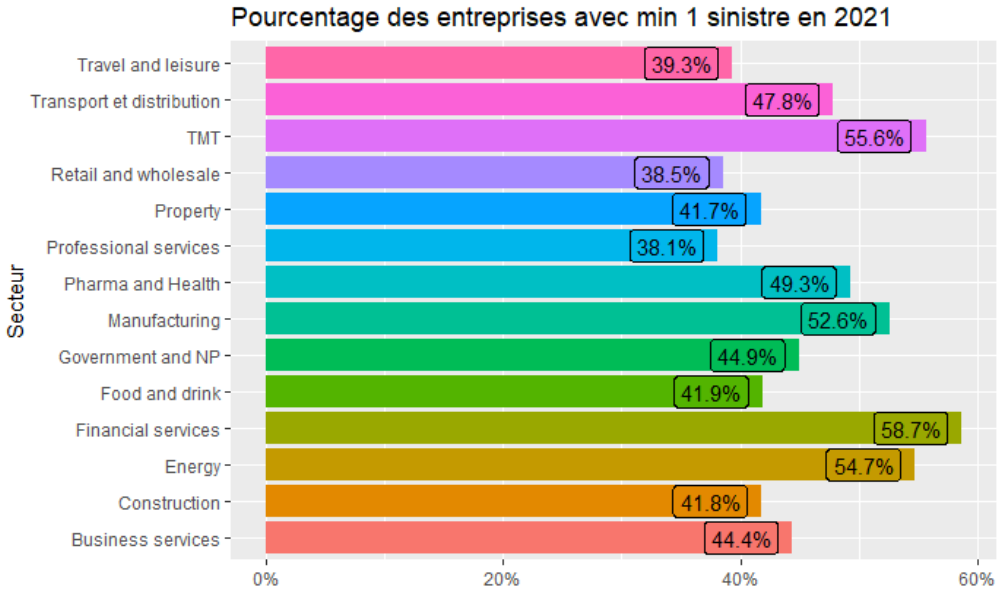

FIGURE 2.1.4 – Pourcentage des entreprises ayant subi au moins une attaque cyber lors de l'année 2021 en fonction du secteur d'activité de l'entreprise.

Jusqu'ici, les résultats semblent assez intuitifs. Les plus grandes entreprises sont davantage touchées ce qui semble logique puisqu'elles possèdent plus d'outils informatiques (donc plus de failles) et ont des ressources plus importantes. Le fait que le secteur des services financiers et le secteur de la télécommunication soient plus ciblés que les autres n'est pas surprenant. Cependant,

le résultat qui va suivre est plus surprenant. En effet, on aurait tendance à dire que les entreprises ayant la plus grande maturité informatique sont celles qui ont une probabilité plus faible par rapport aux entreprises avec un système informatique novice. Ce n'est pourtant pas ce que les données traduisent. Comme illustré sur la figure 2.1.5, les entreprises expertes informatiquement semblent les plus touchées par des attaques du type cyber. En effet, elles possèdent une probabilité de 61% alors que les entreprises avec un système informatique novice ont une probabilité de seulement 36.3%. Les entreprises avec un niveau intermédiaire ont, quant à elles une probabilité de 51.6%. Comment expliquer ce résultat qui semble contre-intuitif ?

— Premièrement, il n'y a pas de corrélation entre la taille de l'entreprise et le niveau de maturité informatique comme illustré sur la figure 2.1.6. Il n'y pas de différence significative entre les barplots. En effet, il n'y pas une plus grande proportion d'entreprises expertes parmi les entreprises énormes que parmi les petites sociétés.

— Il se pourrait que, puisqu'elles sont plus développées d'un point de vue informatique, la moindre faille dans le système informatique soit détectée.

— La différence se fait sans doute dans le coût du sinistre qui est peut-être plus bas pour les entreprises à maturité informatique élevée.

Voici des variables intéressantes qui semblent être cohérentes à introduire dans un modèle pour le cyber risque. Cependant, étant donné le nombre limité de données, elles ne permettent pas de développer un modèle poussé.

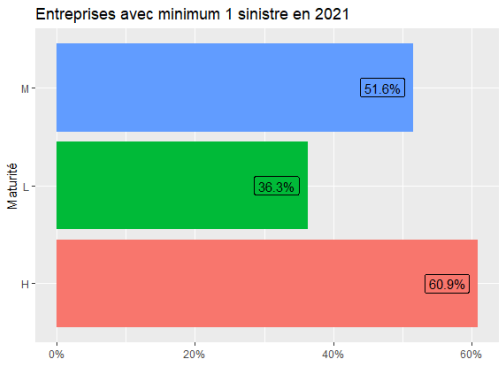

FIGURE 2.1.5 – Pourcentage des entreprises ayant subi au moins une attaque cyber lors de l'année 2021 en fonction de la maturité informatique.

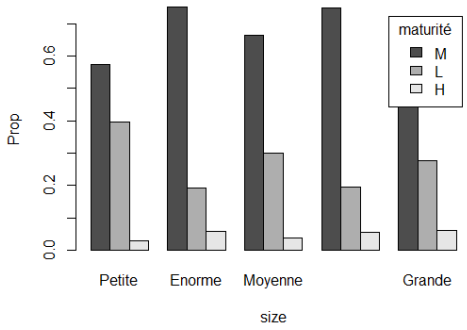

FIGURE 2.1.6 – Proportion des entreprises en fonction de la maturité informatique par la taille d'entreprise.

2.1.1 Entretien avec une personne travaillant dans la tarification du produit d'assurance cyber

Etant à la fois interpellé par les résultats précédents et à la recherche de données sur le risque cyber en assurance, nous avons eu un entretien avec Gordon Clark, employé chez Hiscox dans

le département risque cyber en Angleterre. L'expert n'a malheureusement pas pu partager les données sur lesquelles Hiscox se base pour tarifer leur produit cyber. Cependant, il a pu fournir plus de détails sur les données et commenter certains résultats.

Selon lui, les résultats sont assez cohérents par rapport à la réalité pour les variables "taille" et "secteur". Comme pressenti, le fait que les entreprises les plus matures informatiquement soient plus touchées que les entreprises avec un système informatique novice est contraire à ce qui est observé dans la réalité. L'expert ne savait pas vraiment donner une véritable raison quant à ce résultat. Il ajoute cependant qu'en réalité, une variable indiquant la maturité informatique d'une entreprise est utilisée uniquement pour les entreprises de petite taille.

La discussion s'est poursuivie de manière plus générale sur la tarification du produit cyber. La procédure pour la tarification du produit cyber en assurance va dépendre de la taille de l'entreprise. Il est possible de faire un modèle général pour ce produit pour les entreprises de petite taille mais en réalité, pour les entreprises de grande taille, la tarification se fait de manière individuelle par expert judgement. Selon lui, pour un assureur, la variable la plus importante à incorporer dans un modèle est celle qui indique les revenus de l'entreprise. C'est un produit qui n'est pas encore bien maîtrisé par les assureurs car les frais sont compliqués à estimer. Selon lui, il faut entre trois et six ans pour connaître le vrai coût d'une attaque. Il ajoute que ce sont les rançongiciels qui ont le plus gros impact sur le loss ratio de l'assureur pour le produit cyber. Pour l'interruption de business, un montant journalier est fixé à la conception du contrat par jour où l'activité est interrompue avec un nombre maximum de jours.

2.2 Modèle de Hawkes

Le modèle de fréquence du risque cyber aura pour principal objectif de prendre en considération la dépendance entre les différentes polices d'un portefeuille. Ce qui intéresse l'assureur est de pouvoir modéliser le plus précisément possible le nombre de sinistres que son portefeuille d'assurés aura subi. Le processus de Hawkes est une manière d'incorporer l'aspect "contagion" propre au risque cyber. Il existe d'autres méthodes pour inclure la dépendance entre les différentes polices notamment via un modèle épidéomologique comme développé dans l'article [15]. Ce dernier semble moins propice à la tarification pour deux raisons :

— Les paramètres sont compliqués à calibrer en ce sens où il y a souvent plus d'inconnues que de paramètres.

— Le modèle suppose que lorsqu'une entreprise est attaquée, elle est nécessairement infectée pendant une certaine période or il est probable qu'une entreprise soit touchée sans pour autant avoir un virus introduit dans leur système pour une certaine période. Par exemple, un vol de données.

C'est donc pourquoi ce mémoire étudiera le processus de Hawkes comme dans l'article [10].

2.2.1 Présentation du modèle

L'assureur souhaite prédire le nombre de sinistres qu'il devra couvrir sur une période donnée. La variable qu'il souhaite donc modéliser est une variable de comptage : il veut connaître le nombre d'attaques cyber subi par son portefeuille.

Définition 2.2.1. Un processus de comptage est un processus spécifié par une séquence de points $\{T_1, T_2, \cdots, T_N\}$ sur $\mathbb{R}^+$ tel que pour tout $t \in \mathbb{R}^+$,

$$N_t = \sum_{n=1}^{N} 1_{T_n \leqslant t}$$

représente le nombre d'évènements avant t. La distribution de N_t est caractérisée par une intensité $\lambda(t, \omega)$ également appelée taux de hasard et définie comme

$$\lim_{h \to 0} \frac{P(N_t - N_{t+h} > 0 | \mathcal{F}_t)}{h} = \lambda(t, \omega) \mathrm{d}t.$$

où $\lambda(t, \omega)$ peut être constante, déterministe ou stochastique.

Le processus de comptage dans le cas du risque cyber va donc être le nombre d'évènements cyber dans un portefeuille. N représentera le nombre d'attaques cyber et $(T_i)_{i=1,\cdots,N}$ représente l'instant auquel la ième attaque s'est produite. Trois processus seront définis dans cette section. Ce sont des processus de Poisson qui sont utiles lorsqu'un actuaire souhaite modéliser une fréquence. Cependant, certains ne seront pas adéquats au risque cyber car les écarts entre chaque attaque sont indépendants les uns des autres.

Définition 2.2.2. Un processus de Poisson homogène est un processus de comptage N_t avec une intensité λ constante caractérisé par des écarts entre deux arrivées $(T_n - T_{n-1})$ indépendants et identiquement distribués avec une distribution exponentielle de paramètre $\lambda \in \mathbb{R}$.

$$P(T_n - T_{n-1} > t) = exp(-\lambda t).$$

Exemple 2.2.3. La figure 2.2.1 illustre la simulation de cinq processus de Poisson homogènes avec différentes intensités sur l'intervalle $[0, 10]$. La simulation d'un processus de Poisson homogène se fait facilement étant donné que les temps entre chaque arrivée sont indépendants et identiquement distribués selon une loi exponentielle de paramètre $\lambda \in \mathbb{R}$. En pratique, pour simuler un processus de Poisson homogène, il suffit donc de simuler des lois exponentielles avec un paramètre λ correspondant. Chaque simulation correspondra donc à l'écart entre chacune des attaques.

Le but étant de trouver un modèle qui prend en compte cet effet d'accumulation (aussi appelé auto-excitation) du risque cyber, le processus de Poisson homogène n'est pas intéressant à développer. Le fait que l'intensité soit constante signifie que la fréquence des sinistres est stable dans le temps. On souhaiterait que l'intensité évolue avec le temps de telle sorte que l'intensité soit ,par moment, plus grande qu'à d'autres. Lorsque l'intensité est une fonction déterministe en fonction du temps,

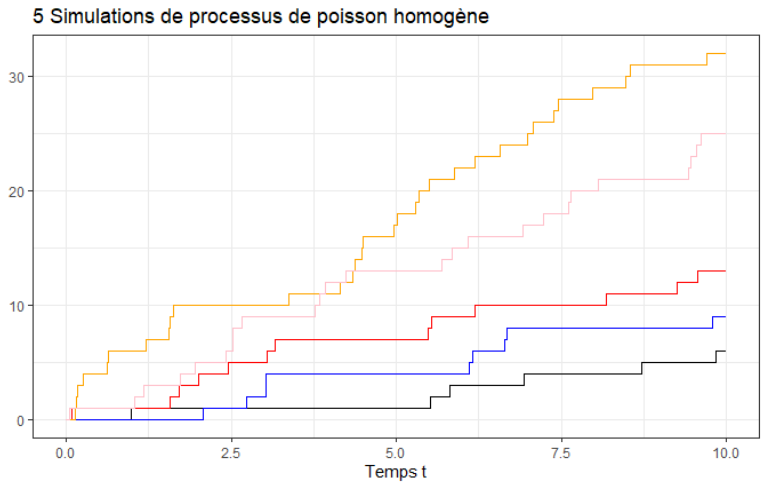

FIGURE 2.2.1 – Simulations de processus de Poisson homogènes avec différentes intensités : 0.5 (noir), 1 (bleu), 1.5 (rouge), 2(orange) et 2.5 (rose).

le processus est appelé un processus de Poisson non-homogène.

Définition 2.2.4. Un processus de poisson non-homogène est un processus de comptage N_t avec une intensité $\lambda(t)$ déterministe caractérisé par des écarts entre deux arrivées $(T_n - T_{n-1})$ indépendants et identiquement distribués selon une distribution exponentielle telle que

$$P(T_n - T_{n-1} > t|\mathcal{F}_{n-1}) = exp(-\int_{T_{n-1}}^{T_n} \lambda(u)\mathrm{d}u).$$

Exemple 2.2.5. Il est possible de simuler un processus de Poisson non-homogène. Supposons que l'intensité soit définie au cours du temps par la fonction $\lambda(t) = 2 + sin(t)$ sur $[0, 10]$ illustrée sur le graphique 2.2.2. Cinq simulations sont illustrées sur la figure 2.2.3. L'intensité atteint un pic lorsque $t \approx 2$ et $t \approx 8$, c'est également à ce moment-là qu'il y a un effet de clustering sur les simulations puisque qu'il y a un plus grand nombre de sauts à ces moments-là. A l'inverse, il y a moins de sauts en $t \approx 5$ puisque l'intensité est plus faible à ce moment là du temps. La simulation d'un processus de Poisson non-homogène d'intensité $\lambda(t)$ est réalisée à l'aide de l'algorithme d'amincissement. Le détail de ce dernier se trouve sur l'algorithm 1.

A l'instar du processus de Poisson homogène, c'est un processus qui ne convient pas dans le contexte qui est étudié puisque l'écart entre chaque attaque est indépendant. Or, le risque cyber comprend un risque d'accumulation c'est-à-dire que lorsqu'un évènement a lieu, la probabilité qu'il y ait une nouvelle attaque dans le portefeuille doit augmenter pour un avenir proche du moins. L'idéal serait donc que l'intensité λ soit stochastique.

Algorithm 1 Algorithme pour simuler un processus de Poisson inhomogène entre [0,T]

Require: $T \geqslant 0 \ \ \& \ \ \lambda(t) \geqslant 0 \ \forall t$ ▷ Il faut que $\lambda(t)$ soit toujours positif
  sauts $\leftarrow$ [] ▷ On initialise la liste dans laquelle on stockera le temps des attaques
  $\bar{\lambda} \leftarrow \sup \lambda(t)$ ▷ On trouve un supremum de $\lambda(t)$
  $t \leftarrow 0$
  while $t \leqslant T$ **do** ▷ On continue tant qu'on n'atteint pas T
    $u \leftarrow unif(1)$
    $w \leftarrow -\frac{\ln(u)}{\bar{\lambda}}$ ▷ Simulation d'une exponentielle de paramètre $\bar{\lambda}$
    $t \leftarrow t + w$ ▷ saut potentiel du processus
    $D \leftarrow unif(1)$
    if $D \leqslant \frac{\lambda(t)}{\bar{\lambda}}$ **then**
      sauts.*append*(t) ▷ Le saut est ajouté à la liste avec une probabilité $\frac{\lambda(t)}{\bar{\lambda}}$
    end if
  end while

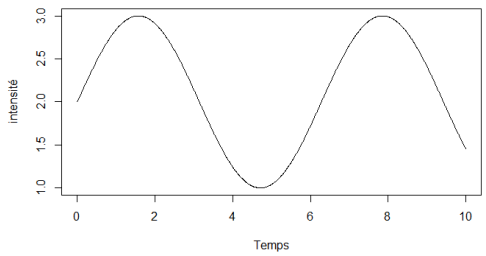

FIGURE 2.2.2 – Evolution au cours du temps de $\lambda(t) = 2 + sin(t)$.

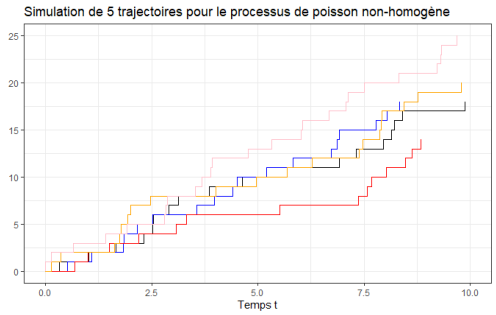

FIGURE 2.2.3 – Cinq simulations d'un processus de Poisson non-homogène avec $\lambda(t) = 2+cos(t)$.

Définition 2.2.6. Un processus de Cox est un processus de comptage avec une intensité stochastique $\lambda(t, \omega)$ et qui est $\mathcal{F}$-adaptée tel que

$$P(T_n - T_{n-1} > t | T_{n-1} = s, \mathcal{F}_s) = exp(-\int_s^{s+t} \lambda(u, \omega) \mathrm{d}u).$$

Modéliser la fréquence d'un portefeuille avec un processus de Cox n'est pas aisé. C'est le processus qui correspond idéalement le mieux à la modélisation du risque cyber. Il est cependant très compliqué, voire impossible de pouvoir calibrer l'intensité à partir du peu de données qui sont disponibles. Cette intensité doit en effet être stochastique. Bien identifier la dynamique de l'intensité stochastique est difficile. Pour modéliser, la fréquence sinistres du risque cyber, il existe un processus qui se situe entre le processus de Poisson homogène et le processus de Cox : appelé le processus de Hawkes.

Définition 2.2.7. Le processus de Hawkes est un processus défini par une intensité stochastique,

qui est entièrement spécifié par le processus de comptage lui même

$$\lambda(t) = \mu(t) + \sum_{T_n<t} \phi(t - T_n), \tag{2.2.1}$$

où ϕ est le noyau d'excitation qui est une fonction déterministe, continue et positive et μ est l'intensité de base qui est une fonction déterministe, positive et continue.

La première partie de l'équation (2.2.1) représente l'intensité de base. En effet, la fonction $\mu(t)$ représente le taux d'un saut spontané. La somme représente l'impact des évènements passés et capture le risque d'accumulation du risque cyber.

2.2.2 Exemples de noyaux pour le processus de Hawkes

Dans le cadre de ce travail, l'estimation de l'intensité $\lambda(t)$ sera paramétrique. Il faudra faire l'hypothèse sur la forme du noyau. L'estimation des paramètres se fera ensuite via une maximisation de la fonction de vraisemblance. Cette section présente quelques exemples de noyaux.

Noyau exponentiel

Le noyau exponentiel ϕ_{exp} est défini sur $[0, +\infty[$ comme

$$\phi_{exp}(t) = \alpha exp(-\beta t)$$

où α et β sont des constantes positives.

Noyau d'Erlang

Le noyau d'Erlang ϕ_{erl} est défini sur $[0, +\infty[$ comme

$$\phi_{erl}(t) = \alpha t \ exp(-\beta t)$$

où α et β sont des constantes positives.

Noyau Power Law

Le noyau Power Law ϕ_{pow} est défini sur $[0, +\infty[$ comme

$$\phi_{pow}(t) = \frac{K}{(t + c)^p}$$

où K , p et c sont des constantes positives. C'est un noyau de forme identique au noyau exponentiel, il modélise cependant moins vite la décroissance.

Noyau de Rayleigh

Le noyau de Rayleigh ϕ_{ray} est défini sur $[0, +\infty[$ comme

$$\phi_{ray}(t) = \alpha t \; exp(-\beta t^2)$$

où α et β sont des constantes positives. C'est un noyau avec une décroissance plus rapide que le noyau d'Erlang.

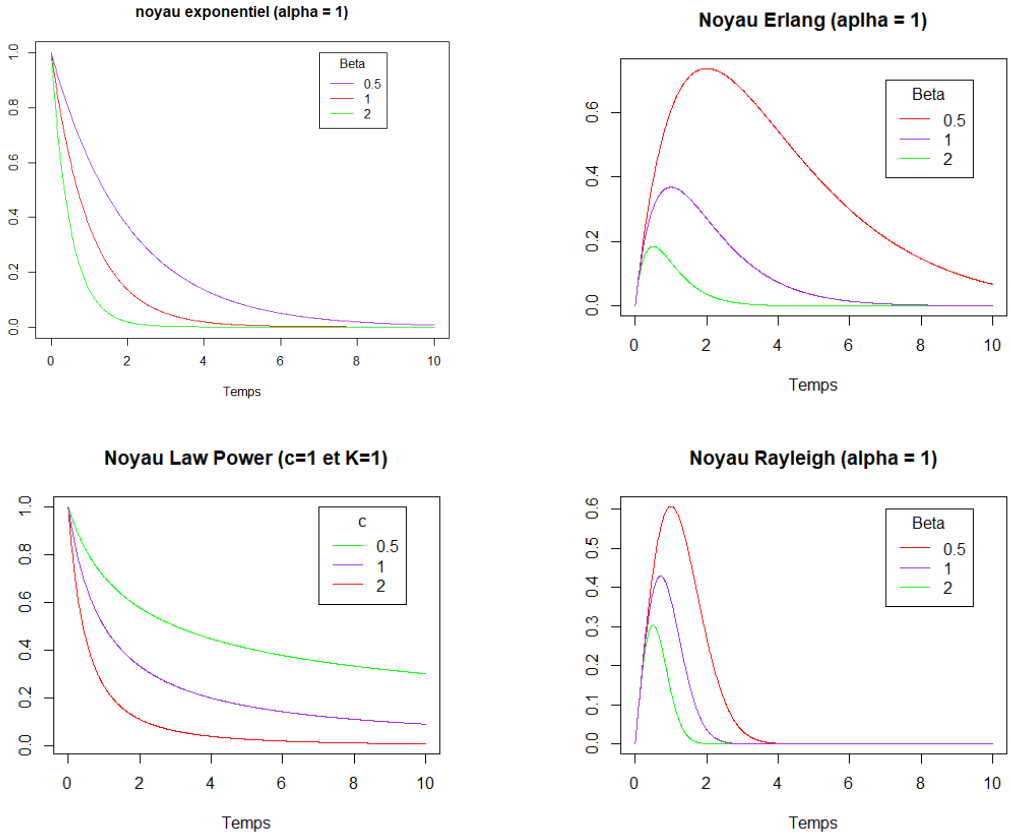

FIGURE 2.2.4 – Représentation des quatre types de noyaux pour différentes valeurs des constantes.

2.2.3 Simulation de processus de Hawkes

Au cours de cette section, plusieurs processus de Hawkes seront simulés. Pour chaque processus, l'intensité au cours du temps sera analysée. Il est intéressant d'observer en détail comment elle évolue puisque c'est ce qui différencie le processus de Hawkes d'un processus de Poisson homogène ou un processus de Poisson non-homogène. Dans un premier temps, l'intensité de base

μ dans le processus est supposée constante. En d'autres mots, l'intensité du processus de Hawkes est définie par

$$\lambda(t) = \mu + \sum_{T_n < t} \phi(t - T_n),$$

avec μ une constante.

Simuler un processus de Hawkes est différent de simuler un processus de Poisson non-homogène. Il y a effectivement de la dépendance entre chaque saut. L'algorithme pour simuler le processus de Hawkes pour un intervalle de temps $[0, T]$ va donc changer mais l'idée est la même dans le fond.

Algorithm 2 [11] Algorithme pour simuler un processus de Hawkes entre [0,T]

Require: $T \geqslant 0, \mu(t) \geqslant 0 \;\&\; \phi(t) \geqslant 0 \; \forall t$ ▷ Il faut que $\lambda(t)$ soit toujours positif
  sauts $\leftarrow$ [] ▷ On initialise la liste dans laquelle on stockera le temps des attaques
  $\bar{\lambda} \leftarrow \mu$ ▷ En 0, l'intensité maximum est égale à μ
  $u \leftarrow unif(1)$
  $t \leftarrow -\frac{\ln(u)}{\bar{\lambda}}$ ▷ Simulation du moment du premier saut
  while $t \leqslant T$ **do** ▷ On continue tant qu'on n'atteint pas T
    $\bar{\lambda} \leftarrow \sup \lambda(s)$ ▷ On trouve un supremum de $\lambda(s)$, le supremum est donc modifié après chaque ajout d'une attaque dans la liste
    $u \leftarrow unif(1)$
    $w \leftarrow -\frac{\ln(u)}{\bar{\lambda}}$ ▷ Simulation d'une exponentielle de paramètre $\bar{\lambda}$
    $t \leftarrow t + w$ ▷ saut potentiel du processus
    $D \leftarrow unif(1)$
    if $D \leqslant \frac{\lambda(t)}{\bar{\lambda}}$ **then**
      sauts.*append*(t) ▷ Le saut est ajouté à la liste avec une probabilité $\frac{\lambda(t)}{\bar{\lambda}}$
    end if
  end while

Afin de réaliser l'algorithme de simulation, il faut mettre à jour $\bar{\lambda}$ après chaque attaque. Il faut donc pour, chaque noyau ϕ, trouver l'argument qui maximise le noyau

$$argmax\ \phi(t)$$

Pour le noyau exponentiel, c'est trivial (que cela soit graphiquement ou numériquement) de trouver l'argument qui maximise la fonction ϕ_{exp}

$$argmax\ \phi_{exp}(t) = 0.$$

C'est-à-dire que lorsqu'une attaque a lieu, l'intensité augmente pour atteindre directement un pic. De manière grossière, le taux de contagion d'une attaque atteint son maximum au moment où l'attaque a eu lieu. Sur la figure 2.2.5, une simulation du processus de Hawkes avec un noyau exponentiel est illustré à gauche. Le graphique de droite reprend l'évolution de l'intensité de ce

processus au cours du temps. Il y a certaines périodes du temps où des clusters sont observables. Il y a entre $t = 8$ et $t = 10$ un grand nombre d'attaques pour un intervalle qui est assez court. L'intensité croît également rapidement sur cet intervalle. De plus, une attaque se produit en $t \approx 5$ et ensuite, aucune attaque n'a lieu jusqu'en $t \approx 6.5$. Sur le graphique qui décrit l'intensité, celle-ci saute en $t \approx 5$ pour ensuite décroitre jusqu'en $t \approx 6.5$. La forme de la décroissance est bien identique que sur la figure 2.2.4 pour le noyau exponentiel.

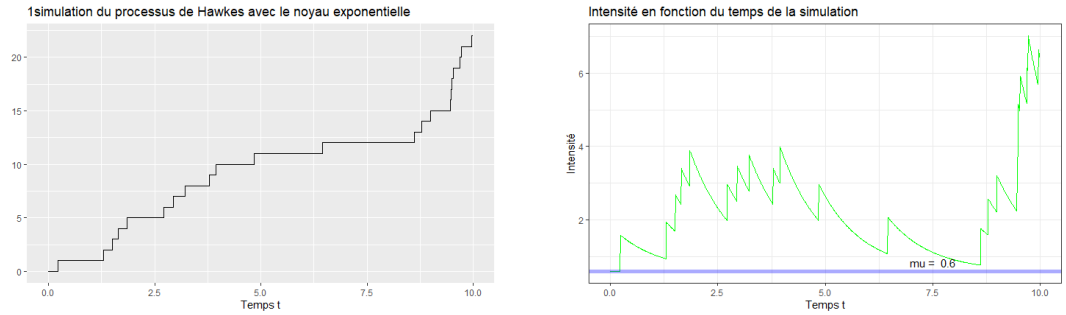

FIGURE 2.2.5 – La simulation d'un processus de Hawkes (noyau exponentiel) avec $\alpha = 1$, $\beta = 1$ et $\mu = 0.6$ (gauche) et l'évolution de l'intensité de ce processus au cours du temps (droite).

Il est possible de simuler un grand nombre de processus et ensuite de regarder le nombre moyen de sauts. Pour 5000 simulations, le nombre moyen d'attaques par évènement est égal à 24.074 . Il est possible de modifier les valeurs des paramètres et ainsi observer les tendances. Logiquement, une augmentation de μ ou α augmenterait le nombre d'attaques. A l'inverse une augmentation de β aura une bonne influence sur le nombre d'attaques puisque le nombre moyen d'attaques va diminuer.

Dans le cas du noyau d'Erlang, il faut également trouver l'argument qui maximise ϕ_{erl}. Ceci est nécessaire pour pouvoir exécuter l'algorithme explicité plus haut. Il est facile de trouver que

$$argmax\ \phi_{erl}(t) = \frac{1}{\beta}.$$

Cela signifie que lorsqu'une attaque a lieu, l'impact de l'attaque sur l'intensité n'est pas direct. Il augmente jusqu'en $t = 1/\beta$ pour ensuite diminuer comme illustré sur la figure 2.2.4. L'impact de l'attaque sur le reste des entreprises atteindra son pic plus tard alors que, dans le cas du noyau exponentiel, l'impact était direct. De manière informelle, le taux de contagion d'une attaque sur les autres entreprises est maximal $1/\beta$ jours après le moment de l'attaque. Sur la figure 2.2.6, une simulation du processus de Hawkes avec un noyau d'Erlang est illustrée à gauche. Le graphique de droite reprend l'évolution de l'intensité de ce processus au cours du temps. La forme du graphique qui représente l'intensité est clairement différente que dans le cas de la simulation du processus de Hawkes avec le noyau exponentiel. Il n'y a plus de "pointes", la forme du graphe est plus lisse. Ceci est dû à l'explication donnée auparavant. L'impact d'une attaque sur l'intensité est fortement visible après un certain laps de temps pour le noyau d'Erlang contrairement au noyau exponentiel où il est direct. Sur la figure 2.2.6, certains clusters sont observables, notamment entre

$t \approx 2.5$ et $t \approx 5$. Lors de chaque attaque, l'intensité croît et ainsi la probabilité qu'une nouvelle attaque surgisse augmente. C'est exactement ce qui était souhaité pour modéliser le risque cyber car comme expliqué, ce risque présente un gros risque d'accumulation. La première attaque a lieu en $t \approx 0.5$, ensuite l'intensité augmente avec une courbe qui est bien similaire à celle sur la figure 2.2.4.

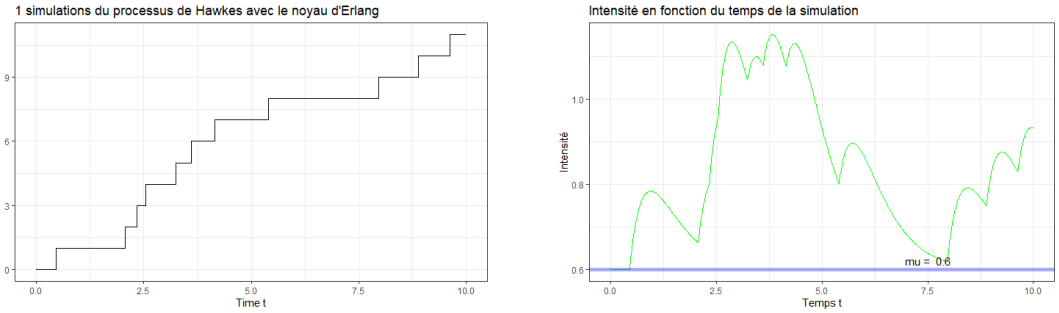

FIGURE 2.2.6 – La simulation d'un processus de Hawkes (noyau d'Erlang) avec $\alpha = 1$, $\beta = 2$ et $\mu = 0.6$ (gauche) et l'évolution de l'intensité de ce processus au cours du temps (droite).

A l'instar du noyau exponentiel, il est possible de fournir une analyse un peu plus poussée pour des paramètres donnés en regardant la distribution du nombre de sauts lorsqu'un grand nombre de simulations est effectué ou en regardant le nombre moyen de sauts. Dans le cas du noyau Power law et du noyau de Rayleigh, le principe est le même que pour les exemples précédents. Il est quand même bon de noter que

$$argmax\ \phi_{pow}(t) = 0$$

et

$$argmax\ \phi_{ray}(t) = \frac{1}{\sqrt{2\beta}}.$$

Jusqu'ici, les simulations des processus de Hawkes furent réalisées avec μ constant mais il est également possible de considérer que μ n'est pas constant. Supposons par exemple que $\mu(t)$ est défini sur $\mathbb{R}^+$ comme

$$\mu(t) = 1 + sin(t).$$

Afin d'utiliser un autre noyau que ceux utilisés pour les exemples précédents, c'est le processus de Hawkes avec un noyau de Rayleigh qui sera simulé, $\forall t \in [0, 10]$,

$$\begin{aligned} \lambda(t) &= \mu(t) + \sum_{T_i<t} \phi_{ray}(t - T_i) \\ &= 1 + sin(t) + \sum_{T_i<t} \phi_{ray}(t - T_i). \end{aligned} \tag{2.2.2}$$

La simulation est reprise sur la figure 2.2.7. Sur la figure de gauche, il y a un cluster d'attaques entre $t \approx 1.5$ et $t \approx 3.5$. La figure de droite représente en vert l'évolution de l'intensité $\lambda(t)$, en rouge l'évolution de la deuxième partie de l'équation (2.2.2) et en bleu, l'évolution de la première

partie de l'équation (2.2.2). La courbe bleue indique l'évolution de l'intensité sans prendre en considération les évènements passés. Celle en rouge représente l'intensité uniquement liée aux attaques passées. La courbe verte est la somme de la courbe bleue et de la courbe rouge. Les premières attaques sont uniquement liées à $\mu(t)$ puisque à ce moment-là, l'intensité liée aux attaques précédentes est nulle. Par exemple en $t \approx 4.5$, une attaque a lieu mais à ce moment-là $1 + sin(t) \approx 0$. Cela signifie que c'est une attaque survenue "à cause" des évènements passés. Cette attaque aurait été très peu probable lors d'un processus de Poisson non-homogène.

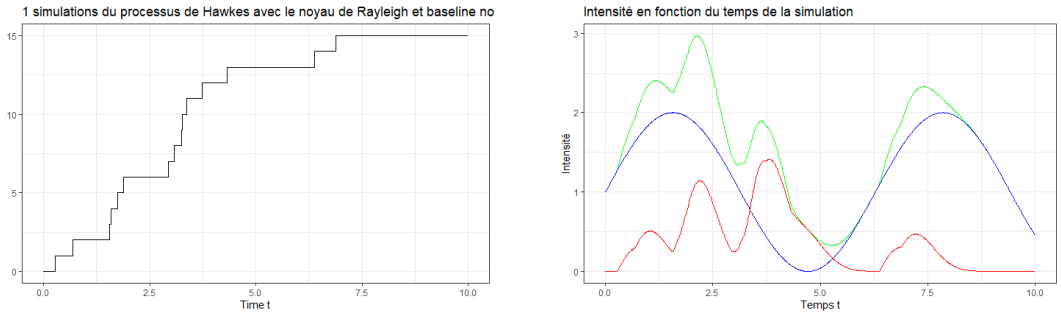

FIGURE 2.2.7 – La simulation d'un processus de Hawkes en utilisant comme fonction noyau le noyau de Rayleigh avec $\alpha = 1$, $\beta = 2$ et $\mu = 1 + sin(t)$ (gauche). L'évolution de l'intensité de ce processus au cours du temps en vert, $1 + sin(t)$ (partie baseline) en bleu et $\sum_{T_i<t} \phi_{ray}(t - T_i)$ (partie contagion) en rouge (droite).

2.2.4 Calibration du modèle

L'objectif est à présent de pouvoir calibrer les différents paramètres à partir de données. Pour ce faire, les seuls inputs requis sont les dates des attaques. Les données qui sont à disposition sont des dates. Chaque date est le moment de l'attaque. C'est aussi pour cela que ce modèle est étudié. Il faudra ensuite tester le modèle sur les données, ce qui sera réalisé dans une section ultérieure. Cette section se concentrera sur l'aspect théorique de la calibration.

L'idée est de calibrer les paramètres en utilisant la méthode de maximisation de la fonction de vraisemblance.

Proposition 2.2.8. *La fonction de log-vraisemblance pour un processus de Hawkes N_t avec une fonction d'intensité de la forme de (2.2.1) est*

$$\log \mathcal{L} = -\int_0^{T_{max}} \mu(u)\mathrm{d}u - \sum_{i=1}^{n}\sum_{j=1}^{i}\int_{T_i}^{T_{i+1}} \phi(u-T_j)\mathrm{d}u + \sum_{i=1}^{n}\left(\log(\mu(T_i) + \sum_{j=1}^{i-1}\phi(T_i-T_j))\right). \quad (2.2.3)$$

Démonstration. Soit un processus de Hawkes défini par la séquence suivante $\{T_1 = t_1, T_2 = t_2, \cdots, T_n = t_n\}$ et un intervalle de temps $[0, T_{max}]$. Avant le début des attaques, $\mathcal{F}_0 = \varnothing$. Par

définition, la fonction de vraisemblance est

$$\mathcal{L} = \text{"}P(T_1 = t_1 \text{ et } T_2 = t_2 \text{ et } \cdots \text{ et } T_n = t_n \text{ et } T_{n+1} > T_{max})\text{"}.$$

Etant donné que chaque saut dépend des sauts précédents, la fonction de vraisemblance dans le cas du processus de Hawkes est donc

$$\mathcal{L} = P(T_{n+1} > T_{max}|\mathcal{F}_n) \prod_{i=1}^{n} P(T_i = t_i \mid \mathcal{F}_{i-1}). \tag{2.2.4}$$

Un processus de Hawkes est un processus de Poisson qui suit une loi exponentielle d'intensité $\lambda(t)$ tel que

$$P(T_n - T_{n-1} > t|\mathcal{F}_{n-1}) = \exp(-\int_{T_{n-1}}^{T_{n-1}+t} \lambda(u)\mathrm{d}u) \tag{2.2.5}$$

avec

$$\lambda(t) = \mu(t) + \sum_{T_n<t} \phi(t - T_n). \tag{2.2.6}$$

Par (2.2.4) et (2.2.5),

$$\begin{aligned}\mathcal{L} &= \exp(-\int_{T_n}^{T_{max}} \lambda(u)\mathrm{d}u) \prod_{i=1}^{n} \exp(-\int_{T_{i-1}}^{T_i} \lambda(u)\mathrm{d}u)\lambda(T_i) \\ &= \exp(-\int_{0}^{T_{max}} \lambda(u)\mathrm{d}u) \prod_{i=1}^{n} \lambda(T_i).\end{aligned}$$

En passant au logarithme,

$$\log \mathcal{L} = -\int_{0}^{T_{max}} \lambda(u)\mathrm{d}u + \sum_{i=1}^{n} \log(\lambda(T_i)).$$

Par (2.2.6),

$$\log \mathcal{L} = -\int_{0}^{T_{max}} \mu(u)\mathrm{d}u - \int_{0}^{T_{max}} \sum_{T_i<u} \phi(u - T_i)\mathrm{d}u + \sum_{i=1}^{n} \Big(\log(\mu(T_i) + \sum_{j=1}^{i-1} \phi(T_i - T_j))\Big).$$

Il est évident que

$$\begin{aligned}\int_{0}^{T_{max}} \sum_{T_i<u} \phi(u - T_i)\mathrm{d}u &= \sum_{i=1}^{n} \int_{T_i}^{T_{i+1}} \sum_{j=1}^{i} \phi(u - T_j)\mathrm{d}u \\ &= \sum_{i=1}^{n} \sum_{j=1}^{i} \int_{T_i}^{T_{i+1}} \phi(u - T_j)\mathrm{d}u.\end{aligned}$$

□

La fonction de vraisemblance est facile à coder si l'intégrale du noyau est connue. En effet, c'est une somme de fonctions. Pour calibrer les différents paramètres, il faut maximiser la fonction de vraisemblance. Dans le cadre ce travail, on se concentrera sur les noyaux avec trois paramètres μ, α et β. Il se peut qu'il y ait plus de paramètres en fonction du noyau qui sera choisi. C'est donc de l'estimation paramétrique. Par exemple pour le noyau Power Law, le nombre de paramètres à estimer sera de quatre. Dans la suite de ce document, l'intensité de base μ sera supposée constante au cours du temps. Pour chaque noyau passé en revue dans la section 2.2.2, la fonction de vraisemblance sera différente.

Dans la première section de l'annexe se trouve le calcul des différentes intégrales. Il est maintenant facile pour chaque noyau d'en tirer l'estimation des paramètres. L'avantage d'avoir des formules fermées pour les intégrales est que le temps de calcul est plus court lorsqu'il s'agira d'exécuter le code. C'est en effet plus rapide et plus précis que d'utiliser une fonction qui approximise l'intégrale.

Etant donné que la fonction de vraisemblance peut être fastidieuse à coder, une erreur peut vite survenir. Pour vérifier qu'aucune erreur n'est faite, il faut simuler un processus de Poisson à l'aide de la méthode de simulation pour le processus de Hawkes décrite plus haut. Il faut ensuite vérifier que l'estimation pour des paramètres est cohérente par rapport aux paramètres utilisés pour faire l'estimation. Ceci est réalisé pour trois noyaux : le noyau exponentiel, le noyau d'Erlang et le noyau de Rayleigh. Le détail est fourni dans la première section de l'annexe.

L'estimation des paramètres est assez bonne pour tous les noyaux. Logiquement, plus le nombre de sauts dans l'échantillon $\{T_1, \cdots, T_n\}$ est grand plus l'estimation des paramètres sera bonne.

Cette partie concernant l'estimation des paramètres est importante. Pour réaliser cela, la fonction *constOptim* est utilisée. Afin que ce soit utile pour une suite éventuelle, il faut que l'estimation des paramètres soit assez rapide. Afin de gagner du temps, il est conseillé d'utiliser la fonction *sum* dans R au lieu de boucles *for* lorsque la fonction de maximum de vraisemblance est codée.

2.3 Test d'adéquation

Il est important de pouvoir vérifier les hypothèses. Cette section explicitera un test qui vérifiera que les données suivent bien un processus de Poisson avec les bons paramètres. Pour commencer, voici une proposition intéressante qui sera utile pour tester le modèles.

Proposition 2.3.1. *[11] Soit un processus de Poisson N_t avec une intensité $\lambda(t)$ et un échantilllon de N_t $\{T_1, \cdots, T_n\}$. Soit*

$$\tau_i = \int_0^{T_i} \lambda(t)\mathrm{d}t, \quad i \in \{1, \cdots, n\}.$$

Alors la séquence $(\tau_i)_{i\in\{1,\cdots,n\}}$ sont des pas de temps distribués selon un processus de Poisson homogène d'intensité 1.

Cette proposition est intéressante pour voir si les données suivent bien le modèle de Hawkes. Cela va permettre de vérifier qu'il n'y ait pas d'erreur qui est faite dans les hypothèses. Puisque la proposition précédente dit que les τ_i sont des pas de temps distribués selon un processus de Poisson homogène d'intensité 1, cela signifie que l'écart entre chaque τ_i suit une loi de Poisson d'intensité 1. Notons

$$\epsilon_i = \tau_{i+1} - \tau_i, \qquad i \in \{1, \cdots, n-1\}.$$

Dés lors, par la proposition 2.3.1, l'échantillon ϵ_i suit une loi exponentielle de paramètre 1. Il existe un test en statistique qui permet de vérifier si un échantillon suit bien une distribution exponentielle. Ce test est celui de Kolgmogorov-Smirnov. L'hypothèse nulle et l'hypothèse alternative sont

H_0 : la distribution est une distribution exponentielle de paramètre 1.

vs

H_1 : la distribution n'est pas une distribution exponentielle de paramètre 1.

Kolmogorov-Smirnov se base sur la différence entre la fonction de répartition théorique et la fonction de répartition empirique construite à l'aide de l'échantillon. Il est également possible d'effectuer ce test pour d'autres distributions telles que la loi normale ou la loi gamma par exemple. La fonction en R qui permet d'effectuer ce test est la fonction *ks.test*.

En plus de vérifier que l'écart entre chaque τ_i suit bien une loi exponentielle, il est également nécessaire de tester l'indépendance entre chaque τ_i. On rappelle que les écarts d'un processus de Poisson homogène sont indépendants les uns des autres. L'hypothèse nulle et l'hypothèse alternative pour le test Ljung-Box sont

H_0 : il n'y a pas d'autocorrélation.

vs

H_1 : il y a de l'autocorrélation.

La fonction en R qui permet d'effectuer ce test est la fonction *Box.test*. Elle requiert un argument : le lag. Quelle est l'interprétation de ce lag ? Le test de Ljung-Box cherche à vérifier si une observation faite aujourd'hui est influencée par les observations passées et va chercher cette information dans les corrélations. Si on suppose que des données sont annuelles, on peut commencer par vérifier si elles sont corrélées à celles mesurées 1 an plus tôt. On a dans ce cas un lag d'un an et on va regarder la corrélation entre les vecteurs X_t et $X_{(t-1)}$. Faire le test uniquement sur un lag d'un an (ou plus généralement, si les résultats sont faits sur un seul lag) ne donne par contre pas toute l'information. Les données pourraient très bien être influencées par les mesures faites n années plus tôt. Il vaut donc mieux répéter le test pour plusieurs lags n, où l'on va vérifier les corrélations entre les vecteurs X_t et $X_{(t-n)}$. Pour un lag m, on va donc vérifier que toutes les corrélations de pas de temps inférieur à m sont nulles. Donc au lag m, on ne regardera pas juste aux corrélations entre X_t et $X_{(t-m)}$, mais aussi $X_{(t-k)}$ avec $k \in [1, \cdots, m]$.

Remarque 2.3.2. Ce test ne teste pas l'indépendance entre les τ_i mais teste s'il y a de l'au-

tocorrélation. Si les τ_i sont autocorrélés, alors ils sont forcément dépendants mais l'absence d'autocorrélation n'implique pas forcément l'indépendance.

Pour la suite, il peut être utile de vérifier que le code fonctionne bien avant d'appliquer cela à une base de données. Si la fonction qui simule un chemin pour un processus de Hawkes et la fonction qui calcule les τ_i pour un chemin donné sont correctes, le test de Kolmogorov-Smirnov devrait renvoyer une bonne p-valeur (p-valeur assez grande) afin d'accepter l'hypothèse nulle. La p-valeur est grande (72%), ce qui signifie que le chemin simulé suit bien un processus de Hawkes.

2.4 Limite du modèle de Hawkes

Le modèle de Hawkes est intéressant car il gère la dépendance entre les polices. Le premier chapitre insiste bien sur cette particularité que contient le risque cyber. La dépendance ne peut pas être gérée par les modèles plus classiques tels que les GLM, les forêts aléatoires, le gradient boosting,... Comme souvent, un modèle contient certaines limites et n'est pas parfait. Le modèle de Hawkes n'y échappe malheureusement pas. Cette partie du travail pointera les désavantages du modèle de Hawkes qui devront être pris en considération par celui qui voudra l'utiliser.

Absence de notion d'exposition Jusqu'ici, aucune notion d'exposition n'a été prise en considération. C'est assez ennuyant lorsqu'une fréquence veut être modélisée. La principale raison qui explique ce désavantage est le manque de données disponibles. Effectivement, le manque de données sur le risque cyber ne permet pas d'inclure l'exposition dans le modèle. Les rares bases de données disponibles sont des bases de données avec uniquement des entreprises ayant eu un sinistre. Il n'y a donc aucune information sur le nombre d'entreprises sans sinistres. Il n'y a pas de bases de données telles que les bases de données disponibles pour une assurance RC auto avec l'ensemble des polices dans la base données même celles qui n'ont pas eu de sinistres. Si tel était le cas,l'exposition pourrait s'insérer dans le modèle. Jusqu'ici, l'intensité du processus est définie par

$$\lambda(t) = \mu(t) + \sum_{T_n<t} \phi(t - T_n).$$

Une hypothèse pour intégrer l'exposition au modèle serait d'avoir une intensité de la forme

$$\lambda(t) = \mu(t)w(t) + \sum_{T_n<t} \phi(t - T_n)w(t)$$

où w est la fonction qui renvoie le nombre d'entreprises en portefeuille au moment t. Au final, en ne prenant pas l'exposition en compte, celle-ci est supposée constante en fonction du temps.

Beaucoup de paramètres à estimer Si la segmentation du portefeuille est grande, il peut y avoir beaucoup de paramètres à estimer. Cela implique par conséquent un temps de calcul plus élevé. Il est donc important d'avoir des algorithmes qui optimisent le temps de calcul au maximum.

Précision des données Pour réaliser un tel modèle, il faut avoir les dates des attaques. C'est assez compliqué d'avoir le moment exact d'une attaque. Une intrusion d'un virus au sein d'un logiciel informatique peut être détectée bien après que celui-ci se soit réellement introduit dans le logiciel.

2.5 Modèle de Hawkes pour plusieurs groupes

Dans la section précédente, le modèle de Hawkes a été modélisé pour l'ensemble d'un portefeuille de sinistres. Prochainement, l'objectif sera de pouvoir diviser le portefeuille en plusieurs groupes afin d'avoir des paramètres propres à chaque sous-groupe. Chaque groupe k aura une intensité de base μ_k qui lui est propre. Les paramètres représentant la contagion vont également différer. La survenance d'une attaque d'un groupe pourrait ne pas avoir la même conséquence sur tous les groupes. Cela semble assez intuitif dans le sens où une attaque dans une entreprise dans le secteur informatique n'aura pas la même influence sur une société pharmaceutique que sur une autre société informatique par exemple.

Supposons que la base de données soit subdivisée en n groupes distincts. L'intensité du processus de Poisson pour le groupe $i \in \{1, \cdots, n\}$ au moment $t \in \mathbb{R}^+$ se définit donc par

$$\lambda^{(i)}(t) = \mu_i(t) + \sum_{t>T_j^{(1)}} \phi_{i1}(t - T_j^{(1)}) + \sum_{t>T_j^{(2)}} \phi_{i2}(t - T_j^{(2)}) + \cdots + \sum_{t>T_j^{(n)}} \phi_{in}(t - T_j^{(n)}),$$

où $T_j^{(i)}$ représente le temps des attaques du groupe i.

C'est-à-dire que pour chaque groupe, il y a $2n + 1$ paramètres à estimer lorsque le noyau est exponentiel, de Rayleigh ou d'Erlang. Le nombre augmente de n lorsqu'il s'agit du noyau Power Law. Etant donné qu'il faut faire cela pour chaque groupe, il y a au total $2n^2 + n$ paramètres à estimer. C'est un nombre qui peut vite grimper lorsque la base de données est subdivisée de manière précise.

2.5.1 Simulation de chemins

L'idée est dans ce cas-ci de pouvoir modéliser l'effet d'auto-excitation et l'effet d'excitation qui provient des autres groupes. Il se pourrait que par exemple l'attaque dans un groupe n'ait pas d'impact sur l'intensité du processus d'un autre groupe. Supposons qu'il y ait deux groupes d'entreprises. Pour la simulation, les paramètres sont les suivants

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} = \begin{bmatrix} 0.6 & 1 \\ 0 & 0.5 \end{bmatrix}.$$

Pour plus d'intuition, les paramètres α_{ij} et β_{ij} représentent l'impact d'une attaque dans le groupe j sur le groupe i. Dans le cas des paramètres pour l'exemple ci-dessus, $\alpha_{21} = 0$. Cela signifie que la survenance d'une attaque dans le groupe 1 n'a pas d'impact sur l'intensité du groupe 2. Cependant, une attaque dans le deuxième groupe va avoir un effet d'excitation sur le premier groupe. Une simulation de ce processus est reprise sur la figure 2.5.1.

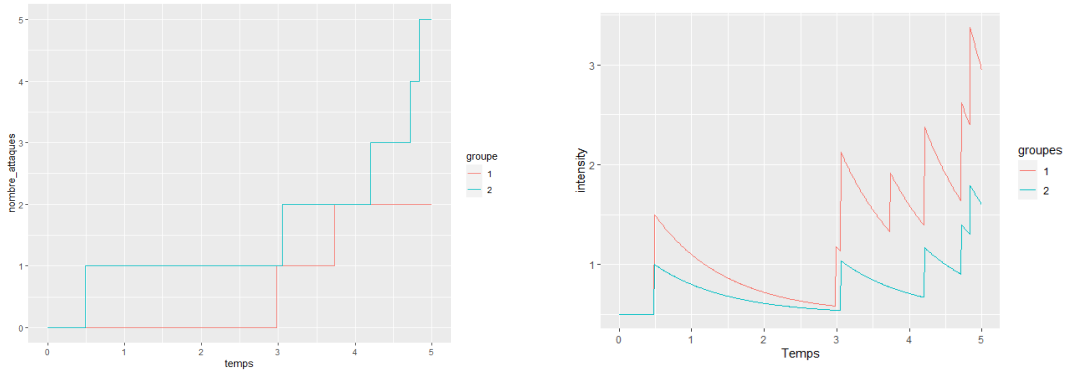

FIGURE 2.5.1 – Le graphique de gauche indique le nombre d'attaques au cours du temps pour chaque groupe et la graphique de droite représente l'intensité au cours du temps pour chaque groupe.

La simulation sur la figure 2.5.1 a été faite en utilisant le noyau exponentiel. Ceci illustre totalement ce qui est expliqué plus haut. La survenance d'une attaque dans le groupe 2 a pour conséquence une augmentation de l'intensité du groupe 1 et du groupe 2. Les intensités n'augmentent cependant pas de la même manière puisque $\alpha_{12} \neq \alpha_{22}$. Il se pourrait également (ce qui n'est pas le cas ici) que $\beta_{12} \neq \beta_{22}$, cela aurait un impact sur la vitesse de décroissance de l'intensité après une attaque au sein du deuxième groupe. Une attaque dans le premier groupe influence uniquement sa propre intensité. Ceci est dû au fait que $\alpha_{21} = 0$. Il y a uniquement deux attaques au sein du premier groupe, en $t \approx 3$ et $t \approx 3.7$. L'intensité du groupe 1 à ces moments-là s'accroît alors que pour le deuxième il n'y a pas d'augmentation de l'intensité.

2.5.2 Calibration de Hawkes avec plusieurs groupe

De manière similaire que pour le cas avec une seule dimension, il est possible de pouvoir calibrer les différents paramètres du modèle. La calibration des paramètres se fait à nouveau par la méthode de maximum de vraisemblance.

Proposition 2.5.1. *La fonction de log-vraisemblance pour un processus de Hawkes multivarié*

Algorithm 3 Algorithme pour simuler un processus de Hawkes multivarié avec k groupes entre [0,T] [6]

Require: $T \geqslant 0, \mu(t) \geqslant 0 \ \ \& \ \ \phi(t) \geqslant 0 \ \forall t$ ▷ Il faut que $\lambda(t)$ soit toujours positif
  sauts ← [] ▷ Pour stocker le temps des attaques.
  groupes ← [] ▷ Pour stocker le groupe dans lequel l'attaque a lieu.
  $t \leftarrow 0$
  while $t \leqslant T$ **do** ▷ On continue tant qu'on n'atteint pas T
    $\bar{\lambda} \leftarrow \sup \sum_{i=1}^{k} \lambda_i(s)$ ▷ On trouve un supremum de $\lambda(s)$, le supremum est donc modifié après chaque ajout d'une attaque dans la liste
    $u \leftarrow unif(1)$
    $w \leftarrow -\frac{\ln(u)}{\bar{\lambda}}$ ▷ Simulation d'une exponentielle de paramètre $\bar{\lambda}$
    $t \leftarrow t + w$ ▷ saut potentiel du processus
    $D \leftarrow unif(1)$
    if $D \leqslant \frac{\lambda(t)}{\bar{\lambda}}$ **then**
      $j \leftarrow 1$
      $\lambda \leftarrow \lambda_1(t)$
      while $D\bar{\lambda} > \lambda$ **do**
        $j \leftarrow j + 1$
        $\lambda \leftarrow \lambda + \lambda_j$
      end while
      sauts.*append*(t) ▷ Le saut est ajouté à la liste avec une probabilité $\frac{\lambda(t)}{\bar{\lambda}}$
      groupes.*append*(j) ▷ Le saut a lieu dans le groupe j avec une probabilité $\frac{\lambda_j(t)}{\lambda(t)}$
    end if
  end while

$(N_t^i)_{i=1,\cdots,d}$ *est*

$$\begin{aligned}\log \mathcal{L}^{(i)} = &- \int_0^{T_{max}^{(i)}} \mu_i(u)\mathrm{d}u \\ &- \sum_{k=1}^{d} \sum_{m=1}^{n^{(k)}} \sum_{n=1}^{m} \int_{T_m^{(k)}}^{T_{m+1}^{(k)}} \phi_{ik}(u - T_n^{(k)})\mathrm{d}u \\ &+ \sum_{m=1}^{n^{(i)}} \left[\log \left\{ \mu_i(T_i^{(m)}) + \sum_{k=1}^{d} \sum_{T_n^{(k)} < T_m^{(i)}} \phi_{ik}(T_m^{(i)} - T_n^{(k)}) \right\} \right],\end{aligned} \tag{2.5.1}$$

où $n^{(k)}$ est le nombre total de sauts dans le groupe k avec $k \in \{1, \cdots, d\}$ et T_j^i est le $j$ème saut dans le groupe i.

Démonstration. Le début de la preuve est identique à celle pour le cas univarié. Il est facile de

vérifier que

$$\log \mathcal{L}^{(i)} = -\int_0^{T_{max}} \lambda^{(i)}(u)\mathrm{d}u + \sum_{m=1}^{n^{(i)}} \log(\lambda^{(i)}(T_m^{(i)})).$$

La première partie de l'expression ci-dessus devient

$$\begin{aligned} -\int_0^{T_{max}} \lambda^{(i)}(u)\mathrm{d}u &= -\int_0^{T_{max}} \mu_i(u)\mathrm{d}u - \int_0^{T_{max}} \sum_{k=1}^{d} \sum_{T_n^{(k)}<u} \phi_{ik}(u - T_n^{(k)})\mathrm{d}u \\ &= -\int_0^{T_{max}} \mu_i(u)\mathrm{d}u - \sum_{k=1}^{d}\sum_{m=1}^{n^{(k)}}\sum_{n=1}^{m} \int_{T_m^{(k)}}^{T_{m+1}^{(k)}} \phi_{ik}(u - T_n^{(k)})\mathrm{d}u. \end{aligned}$$

Il est facile de vérifier que la seconde partie de la somme devient

$$\sum_{m=1}^{n^{(i)}} \log(\lambda^{(i)}(T_m^{(i)})) = \sum_{m=1}^{n^{(i)}} \left[\log\left\{ \mu_i(T_i^{(m)}) + \sum_{k=1}^{d} \sum_{T_n^{(k)}<T_m^{(i)}} \phi_{ik}(T_m^{(i)} - T_n^{(k)}) \right\} \right].$$

□

Il est possible de coder cette fonction en R. L'objectif étant d'allier vitesse de calcul et précision, il est conseillé d'utiliser les formes fermées des intégrales ainsi que de coder la fonction gradient de l'équation (2.5.1). Le détail de ces calculs se trouve dans l'annexe. La forme de la fonction de maximum contient beaucoup de sommes. Faire un code avec des boucles *for* engendrerait un temps de calcul assez conséquent. Il est possible d'économiser du temps en réorganisant le code et en utilisant des fonctions *sum*.

Par exemple, pour vérifier si les codes sont cohérents et qu'il n'y a pas d'erreur, il est toujours utile de vérifier à l'aide de simulations. Comme pour le cas univarié, on exécute une simulation avec certains paramètres. On essaye de calibrer ces paramètres en utilisant les données de la simulation et en théorie, les paramètres estimés devraient être proches de ceux qui ont été utilisés pour la simulation.

2.6 Application du processus de Hawkes

Au cours de cette section, l'application du processus de Hawkes sera exécutée sur une base de données. La base de données est celle du PRC. Dans la première partie de cette section, une présentation des données sera faite. Ensuite, une application du modèle de Hawkes univarié sera réalisée sur le jeu de données. Enfin, une analyse plus poussée avec plusieurs groupes sera établie.

2.6.1 Présentation des données

Dans la base de données du PRC, il y a 14 variables. Elle reprend la date de 9015 vols de données au sein d'entreprises américaines entre 2006 et 2019 et parle donc uniquement de vols de données, c'est juste pour un sous groupe des attaques cyber. Parmi les variables se trouvent la date à laquelle le vol fut déclaré, le nom de la compagnie, la ville de l'entreprise ainsi que l'Etat, le type de vol, le secteur d'activité de l'entreprise, le nombre de données volées, une description de l'incident, l'année du vol, la latitude et longitude de l'entreprise et la source qui a rendu le vol public. Pour la suite, on utilisera uniquement les variables suivantes :

- **Date à laquelle le vol fut déclaré**. Ce sont des dates qui vont de mars 2006 à 2019.
- **Le type de vol**. Il y a 7 types de vols repris dans cette base de données : Une fraude différente du hacking, le hacking, vol d'un employé, vol physique (vol de documents), source du vol inconnue, vol lié à une perte d'un objet portatif (téléphone, ordinateur portable,...) ou acte involontaire (hormis une perte).
- **Le secteur d'activité de l'entreprise**. Dans cette base de données, huit secteurs d'activité différents sont repris : les business qui sont subdivisés eux-mêmes en trois (business financiers (BSF), commerces (BSR) et les autres business(BSO)), institutions éducatives (EDU), militaires et gouvernement (GOV), entreprises liées aux soins de santé, association sans but lucratif (NGO) et les entreprises dont on ne connait pas le secteur d'activité (UNK).
- **L'année de l'attaque**. L'année durant laquelle s'est produit le vol.
- **L'Etat** Étant donné que ce sont des évènements qui ont eu lieu aux États-Unis, la base de données fournit l'Etat dans lequel l'attaque a eu lieu.

L'objectif étant d'appliquer le modèle de Hawkes, il faut avoir la date exacte du moment où le vol s'est déroulé. Cette information, la base de données ne la fournit pas, seule l'année est mentionnée. En général, l'année du vol correspond à l'année où l'attaque a été rendue public. On fait l'hypothèse que la date de l'attaque est identique à celle de la date de publication de l'attaque. Cela fait également sens si on se réfère à la police d'AXA étudiée dans le chapitre 2 puisque l'assureur couvrira à partir du moment où l'attaque est déclarée.

2.6.2 Modèle avec Hawkes univarié

La base de données décrite ci-dessous concerne uniquement les vols de données. Or, l'objectif de ce travail est d'étudier le risque cyber. De ce fait, il y a des types de vols qui ne sont pas intéressants pour le risque cyber. Le vol physique par exemple n'est pas pertinent pour étudier le risque cyber. Les vols liés à un objet ne seront pas étudiés car bien que ce soit des objets électroniques, il semble peu probable qu'une assurance pour le risque cyber couvre ce type de sinistres. De plus, comme longuement expliqué dans différentes sections, le modèle de Hawkes est un modèle qui prend en compte la dépendance ainsi que l'auto-excitation. Un acte involontaire d'un employé ou une erreur de communication ne devront pas être modélisés avec Hawkes car il n'y a pas de raison que ces types d'attaque aient un effet auto-excitant. Le fait qu'un employé fasse une erreur dans une entreprise A n'augmente pas la "probabilité" qu'un employé fasse une erreur dans une entreprise B. L'application du modèle de Hawkes va se faire en utilisant uniquement

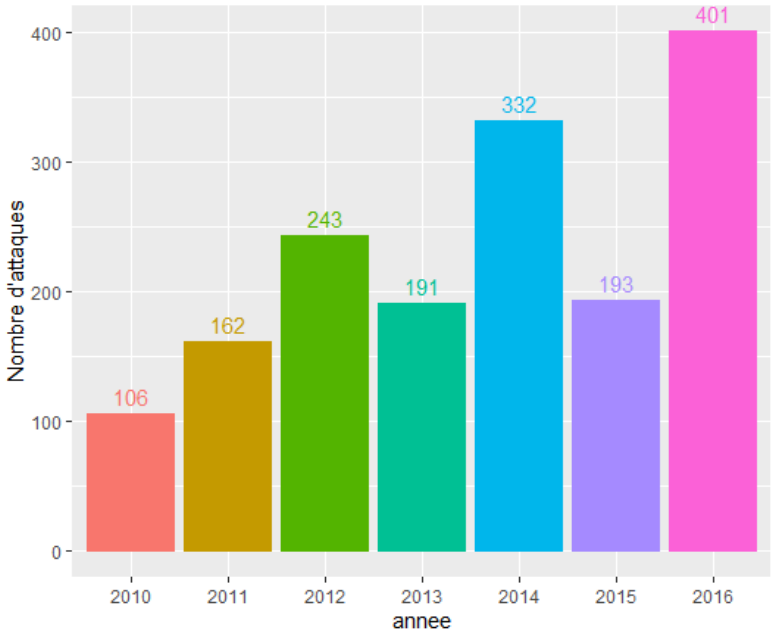

FIGURE 2.6.1 – Le nombre d'attaques par année dans la base de données du PRC.

les vols liés au hacking. C'est assez intuitif que le hacking présente un effet auto-excitant. Vu qu'il n'y a que très peu de données avant 2010, seuls les évènements qui ont eu lieu après seront utilisés. La base de données n'utilisera pas les données de 2019 puisqu'elle a été publiée en 2019 ce qui implique que des évènements ayant été reportés en 2019 n'ont pas été ajoutés. Dans un premier temps, regardons comment le nombre de sinistres évolue en fonction du temps.

Sur la figure 2.6.1, il est facile de faire le constat que le nombre de hacking a tendance à augmenter avec le temps. Peut-on calibrer le modèle pour une année en utilisant les données de l'année précédente ? Cela pose quand même question par rapport à ce qui a été réalisé jusqu'ici. En effet, dans les parties précédentes de ce travail, la partie "baseline" de l'intensité $\lambda(t)$ était considérée constante. Il serait peut-être plus judicieux de trouver une "baseline" qui dépende du temps. Au lieu de prendre

$$\lambda(t) = \mu + \sum_{T_n<t} \phi(t - T_n), \tag{2.6.1}$$

il serait intéressant de considérer

$$\lambda(t) = \mu + \frac{\gamma t}{365} + \sum_{T_n<t} \phi(t - T_n). \tag{2.6.2}$$

De cette manière, l'intensité à laquelle sont soumises toutes les entreprises augmenterait avec le temps si γ est positif et diminuerait si γ est négatif.

	p-valeur test KS	p-valeur Box-Ljung
Intensité de base constante ($\gamma = 0$)	25.45%	1.54%
Intensité de base qui évolue avec le temps ($\gamma \neq 0$)	42.78%	57.74%

FIGURE 2.6.2 – Résultats pour les tests d'adéquation du modèle aux données du PRC entre 2010 et 2016.

	Min.	Q.25%	Médiane	Moyenne	Q.75%	Max
$\gamma = 0$	164	217	231	231.4	245	310
$\gamma \neq 0$	273	348	365	364.9	381	466

FIGURE 2.6.3 – Synthèse du nombre d'attaques en 2017 selon les simulations avec les paramètres des modèles respectifs.

En considérant une intensité constante au cours du temps, pour les attaques entre 2010 et 2016 et en utilisant des noyaux d'Erlang, les résultats sont les suivants. Lorsqu'on applique le test d'adéquation, la p-valeur est relativement bonne puisqu'elle vaut 25.45% pour le test de Kolmogorov-Smirnov. A première vue, il ne faudrait pas utiliser une intensité de la forme de (2.6.1).

Néanmoins, il est également intéressant de regarder les résultats obtenus avec l'intensité (2.6.2). Lorsque la calibration est faite, le test de Kolmogorov Smirnov renvoie une p-valeur de 42.78%. C'est une p-valeur qui est meilleure que pour le premier test.

Il est utile de regarder si les écarts entre les τ_i sont indépendants en utilisant le test de Box-Ljung. Dans ce cas, les résultats sont frappants. Le deuxième modèle est nettement meilleur puisque la p-valeur du test est de 57.74%. À l'inverse, la p-valeur pour le premier modèle pour le test de Box-Ljung vaut 1.54%. Cela signifie que les écarts entre les τ_i ne sont pas indépendants. Cela nous pousse à penser qu'il est préférable d'utiliser le deuxième type d'intensité. Les différents résultats sont repris à la figure 2.6.2.

Les deux modèles ont été calibrés sur des données entre 2010 et 2016. Il est intéressant de comparer le nombre de sinistres prédits pour 2017 et le nombre de sinistres qui ont réellement eu lieu. Pour ce faire, il faut simuler un grand nombre de processus en utilisant les paramètres estimés. Le nombre de processus simulés est de 10 000 dans ce travail. Les résultats des simulations se trouvent sur la figure 2.6.3.

Le nombre d'attaques en 2017 étant de 385, il est facile de constater que le modèle avec un gamma différent de zéro est meilleur. Le nombre d'attaques maximum de simulation pour le cas $\gamma = 0$ est inférieur au nombre d'attaques en 2017. On remarque facilement qu'en fait, le nombre moyen de simulations pour le cas $\gamma = 0$ est fortement semblable au nombre moyen d'attaques entre 2010 et 2016 qui est de 232.57. Le modèle ne prend donc pas en compte cette augmentation de l'intensité des attaques au fur et à mesure du temps.

Les résultats sont nettement meilleurs pour le modèle qui prend en considération l'augmentation de la "baseline" au fil du temps. Le nombre de sinistres observés est de 20 au-dessus par rapport à la moyenne des attaques des simulations. De plus, le nombre de sinistres observés se trouve entre

les quantiles 5% et 95% (figure 2.6.4). Le backtesting confirme que le deuxième modèle est plus approprié.

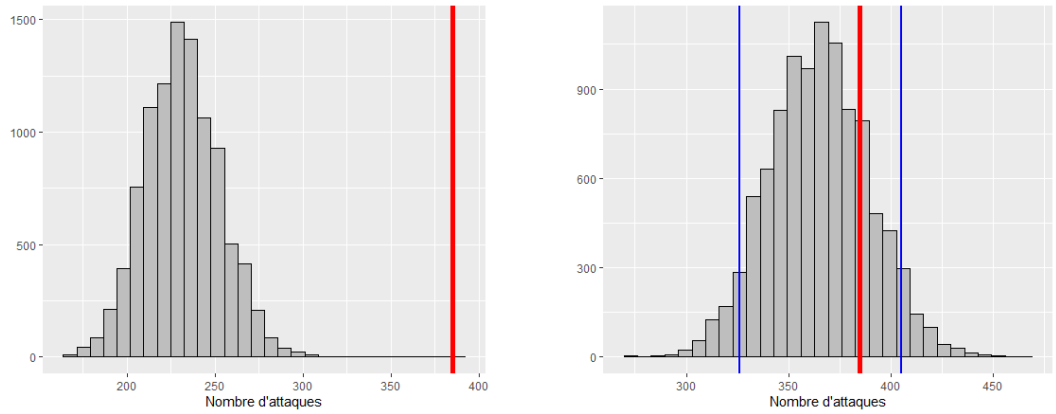

FIGURE 2.6.4 – Graphiques du nombre de simulations en fonction du nombre d'attaques (10 000 simulations au total). À gauche : $\gamma = 0$ et à droite : $\gamma \neq 0$. La barre verticale rouge est le nombre d'attaques en 2017. Les barres bleues sont les quantiles 5% et 95%.

2.6.3 Modèle de Hawkes avec trois groupes

Dans cette section, le modèle de Hawkes sera appliqué sur plusieurs groupes. La figure 2.1.4 indique qu'un facteur important qui impacte le risque cyber est le secteur d'activité de l'entreprise. A priori, il peut y avoir de grandes différences en fonction du secteur d'activité. Dans un premier temps, le portefeuille sera divisé en trois catégories. La catégorie 1 regroupera les business (BSF, BSO, BSR). La catégorie 2 englobera uniquement les entreprises du secteur médical. Enfin, la catégorie 3 regroupera les autres secteurs (EDU, GOV, NGO).

Avant d'appliquer les données, regardons comment évolue le nombre de sinistres par année pour chaque groupe. C'est uniquement une analyse descriprive qui a pour objectif de trouver la fenêtre de calibration. Les graphiques avec le nombre de sinistres par années au fil du temps se trouvent dans l'annexe sur les figures B.0.1, B.0.2 et B.0.3. Pour la première catégorie, on remarque sur la figure B.0.1 que le nombre de sinistres est comparable d'une année à l'autre avec une légère tendance à la hausse avec le temps qui sera prise en compte grâce au paramètre γ du processus. Pour la catégorie 3 sur la figure B.0.3, entre 2010 et 2012, le nombre de sinistres a une tendance claire à augmenter avec le temps. Cependant, après 2012, le nombre de sinistres diminue clairement avec le temps. Pour la catégorie 2, les quatre premières années le nombre de sinistres est comparable d'une année à l'autre avec une légère hausse avec le temps. Ensuite, en 2014, le nombre de sinistres passe à 180 alors qu'en 2013, il était de 44. En 2015, le nombre semble à nouveau en adéquation avec les quatre premières années puisqu'il vaut 73. En 2016, c'est à nouveau bizarre puisqu'il saute à nouveau pour atteindre 225 sinistres. Le nombre de sinistres dans ce groupe est assez volatile à en croire cette analyse descriptive. En analysant cela, il serait

peut-être préférable de faire la calibration, uniquement sur les 3 dernières années et non sur les 6 dernières comme cela a été fait pour l'analyse univariée. En effet, puisque le nombre de sinistres ne reflète plus la tendance des années les plus proches, les introduire dans la calibration pourrait déteriorer les résultats. Le processus de Hawkes est calibré sur une fenêtre de 3 ans et de 6 ans afin de voir si le modèle calibré entre 2014 et 2016 a un meilleur caractère prédictif que celui ajusté entre 2010 et 2016 comme pressenti. Il y a deux objectifs dans cette section. Le premier est de montrer que le risque évolue mais que la tendance de l'évolution peut changer avec le temps et que comme pour un portefeuille auto par exemple, les données trop anciennes ne reflètent plus la réalité. Le second est de mettre en évidence l'importance de l'analyse descriptive avant de lancer le modèle. Il est également important de noter que pour ce mémoire, la tendance linéaire est considérée mais une tendance non linéaire aurait pu être choisie. Dans ce cas, le nombre de paramètres aurait augmenté.

L'ajustement des paramètres est d'abord fait en utilisant les données entre 2010 et 2016. L'objectif est de trouver le meilleur modèle pour chaque groupe. La calibration est faite pour le noyau exponentiel, de Rayleigh et d'Erlang par maximisation de la fonction de vraisemblance. Afin de connaître le noyau le plus adéquat, il faut prendre celui qui a la plus grande valeur pour la fonction de maximum de vraisemblance.

	Exponentiel	Rayleigh	Erlang
Catégorie 1	1634.85	1647.40	1644.79
Catégorie 2	1247.28	1250.36	1243.67
Catégorie 3	851.98	864.33	860.12

FIGURE 2.6.5 – L'inverse du maximum de vraisemblance pour chaque type de noyau avec les données entre 2010 et 2016.

La figure 2.6.5 reprend l'inverse de la fonction de vraisemblance après avoir réalisé la maximisation. Le meilleur noyau est le noyau exponentiel pour la première et troisième catégorie alors que le noyau d'Erlang semble plus propice pour le deuxième groupe.

Avant de commencer le backtesting, il est important de réaliser les différents tests d'adéquation. Pour les catégories 1 et 3, les résultats des tests sont excellents. Pour les catégories 1 et 3, les p-valeurs respectives du test de Kolmogorov-Smirnov sont de 70.1% et 94.08%. De plus, les Ljung-Box tests donnent également des résultats performants. Peu importe le lag, les p-valeurs sont au dessus de 10%. La figure B.0.4 dans l'annexe reprend les différentes p-valeurs en fonction du lag pour les catégories 1 et 3.

	Test Kolmogorov-Smirnov
Catégorie 1 (noyau exponentiel)	70.1%
Catégorie 3 (noyau exponentiel)	94.08%

FIGURE 2.6.6 – Résultats du test de Kolmogorov-Smirnov pour les catégories 1 et 3 en utilisant les données entre 2010 et 2016.

Pour la catégorie 2, les résultats sont nettement moins bons. Pour le noyau d'Erlang, la p-valeur est de 0.2% pour le test de Kolmogorov-Smirnov. Le test de Ljung-Box donne également des mauvais résultats (voir sur la figure B.0.5). Pour la catégorie 2, le second meilleur noyau selon le critère du maximum de vraisemblance est le noyau exponentiel. Il faut regarder si, pour ce dernier, les tests d'adéquation sont meilleurs. Pour le test de Kolmogorov-Smirnov, la p-valeur est de 1.78%. C'est une p-valeur qui n'est toujours pas bonne. Pour le test de Ljung-Box, les p-valeurs sont également inférieures à 5% quand le lag est supérieur à 2 (voir sur la figure B.0.6). Enfin, regardons si le noyau de Rayleigh passe les tests d'adéquation. Comme pour les deux noyaux précédents, le noyau de Rayleigh obtient une p-valeur de 0.06% pour le test de Kolmogorov-Smirnov. Pour le test de Box-Ljung, les p-valeurs sont inférieures à 1% à chaque fois que le lag est plus grand que 1 (voir figure B.0.7).

	Test Kolmogorov-Smirnov
Noyau exponentiel	1.78%
Noyau de Rayleigh	0.06%
Noyau d'Erlang	0.2%

FIGURE 2.6.7 – Résultats du test de Kolmogorov-Smirnov pour la deuxième catégorie en fonction du noyau en utilisant les données entre 2010 et 2016.

Etant donné qu'aucun noyau ne passe les tests pour la catégorie 2, le noyau avec le meilleur critère selon le maximum de vraisemblance sera choisi pour la suite c'est-à-dire le noyau d'Erlang. Essayons d'expliquer pourquoi la catégorie 2 ne donne pas des résultats satisfaisants par rapport aux deux autres catégories. C'est sans doute en partie lié aux commentaires faits lors de l'analyse descriptive. Le nombre de sinistres en 2014 (180) et en 2016 (225) n'étant pas du tout du même ordre de grandeur que les autres années, le modèle calibré n'est de ce fait pas vraiment optimal. Ainsi, les test d'adéquation ne sont pas bons.

Puisque la calibration fût faite entre 2010 et 2016, une comparaison du nombre de sinistres observés en 2017 avec le nombre de sinistres prédits par le modèle est faite. Pour réaliser cela, 2500 simulations sont faites avec les valeurs des différents estimateurs afin de pouvoir faire la compraison.

On remarque que le modèle a tendance à légèrement sous-estimer le nombre de sinistres pour les catégories 1 et 2 et à légèremment sur-estimer le nombre de sinistres pour la troisième catégorie si l'on compare la valeur moyenne des simulations et le nombre de sinistres observé en 2017. Cependant, le nombre de sinistres observés se trouve toujours entre les quantiles 5% et 95% des sinistres prédits, le nombre de sinistres prédits est donc réaliste. Les résultats sont donc bons alors que les tests d'adéquation ne le sont pas nécessairement.

A présent, le modèle est calibré entre 2014 et 2016. On remarque déjà que le noyau adéquat pour un groupe n'est pas forcément identique si la calibration est faite entre 2010 et 2016 ou entre 2014 et 2016. Le maximum de la fonction de vraisemblance pour chaque noyau et pour chaque groupe est repris sur la figure 2.6.9. Dans ce cas-ci, le noyau de Rayleigh est plus optimal pour

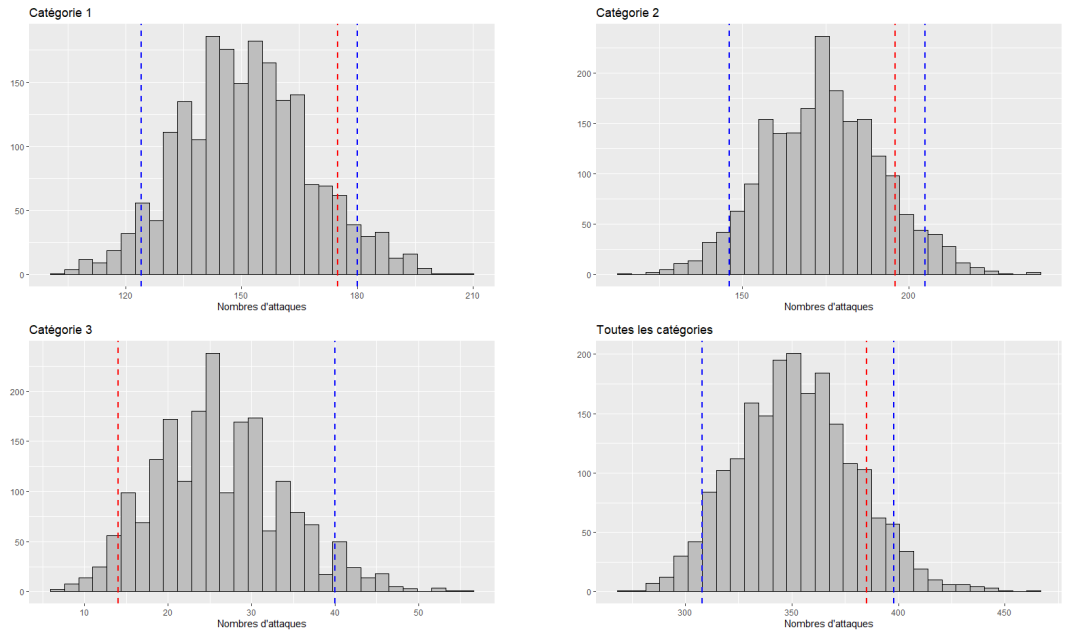

FIGURE 2.6.8 – Comparaison des simulations du nombres d'attaques en 2017 avec le nombre d'attaques observées en 2017 (en rouge). En bleu, ce sont les quantiles 5% et 95% des 2500 simulations. Ceci est fait séparément pour chaque catégorie ainsi que pour toutes les catégories rassemblées. La calibration des paramètres est faite en utilisant les données entre 2010 et 2016.

les deux premières catégories. Pour la catégorie 3, le noyau exponentiel reste le meilleur selon le critère du maximum de vraisemblance.

	Exponentiel	Rayleigh	Erlang
Catégorie 1	756.99	748.80	760.72
Catégorie 2	807.11	807.05	816.46
Catégorie 3	271.17	272.08	272.11

FIGURE 2.6.9 – L'inverse du maximum de vraisemblance pour chaque type de noyau avec la calibration entre 2014 et 2016.

Lorsque la calibration fût faite sur les six dernières années, les p-valeurs du test de Kolmogorov-Smirnov étaient bonnes uniquement pour la catégorie 1 et 3. Maintenant, la figure 2.6.10 montre que les p-valeurs sont correctes pour tous les groupes.

	Test Kolmogorov-Smirnov
Catégorie 1	43.8%
Catégorie 2	18.9%
Catégorie 3	95.2%

FIGURE 2.6.10 – Résultats du test de Kolmogorov-Smirnov pour les différentes catégories lorsque le modèle est calibré sur les données entre 2014 et 2016.

Pour le test de Ljung-Box, les résultats sont également satisfaisants comme en témoignent les figures reprises en annexe sur les figures B.0.8, B.0.9 et B.0.10. Pour ce test, les p-valeurs pour la deuxième catégorie sont mauvaises une fois que le lag est supérieur à 3. Comparons à présent le nombre de sinistres prédits et le nombre de sinistres observés pour l'année 2017 sur la figure 2.6.11.

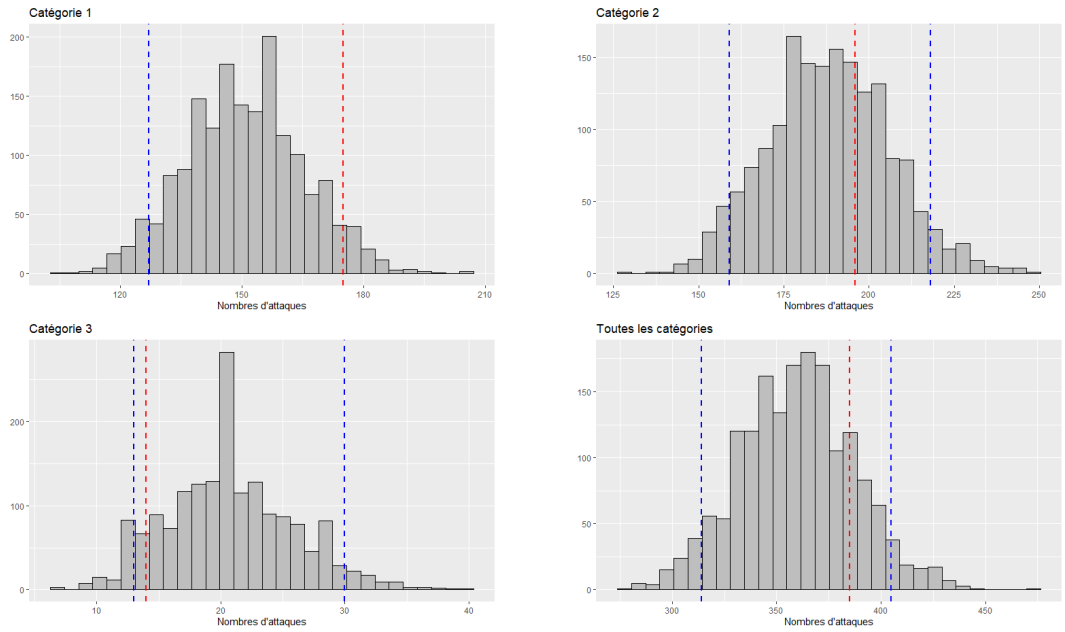

FIGURE 2.6.11 – Comparaison du nombre d'attaques simulées en 2017 avec le nombre d'attaques observées en 2017 (en rouge). En bleu, ce sont les quantiles 5% et 95% des 2500 simulations. Ceci est fait pour chaque catégorie séparément ainsi que pour toutes les catégories rassemblées. La calibration des paramètres est faite en utilisant les données entre 2014 et 2016.

On remarque que le nombre de sinistres observés pour la catégorie 1 est exactement égal au quantile 95% des simulations. Les prédictions pour les deux autres groupes sont réalistes puisqu'elles se trouvent entre les quantiles 5% et 95%.

	2010-2016	2014-2016
Catégorie 1	92%	95 %
Catégorie 2	90.2%	68 %
Catégorie 3	5%	10.8%
Toutes les catégories	88.7%	82.8%

FIGURE 2.6.12 – Quantile correspondant au nombre de sinistres observés en 2017 pour la série de simulation pour les deux sortes de calibration des paramètres.

Le pouvoir prédictif est légèrement meilleur pour le deuxième modèle. Pour aller plus loin, il faudrait étudier d'autres périodes de calibration (prendre les deux dernières années par exemple) et comparer avec d'autres années (estimer le nombre de sinistres en 2014 en calibrant avec les données de 2010 à 2013).

2.6.4 Modèle de Hawkes avec cinq groupes

Dans ce dernier exemple, le modèle sera calibré sur cinq groupes. L'objectif étant d'inclure la variable qui indique l'Etat dans lequel l'entreprise se trouve. Une distinction est faite entre les grands Etats (Californie, Maryland, New-York, Texas, Floride, Massachusetts, Washington, Pennysylvanie et Illinois) et les petits Etats (les autres Etats). Le portefeuille est subdivisé en cinq groupes.

— **Groupe 1** : les Business dans les grands Etats.

— **Groupe 2** : les Business dans les petits Etats

— **Groupe 3** : les entreprises du secteur médical dans les grands Etats.

— **Groupe 4** : les entreprises du secteur médical dans les petits Etats.

— **Groupe 5** :ce sera le même groupe que la catégorie 3 de la section précédente. Etant donné que le nombre d'entreprises dans cette catégorie est bas, aucune distinction entre les Etats sera faite.

La calibration est faite en utilisant les données de 2014 à 2016 puisque dans la section précédente, les résultats semblaient être meilleurs. Un drift est également incorporé dans le modèle.

	Exponentiel	Rayleigh	Erlang
Groupe 1	616.66	610.24	613.35
Groupe 2	371.96	370.61	370.81
Groupe 3	457.14	488.51	458.28
Groupe 4	619.72	622.45	618.92
Groupe 5	265.46	266.90	265.01

FIGURE 2.6.13 – L'inverse du maximum de vraisemblance pour chaque type de noyau avec la calibration entre 2014 et 2016.

Pour chaque groupe, le test d'adéquation est fait pour le noyau qui semble le plus adéquat par la méthode du maximum de vraisemblance. Le test de Kolmogorov-Smirnov passe pour tous les groupes pour un seuil de 10%. Le test de Ljung-Box est bon pour tous les groupes à un seuil de 10% pour un lag de 10 hormis pour le groupe 4 où le test renvoie des p-valeurs inférieures à 5% pour des lags supérieurs à 3. Pour le groupe 4, les p-valeurs sont mauvaises peu importe le noyau. Plus de détails sont disponibles sur la figure C.0.2.

	Test Kolmogorov-Smirnov
Groupe 1	61.3%
Groupe 2	88.5%
Groupe 3	29.5%
Groupe 4	57.2%
Groupe 5	69.8%

FIGURE 2.6.14 – Résultats du test de Kolmogorov-Smirnov pour les différentes catégories lorsque le modèle est calibré sur les données entre 2014 et 2016.

Il est à nouveau possible de prédire le nombre de sinistres et de le comparer au nombre de sinistres observés dans chaque groupe. Les résultats de cette étape sont sur la figure C.0.3 dans l'annexe. Pour les groupes 1 et 2, le modèle a prédit un nombre de sinistres inférieur à celui observé mais le quantile est bien entre 5% et 95% donc le nombre prédit est réaliste. Pour le groupe 2, le nombre de sinistres observés en 2017 est supérieur au quantile 95% des simulations. Le quantile correspondant au nombre de sinistres observés est le quantile 97%. Pour les catégories 3 et 4, les résultats semblent nettement meilleurs. Le quantile correspondant au nombre de sinistres observés est le quantile 67% pour le groupe 3 et le quantile 61% pour le quatrième groupe. Le pouvoir prédictif pour le cinquième groupe est meilleur que dans la modélisation précédente (avec 3 groupes). Le quantile correspondant au nombre de sinistres observés est le quantile 72% pour le groupe 5. Au total, le nombre de sinistres observés correspond au quantile 83% des simulations. On remarque que le quantile pour le nombre de sinistres total avec cinq groupes est plus élevé que lorsque la modélisation a été faite pour trois groupes. Pourtant, le nombre de paramètres est plus grand. En effet, lorsqu'il y a cinq groupes, le nombre de paramètres est égal à 60 tandis que lorsqu'il y a trois groupes, le nombre de paramètres est égal à 24. Le nombre de paramètres a plus que doublé mais les résultats ne sont au final pas nécessairement meilleurs. La variable Etat n'améliore pas le caractère prédictif du modèle.

2.7 Interprétation des résulats

Dans cette section, l'idée est de tirer une interprétation des résultats obtenus. L'interprétation utilisera les résultats pour la modélisation à cinq groupes. Les estimateurs de μ et γ sont sur les figures 2.7.1 et 2.7.2.

μ_1	μ_2	μ_3	μ_4	μ_5
0.1357	0.0727	0.0216	0.0667	0.0658

FIGURE 2.7.1 – Estimateurs des μ.

γ_1	γ_2	γ_3	γ_4	γ_5
0.01236	0.0074	-0.0054	0.0732	-0.0126

FIGURE 2.7.2 – Estimateurs des γ.

Les μ sur la figure 2.7.1 représentent donc l'intensité de base pour chaque groupe au premier janvier 2014 puisque la calibration a été faite à partir de cette date-là. Ensuite, l'intensité de base va augmenter ou diminuer en fonction du γ de manière linéaire. L'intensité de base au temps t (t est exprimé en nombre de jours) pour le groupe i sera égale à $\mu_i + \frac{\gamma_i t}{365}$ où t représente un nombre de jour.

En 2014, ce sont les business situés dans les grands Etats qui semblaient les plus touchés ($\mu_1 = 0.157$) dans le portefeuille. L'intensité de base en 2014 pour les groupes 2, 4 et 5 est assez semblable et est inférieure de moitié par rapport au premier groupe. Les entreprises liées au secteur médical dans les grands Etats semblaient moins touchées. La fréquence augmente avec le temps pour les groupes 1, 2 et 4 alors que pour les groupes 3 et 5, la tendance semble être à la baisse. C'est le groupe 4 qui subit l'évolution la plus importante car l'intensité de base en 2015 aura plus que doublé en comparaison avec 2014.

Remarque 2.7.1. C'est assez compliqué dans le cas de tirer ici de réelles conclusions quant aux réels facteurs de risque pour le risque cyber. Par exemple, l'intensité μ_1 est plus ou moins deux fois plus grande que l'intensité μ_2 alors peut-on dire que les entreprises du groupe 1 sont deux fois plus sujettes à des attaques cyber ? Non, la base de données ne fournit aucune notion quant à l'exposition des entreprises. Le processus de Hawkes permet d'estimer le nombre total de sinistres par groupe sur une période de temps T. Imaginons qu'au départ, il y ait 4 fois plus d'entreprises du groupe 1 que du groupe 2 alors une entreprise du groupe 2 est plus sujette à une attaque que le groupe 1 alors que $\mu_1 > \mu_2$. Toujours dans ce cas-ci, λ_i peut à la fois correspondre à l'augmentation (ou diminution) du risque qu'une entreprise du groupe i a de subir une attaque au fil du temps mais cela pourrait juste être l'augmentation (ou diminution) de l'exposition totale du groupe correspondant au fil du temps.

On peut également regarder l'impact d'une attaque dans le groupe j sur le groupe i. Pour s'aider, la figure C.0.6 montre la forme des différents noyaux pour chaque groupe c'est-à-dire que le premier graphe montre $\phi_{11}, \phi_{12}, \phi_{13}, \phi_{14}$ et ϕ_{15} avec les paramètres estimés par maximum de vraisemblance.

Pour le groupe 1, une attaque dans le groupe 1 ou 5 impacte l'intensité du groupe fortement pendant 1 jour. Tandis qu'une attaque dans le groupe 3 ou 4 a un impact plus faible mais qui dure environ 2 jours. L'impact du groupe 2 est léger mais dure une dizaine de jours. Ce sont

des attaques dans le groupe 5 et 2 qui auront la plus grosse influence sur l'intensité du groupe 2. Notons que l'influence ne dure pas longtemps (environ 2 jours). L'intensité du groupe 3 est uniquement influencée par les attaques dans son propre groupe. De plus, l'influence d'une attaque est encore visible plus de 20 plus tard. Ce sont les attaques dans le groupe 5 et 4 qui auront la plus grosse influence sur l'intensité du groupe 4. Les autres groupes n'ont quasiment pas d'influence. Le groupe 5 n'est quasiment pas contaminé par une attaque dans un autre groupe. Une attaque dans le groupe 5 augmente fortement l'intensité mais cette influence dure moins de 0.2 jours. Une attaque dans le groupe 3 a une très légère influence sur le groupe 5. On remarque donc qu'un groupe i peut avoir une influence sur un groupe j mais pas inversément. L'effet auto-excitation est toujours intra-groupe c'est-à-dire qu'une attaque dans un groupe i influence de manière significative l'intensité du processus de son propre groupe. L'auto-excitation peut être inter groupe (une attaque dans un groupe i influence de manière significative l'intensité du processus d'un groupe j) mais c'est plus rare.

En pratique

Supposons que l'exposition de la base de données est constante dans le temps. Il est donc supposé que le nombre d'entreprises est le même entre 2010 et 2017 et que le nombre d'entreprises pour chaque groupe est également constant dans le temps tel que $n_1 = 2500, n_2 = 1000, n_3 = 4000, n_4 = 2000$ et $n_5 = 500$ où n_i est le nombre d'entreprises dans le groupe i.

L'assureur sera donc capable d'obtenir une fréquence de sinistres annuelle pour chaque catégorie d'assurés. En effet, à l'aide d'un grand nombre de simulations, il peut avoir une idée de la distribution de la fréquence pour chaque groupe. Cela revient donc à faire exactement comme sur la figure C.0.3 et en suite diviser l'axe x par le nombre d'entreprises dans le groupe et diviser l'axe y par le nombre de simulations.

Dans ce cas-ci, les fréquences de sinistres moyennes prédites pour l'année 2017 seraient :

groupe 1	groupe 2	groupe 3	groupe 4	groupe 5
4.72%	5.8%	0.7%	8.4%	2.8%

FIGURE 2.7.3 – Fréquence de sinistre annuelle moyenne pour chaque groupe.

Chapitre 3

La sévérité du risque cyber

Comme expliqué dans l'introduction et dans la deuxième section de ce chapitre, le risque cyber est un risque difficilement quantifiable. Par exemple, un vol de données, est-ce vraiment quantifiable ? Comment quantifier une dégradation de l'image d'une entreprise suite à un piratage de ses réseaux sociaux ? Comme détaillé dans la partie sur les polices d'assurance, la police d'assurance doit être précise. En effet, le risque cyber est un terme assez large. Il faut donc définir le périmètre de manière précise. Intuitivement, lorsqu'on parle de risque cyber, les gens pensent au hacking. Mais si on se base sur la définition 1.1.1 du risque cyber, cela comprend également la destruction d'appareils électroniques. Alors, est-ce que cette assurance va vraiment intervenir lorsqu'un employé dégrade une souris d'ordinateur par exemple ? L'assurance du risque cyber est une assurance où il y a plus un aspect service tout aussi important que l'aspect financier. L'assureur va fournir un service d'assistance (comme un call center, par exemple) lors de la survenance d'un sinistre. L'assureur peut également aider à récupérer des données volées et à construire un nouveau système d'information lorsqu'un vol a eu lieu. Il faut absolument le changer sinon les personnes malveillantes recommenceront puisqu'ils connaissent la faille. Toutes ces choses là, à savoir le prix d'une assistance à la crise, le prix du remplacement des serveurs ou le prix d'une récupération de données, sont spécifiques à chaque entreprise. En fonction de la taille de l'entreprise, une tarification au cas par cas peut s'avérer plus pertinente. L'objectif de cette section est d'utiliser une base de données fournies par SAS et d'étudier une manière de segmenter la queue de distribution comme dans l'article [9]. Cette section montrera que les variables qui font augmenter la moyenne de la partie centrale, qui est définit comme les sinistres en dessous du seuil de la loi de Pareto, de la distribution ne sont pas les mêmes variables que celles qui font augmenter la moyenne de la queue de la distribution. Cette partie mettra en évidence une caractéristique de la Pareto Généralisée : le principe de mutualisation n'est pas applicable sur la queue de distribution des risques avec une grande queue de distribution.

3.1 Données pour la modélisation

Dans ce chapitre sur la sévérité, les données qui seront utilisées sont des données fournies par SAS. Ce sont des données qui rassemblent des sinistres relatifs à différents risques opérationnels. Les risques qui dans le contexte de ce mémoire sont intéressants sont les risques liés aux évènements cyber. La base de données reprend 39 000 risques opérationnels. Après avoir trié les données qui nous intéressent, il reste 1571 données. C'est une base de données qui reprend seulement les gros risques opérationnels excédant 100 000 dollars. Le montant des sinistres a été ajusté en tenant compte de l'inflation sur base de l'indice des prix à la consommation. Comme expliqué dans l'introduction de ce chapitre, l'étude de la queue de distribution des montants de sinistre veut être réalisée. Il faut donc trouver le seuil à partir duquel il faut étudier en détail la distribution après ce seuil. Pour réaliser cela, il faudrait pouvoir faire une estimation de la fonction de distribution du coût de sinistre. Cependant, puisque les données sont sujettes à de la troncature, ce n'est pas vraiment possible de trouver une estimation.

Définition 3.1.1. [12] On dit qu'il y a troncature gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à T. T est alors la variable aléatoire de troncature à gauche. X est observée uniquement si $X \geqslant T$.

L'estimateur pour les données tronquées est le même que si les données n'étaient pas tronquées. Le détail se trouve dans l'annexe D. Pour la suite, il est difficile de pouvoir utiliser ces données pour un assureur mais cela pourrait très bien être le portefeuille d'un réassureur dont le seuil d'intervention commence à 100 000 dollars. L'objectif étant de mettre en avant la méthodologie plus que les résultats, les données seront utilisées telles quelles.

3.2 Les arbres pour la théorie des valeurs extrêmes

Cette section sera surtout consacrée sur l'aspect méthodologique plutôt que sur l'aspect résultat. Le but étant de vouloir étudier la sévérité et plus particulièrement la queue de la distribution de la sévérité du risque. L'idée générale de cette section est de montrer que les variables qui augmentent la moyenne de la partie centrale de la distribution ne sont pas les mêmes que celles qui augmentent la moyenne de la queue de distribution. Il faut étudier ces deux choses séparement. Etudier la queue de distribution est important dans le sens où les distributions avec une grande queue de distribution empêchent la mutualisation car la distribution peut prendre de très grandes valeurs avec une probabilité importante.

La distribution de Pareto Généralisé $\mathcal{GP}ar$ est une distribution qui est utilisée pour modéliser la distribution des valeurs extrêmes, les valeurs qui sont au delà d'un certain seuil. Cette distribution a deux paramètres le *scale* paramètre τ et le *shape* paramètre ξ. La fonction de distribution d'une

loi de Pareto Généralisée est donnée par

$$\mathcal{GP}ar(y;\tau,\xi) = \begin{cases} 1 - \left(1 + \xi\frac{y}{\tau}\right)_+^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\tau}\right) & \text{si } \xi = 0. \end{cases}$$

avec $y \in (0,\infty)$ si $\xi \geqslant 0$ et $y \in \left(0, -\frac{\tau}{\xi}\right)$. [7]

Le paramètre qui va être le plus important c'est le *shape* paramètre ξ. Il est également appelé *tail index* et c'est ce paramètre qui indique si la distribution a une queue de distribution *heavy*. Si $\xi > 0$, la distribution est *heavy tail*.

Proposition 3.2.1. *[7] Si $\xi < 1$ et $X \sim \mathcal{GP}ar(\tau,\xi)$ alors*

$$E[X] = \frac{\tau}{1-\xi}.$$

Cette proposition indique donc une manière de connaître la moyenne d'une distribution pour les valeurs qui dépassent le seuil.

Proposition 3.2.2. *[7] Si $\xi > 0$ et $X \sim \mathcal{GP}ar(\tau,\xi)$,*

$$E[X^k] = \infty \quad \textit{quand } k \geqslant \frac{1}{\xi}.$$

Cela signifie que si la sévérité du risque cyber a une distribution qui a un *tail index* ξ plus grand que 1, alors le risque devient "inassurable" en moyenne. Dans ce cas là, l'assureur sera dans l'obligation de mettre un montant maximum pour lequel il interviendra en cas de sinistre afin qu'il puissse répondre à ses obligations. De plus, il est également important de remarquer que par définition de la variance, pour un risque X, $Var[X] = \infty$ si $\xi \geqslant 0.5$. Il est donc assez intuitif que pour l'assureur, le *tail index* devra être inférieur à 0.5 afin qu'il puisse avoir une certaine maîtrise du risque. L'objectif de cette section est de montrer que les facteurs qui augmentent la moyenne de la partie centrale de la distribution ne sont pas les facteurs qui augmentent la moyenne de la queue de la distribution. Pour faire cela, nous allons faire deux arbres. Le premier sera un arbre de la moyenne pour la partie centrale de la distribution. Le deuxième sera construit sur les données qui excèderont le seuil et le critère pour chaque split sera la fonction de vraisemblance de la loi de Pareto Généralisée.

3.2.1 Construction d'un arbre

En R, il existe la fonction *rpart* qui permet de faire des arbres de régression. Cependant, cette fonction ne permet pas d'avoir un arbre pour la loi de Pareto Généralisée. Nous allons donc coder un arbre avec 3 split. On se contentera d'un arbre avec 3 split, et non d'un arbre avec une taille optimale.

Comment choisir pour chaque feuille le critère de *split* ? Le livre [8] et l'article [9] permettent de répondre à cette question. Supposons dans la feuille A d'un arbre qu'il y ait un ensemble $\mathcal{A}$ de données avec n variables. Il faut définir pour chaque arbre, une fonction d'erreur ϕ (moindre carré, déviance, fonction maximum de vraisemblance,...). Typiquement, si l'arbre est construit pour modéliser une moyenne, la fonction d'erreur ϕ sera égale à

$$\phi(x, \mathcal{A}) = \sum_{a \in \mathcal{A}} (x - a)^2.$$

C'est-à-dire que la moyenne pour la feuille A est égale à $x_A = \text{argmin}_x\ \phi(x, \mathcal{A})$. Pour trouver le critère de *split*, il faut que pour chaque variable d trouver les ensembles $\mathcal{B}^d$ et C^d qui maximisent la fonction

$$\phi(x_A, \mathcal{A}) - \phi(x_B, \mathcal{B}^d) - \phi(x_C, C^d) \tag{3.2.1}$$

tel que $\mathcal{B}^d \bigcup C^d = \mathcal{A}$ et $\mathcal{B}^d \bigcap C^d = \varnothing$. Enfin, on choisit la variable $d_{opt} \in \{1, \cdots, n\}$ qui maximise l'équation (3.2.1). La feuille A est donc divisée en deux feuilles B et C qui contiennent respectivement les données $\mathcal{B}^{d_{opt}}$ et $C^{d_{opt}}$. La fonction d'erreur pour l'arbre sur la loi de Pareto sera l'inverse de la fonction de log-vraisemblance. Pour rappel, la fonction de vraisemblance d'une Pareto Généralisée est égale à

$$\ln \mathcal{L}(u, \tau, \xi) = \sum_{i | y_i > u} -\ln \tau - \left(1 + \frac{1}{\xi}\right) \ln \left(1 + \frac{\xi}{\tau}(y_i - u)\right).$$

Il existe plusieurs méthodes pour trouver le seuil u à partir duquel la queue de distribution commence. Plusieurs méthodes sont proposées dans le livre [7]. Pour le cas ici présent, nous allons utiliser un outil graphique : le Gertensgarbe plot. L'idée principale de cette méthode est de se dire que la différence entre les écarts successifs des valeurs extrêmes n'est pas le même que l'écart successif des valeurs avant le seuil. Dans ce cas-ci, le seuil se trouve à 10.1 M$. Le seuil $u = 10.1$M$ ici représente le quantile 70% des données.

3.2.2 Application aux données

Pour étudier cette partie concernant la sévérité, il y aura 4 variables discrètes qui seront utilisées.

— La variable secteur d'activité qui est divisée en deux catégories : les entreprises liées au secteur financier et les entreprises avec un autre secteur d'activité.

— La taille de l'entreprise en fonction du nombre d'employés est divisée en trois catégories : petite (moins de 6000 employés), moyenne (entre 6000 et 50000 employés) et grandes (plus de 50000 employés). Chaque catégorie représente 1/3 du portefeuille.

— Le type d'attaque est divisé en trois catégories : fraude interne, fraude externe et interruption de business et défaillance du système.

— La région de l'entreprise sera divisée en quatre catégories : Etats-Unis, Asie, Europe et Autres (Afrique, Amérique du Nord et Amérique du Sud).

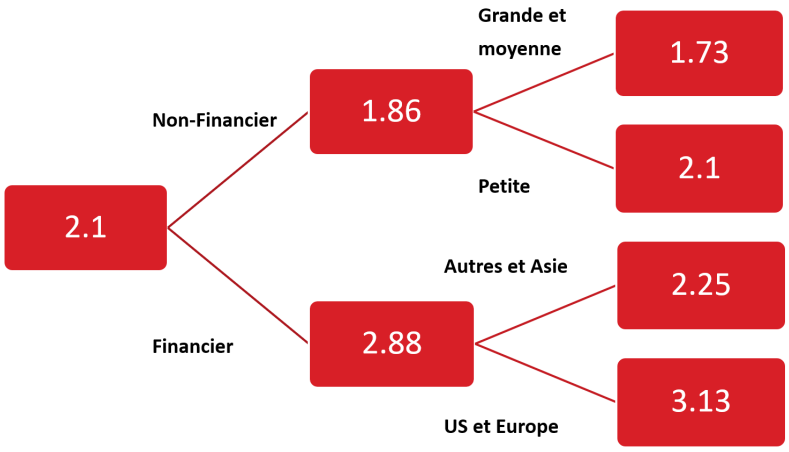

FIGURE 3.2.1 – Arbre de régression où le critère de chaque split est l'erreur quadratique. Cet arbre a été construit sur les données qui sont inférieures au seuil. Le nombre dans chaque feuille correspond à la moyenne des données présentes dans la feuille exprimées en millions de dollars.

La figure 3.2.1 exprime l'arbre qui représente la valeur moyenne de la partie centrale de la distribution c'est-à-dire la moyenne des valeurs qui se trouvent avant le seuil. La figure 3.2.2 est l'arbre pour la loi de la Pareto Généralisée pour les valeurs au-dessus du seuil. Les chiffres dans chaque feuille sont les paramètres estimés de la loi de Pareto Généralisée. Il y a deux enseignements à retenir de ces figures :

1. Les facteurs qui augmentent la moyenne de la partie centrale ne sont pas les mêmes que les facteurs qui augmentent la moyenne de la queue de la distribution. Par exemple, le secteur d'activité de l'entreprise semble jouer un rôle assez important sur la valeur moyenne de la partie centrale de la distribution. C'est le premier split qui est fait pour le premier arbre. A première vue, le type de fraude n'a pas vraiment d'importance sur la valeur moyenne de la partie centrale. Le fait que l'attaque soit une faute externe ou interne n'a pas vraiment de grande influence à première vue. Si on regarde désormais l'arbre concernant la queue de la distribution, on remarque que le facteur le plus important dans le premier arbre (le secteur d'activité de l'entreprise) ne semble pas avoir d'importance sur le tail index de la queue de distribution. Le facteur principal qui va rendre le risque cyber assurable ou non, c'est le type d'attaque. Si le type d'attaque est autre qu'une fraude externe, l'assureur devra mettre une limite d'intervention.
2. Faire une segmentation par rapport aux caractéristiques de l'entreprise permet de ne pas pénaliser l'ensemble du portefeuille à cause des entreprises qui ont un "mauvais" profil. Si une segmentation n'est pas réalisée pour la queue de distribution, tous les sinistres au dessus de 10.1M$ ne seraient pas indemnisés car le tail index est supérieur à 1. Il est égal à 1.27. Ca signifie que dans le cas de ce portefeuille par exemple 30% des sinistres n'auraient pas été couverts. Faire une segmentation juste par rapport au type d'attaque permettrait d'augmenter le nombre de couvertures. Le nombre de sinistres dans la base de données qui

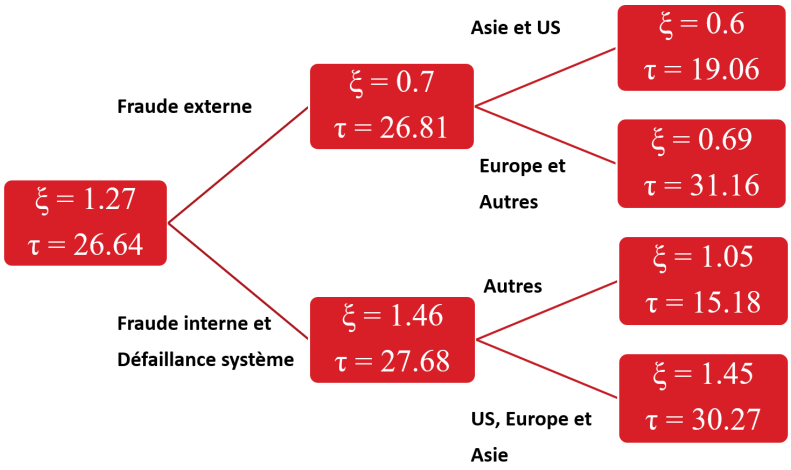

FIGURE 3.2.2 – Arbre de régression où le critère de chaque split est le maximum de vraisemblance d'une Pareto Généralisée. Cet arbre a été construit sur les données qui sont supérieures au seuil. Les nombres dans chaque feuille correspondent aux estimateurs des paramètres de la loi de Pareto Généralisée des données présentes dans la feuille exprimées en millions de dollars.

seraient couverts passerait de 70 à 80%. Cela met en lien une remarque faite lors d'une conférence sur le risque cyber de l'EAA [16].

Remarque 3.2.3. Si Y représente la variable qui modélise la sévérité d'un secteur A avec a un tail index ξ_Y et X représente la variable qui modélise la sévérité d'un secteur B avec a un tail index ξ_X, alors le tail index de la variable $W = \rho X + (1-\rho)Y$ a un tail index $\xi_W \approx \max(\xi_X, \xi_Y)$ avec ρ qui représente la proportion d'entreprise du secteur B dans le portefeuille. Cela montre qu'il n'y a pas de principe de mutualisation pour les queues de distribution. Ajouter des "bons" profils dans un portefeuille avec un *tail index* supérieur à 1 ne va pas permettre de diminuer le tail index de l'ensemble du portefeuille.

Exemple chiffré

Pour appuyer ce qui a été dit ici plus haut, simulons 10000 simulations d'une Pareto Généralisée pour la feuille avec le tail index le plus petit et celle avec le tail index le plus grand.

	Min.	Q.25%	Médiane	Moyenne	Q.75%	Max
$\xi = 0.6, \tau = 19.06$	10.01	16.06	26.20	50.37	53.69	1546.09
$\xi = 1.45, \tau = 30.27$	10.02	20.73	52.13	944.40	165.37	194243.64

Les résultats montrent clairement qu'il y a une énorme différence pour l'assureur entre ces deux distributions. Cela montre que l'assureur devra gérer son produit différemment en fonction du type de couverture (réassurance et/ou limite d'intervention).

Conclusion

Dans le premier chapitre, le risque cyber est présenté de manière assez détaillée. Il y est notamment listé une série de caractéristiques qui font que ce risque est particulier. Pour diverses raisons, il se différencie des risques classiques traités lors du master en sciences actuarielles. L'objectif de ce mémoire étant de modéliser le risque cyber, nous avons essayé de trouver des méthodes qui permettent de prendre en considération les caractéristiques du risque cyber. Pour commencer, il a fallu réfléchir à une modélisation qui soit possible avec le peu de données disponibles sur le risque cyber. Le princpal défaut des rares bases de données disponibles est qu'il n'y a pas d'indicateur par rapport à l'exposition au risque. Certaines propriétés du risque sont gérées par la police avec la présence d'une limite d'intervention, d'une franchise et des exclusions. D'autres peuvent être prises en considération dans le modèle.

Une caractéristique importante du risque cyber est la forte dépendance entre les polices d'un portefeuille. Une manière de modéliser cela pour un portefeuille est d'utiliser le processus de Hawkes qui est un processus de Poisson avec un effet auto-excitant afin de pouvoir prendre en considération cet aspect "contagion" lié au risque cyber. Evidemment, le modèle n'est pas parfait puisque l'exposition n'est pas incluse dans le processus, le nombre de paramètres pouvant très vite augmenter lorsque le portefeuille est finement segmenté. Il faut connaître la date précise du sinistre ce qui n'est pas toujours évident dans le cas du risque cyber. Le principal sujet de ce mémoire est le modèle de Hawkes appliqué au risque cyber. Il est important de noter que ce processus pourrait être utilisé dans un autre contexte que le risque cyber. Il pourrait être utilisé dans la modélisation d'une pandémie par exemple comme le Covid-19. L'utiliser dans ce contexte permettrait de ne plus avoir le problème lié à l'exposition puisque le nombre d'individus en Belgique ainsi que le nombre d'infections par jour sont des données facilement accessibles. Il est important de noter que le processus de Hawkes pourrait inclure l'exposition mais le problème est que les données ne fournissent pas cette information.

Le risque cyber étant un risque lié au numérique et à la technologie, il se voit naturellement évoluer de manière non négligeable. Cela se reflète dans les chiffres avec de plus en plus de sinistres avec le temps. Cette caractéristique fut ajoutée au modèle en introduisant une tendance linéaire dans la partie "baseline" du processus de Hawkes. Pour aller plus loin, il serait intéressant d'aller plus en profondeur sur cette caractéristique en analysant une tendance non-linéaire ainsi que de fournir une analyse plus appronfondie sur la fenêtre de calibration.

Concernant l'aspect sévérité du risque cyber, l'assureur devra découper son modèle en deux

parties. La première sera consacrée à la partie centrale de la distribution. La deuxième analysera la queue de la distribution. C'est en fait une pratique courante de séparer le portefeuille en petits (attritionnels) et grands (larges ou atypiques) sinistres. Ce n'est pas une obligation en soi, mais c'est important pour avoir une meilleure estimation de la sinistralité, les distributions étant assez différentes entre les petits et grands sinistres. En fonction des caractéristiques de l'entreprise, l'assureur devra tarifer son produit différemment. Il a été montré que certaines variables rendent le risque inassurable si aucune limite d'intervention n'est incluse dans la police d'assurance. Une tarification basée sur des variables explicatives peut ne pas être pertinente pour certaines entreprises étant donné que la nature du risque cyber peut changer une fois que l'entreprise est plus importante : le risque réputationnel commence à prendre une place plus importante et les sinistres cyber peuvent devenir plus complexes. Ce point a été confirmé par des intervenants externes, où un modèle « classique » est appliqué jusqu'à un certain montant de revenus annuels. Au-delà de ça, une tarification au cas par cas est appliquée. L'analyse faite pour la sévérité est donc principalement utile pour avoir une idée de l'étendue de la sévérité. Le manque de données limite une modélisation plus fine du risque. Le caractère évolutif de la sinistralité est un défi également : même si on disposait de nombreuses données, il faudrait prendre en compte le caractère évolutif dans la sinistralité, ce qui ajoute une difficulté.

L'article [17] propose des pistes pour étudier ce risque comme le modèle épidéomologique ou la théorie des jeux. Dans une suite à ce mémoire, il peut être intéressant de partir de cet article qui passe en revue la littérature sur les différentes méthodes pour étudier ce risque afin de pouvoir les étudier plus en profondeur.

En conclusion, le risque cyber est un risque pour lequel l'expertise est encore en développement. Les pistes d'exploration sont nombreuses mais le manque de données sur le sujet est un frein pour l'étude du risque cyber. L'exposition au risque et la valeur précise d'une attaque sont difficiles à évaluer. Ce mémoire propose une piste pour étudier la fréquence de ce risque à l'aide du processus de Hawkes et met en évidence la nécessité de faire une séparation dans le modèle entre l'étude de la partie centrale de la sévérité et la partie des valeurs extrême pour la sévérité. Les deux études méritent d'être étudiées plus en profondeur dans une suite à ce mémoire. Il peut être également intéressant de regarder s'il est possible de modéliser avec des méthodes plus classiques notamment pour les risques où il n'y a à priori pas de dépendance entre les polices (fraudes internes, dysfonctionnement de son propre système informatique,...) au lieu d'utiliser le processus de Hawkes qui comporte certains défauts.

Appendices

Annexe A

Calibration

Toujours dans l'objectif d'avoir une méthode qui performe numériquement avec un temps d'exécution le plus court possible, il est intéressant de calculer les gradients des différentes fonctions de vraisemblance. Il est en effet possible d'optimiser la fonction de vraisemblance (2.2.3) sans le gradient mais alors la méthode de calcul sera plus longue. Il faut connaître la dérivée de l'équation (2.2.3) par rapport à α, β et μ. L'objectif de ce travail n'étant pas dans le calcul de dérivée, le détail du calcul est fourni dans l'annexe. Notons qu'il faut que les fonctions de vraisemblence soient convexes pour pouvoir utiliser les gradients et ainsi utiliser une méthode d'optimisation par descente du gradient. Si la fonction de vraisemblance n'est pas convexe, le résultat va renvoyer le maximum le plus proche du point de départ de l'algorithme d'optimisation. Ce maximum ne sera pas nécessairemenet un maximum global.

Voici les différents résultats :

	μ	α	β
Paramètres pour la simulation	1	3	5
Paramètres estimés	1.036326	3.180150	5.157207

FIGURE A.0.1 – Paramètres pour une simulation avec le noyau exponentiel sur $[0, 1000]$.

Paramètres pour la simulation	0.6	2	4
Paramètres estimés	0.5358326	1.7822859	3.1456844

FIGURE A.0.2 – Paramètres pour une simulation avec le noyau d'Erlang sur $[0, 1000]$.

Paramètres pour la simulation	0.6	2	2
Paramètres estimés	0.5710866	2.0839318	2.1305062

FIGURE A.0.3 – Paramètres pour une simulation avec le noyau de Rayleigh sur $[0, 1000]$.

Concernant, l'optimisation de la fonction de maximum de vraisemblance est beaucoup plus rapide si une méthode par descente du gradient est utilisée. C'est donc ce qui a été utilisé pour vérifier que

les fonctions soient correctement codées. Cependant, le gain en temps de calcul a un coût. Ce coût est qu'il faut calculer le gradient de cette fonction de maximum de vraisemblance. La dimension du gradient est trois. Il faut donc dériver la fonction (2.2.3) par rapport à α, β et μ. Lorsqu'on voudra calibrer les paramètres pour une base de données, on n'utilisera pas les gradients pour les raisons citées plus haut.

Pour la simulation avec tous les noyaux qui sont des noyaux d'Erlang et un horizon de temps de 1500, les paramètres suivants sont utilisés :

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix},$$

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}.$$

Le résultat de l'estimation est le suivant :

$$\mu = \begin{bmatrix} 0.19 \\ 0.39 \end{bmatrix}, \quad \beta = \begin{bmatrix} 2.22 & 3.20 \\ 4.27 & 4.5 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 1.26 & 3.31 \\ 1.80 & 4.17 \end{bmatrix}.$$

De plus, si on exécute le test d'adéquation sur la simulation, les p-valeur sont excellentes pour la modélisation du premier groupe et du second groupe.

A.1 Détail des intégrales pour les différents noyaux.

Noyau exponentiel

Par définition,

$$\begin{aligned} \int_a^b \phi_{exp}(t)\mathrm{d}t &= \int_a^b \alpha \exp(-\beta t)\mathrm{d}t \\ &= \Big[\frac{-\alpha \exp(-\beta t)}{\beta}\Big]_a^b. \end{aligned}$$

Noyau d'Erlang

Par l'intégrale d'un produit,

$$\begin{aligned} \int_a^b \phi_{erl}(t)\mathrm{d}t &= \int_a^b \alpha t \exp(-\beta t)\mathrm{d}t \\ &= \Big[-\frac{\alpha t}{\beta}\exp(-\beta t)\Big]_a^b - \int_a^b -\frac{\alpha}{\beta}\exp(-\beta t)\mathrm{d}t \\ &= \Big[-\frac{\alpha}{\beta}\exp(-\beta t)(\frac{1}{\beta}+t)\Big]_a^b. \end{aligned}$$

Noyau de Rayleigh

En posant $u = t^2$,

$$\begin{aligned}\int_a^b \phi_{ray}(t)\mathrm{d}t &= \int_a^b \alpha t \exp(-\beta t^2)\mathrm{d}t \\ &= \int_{a^2}^{b^2} \frac{\alpha}{2}\exp(-\beta u)\mathrm{d}u \\ &= \Big[-\frac{\alpha}{2\beta}\exp(-\beta u)\Big]_{a^2}^{b^2}.\end{aligned}$$

A.2 Dérivées partielles par rapport à α et β des intégrales des noyaux.

Noyau exponentiel

$$\frac{\partial}{\partial\alpha}\Big(\frac{-\alpha\exp(-\beta t)}{\beta}\Big) = \frac{-\exp(-\beta t)}{\beta}$$
$$\frac{\partial}{\partial\beta}\Big(\frac{-\alpha\exp(-\beta t)}{\beta}\Big) = \frac{\beta\alpha t\exp(-\beta t)+\alpha t\exp(-\beta t)}{\beta^2}$$

Noyau d'Erlang

$$\frac{\partial}{\partial\alpha}\Big(-\frac{\alpha}{\beta}\exp(-\beta t)(\frac{1}{\beta}+t)\Big) = -\frac{1}{\beta}\exp(-\beta t)(\frac{1}{\beta}+t)$$
$$\frac{\partial}{\partial\beta}\Big(-\frac{\alpha}{\beta}\exp(-\beta t)(\frac{1}{\beta}+t)\Big) = \frac{\beta\alpha t^2\exp(-\beta t)+\alpha t\exp(-\beta t)}{\beta^2}$$

Noyau de Rayleigh

$$\frac{\partial}{\partial\alpha}\Big(\frac{-\alpha\exp(-\beta t)}{2\beta}\Big) = \frac{-\exp(-\beta t)}{2\beta}$$
$$\frac{\partial}{\partial\beta}\Big(\frac{-\alpha\exp(-\beta t)}{2\beta}\Big) = \frac{\alpha\exp(-\beta t)}{\beta^2}\Big[t+\frac{2}{\beta}+t^2\beta+t\Big]$$

Annexe B

Hawkes avec 3 groupes

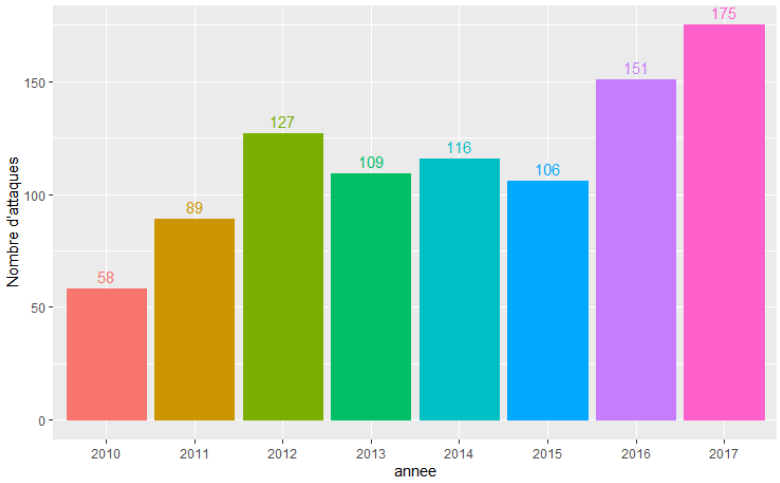

FIGURE B.0.1 – Nombre d'attaque pour la catégorie 1 pour chaque année entre 2010 et 2017.

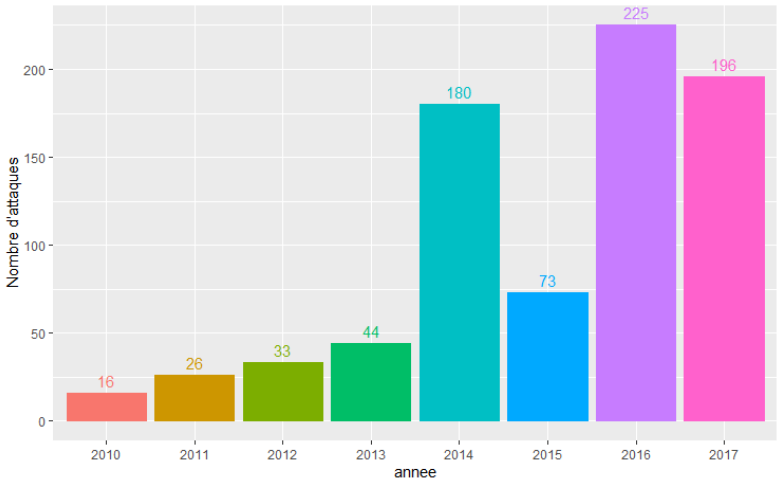

FIGURE B.0.2 – Nombre d'attaque pour la catégorie 2 pour chaque année entre 2010 et 2017.

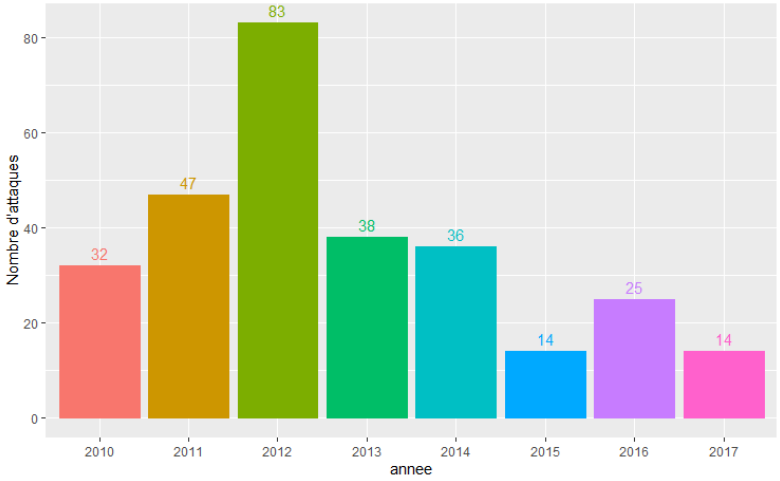

FIGURE B.0.3 – Nombre d'attaque pour la catégorie 3 pour chaque année entre 2010 et 2017.

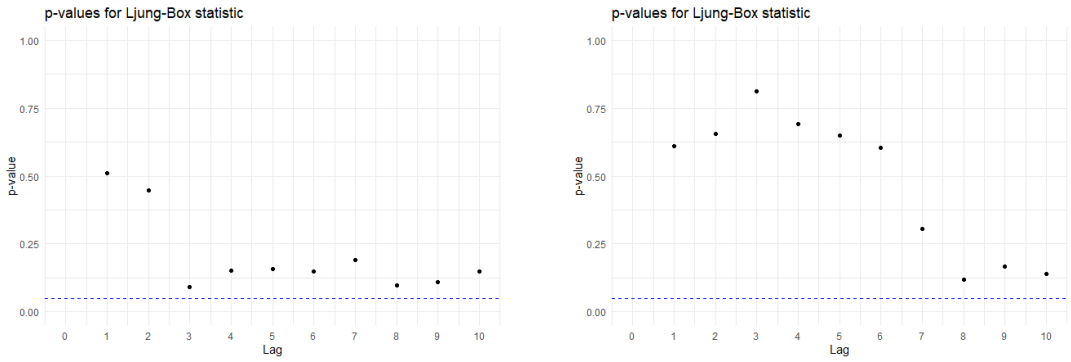

FIGURE B.0.4 – Les p-valeurs du test de Ljung-Box pour la première (à gauche) et la troisième (à droite) catégorie en fonction du lag en utilisant les données entre 2010 et 2016.

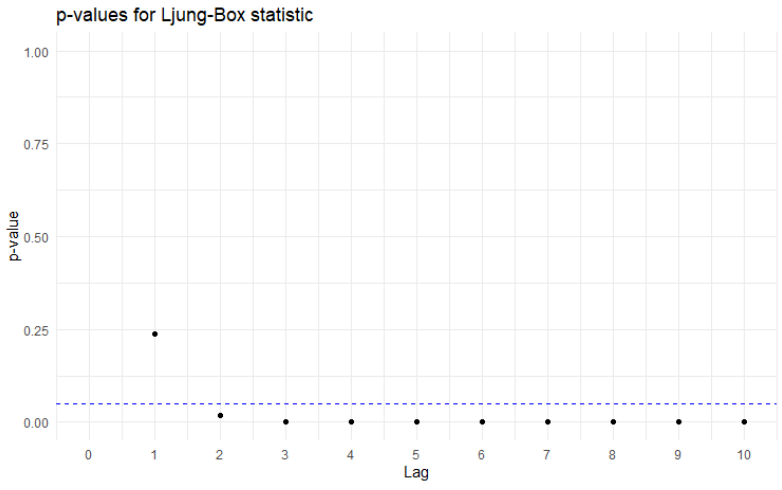

FIGURE B.0.5 – Les p-valeurs pour le test de Ljung-Box pour le noyau d'Erlang pour la catégorie 2 pour la calibration entre 2010 et 2016.

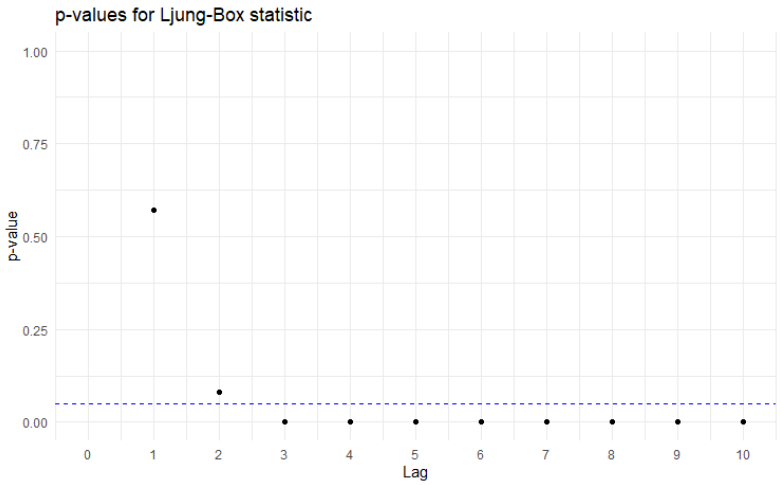

FIGURE B.0.6 – Les p-valeurs pour le test de Ljung-Box pour le noyau exponentiel pour la catégorie 2 pour la calibration entre 2010 et 2016.

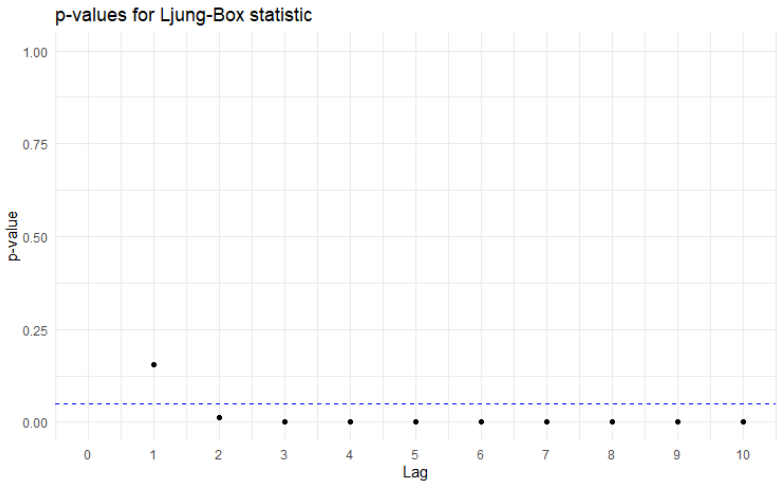

FIGURE B.0.7 – Les p-valeurs pour le test de Ljung-Box pour le noyau de Rayleigh pour la catégorie 2 pour la calibration entre 2010 et 2016.

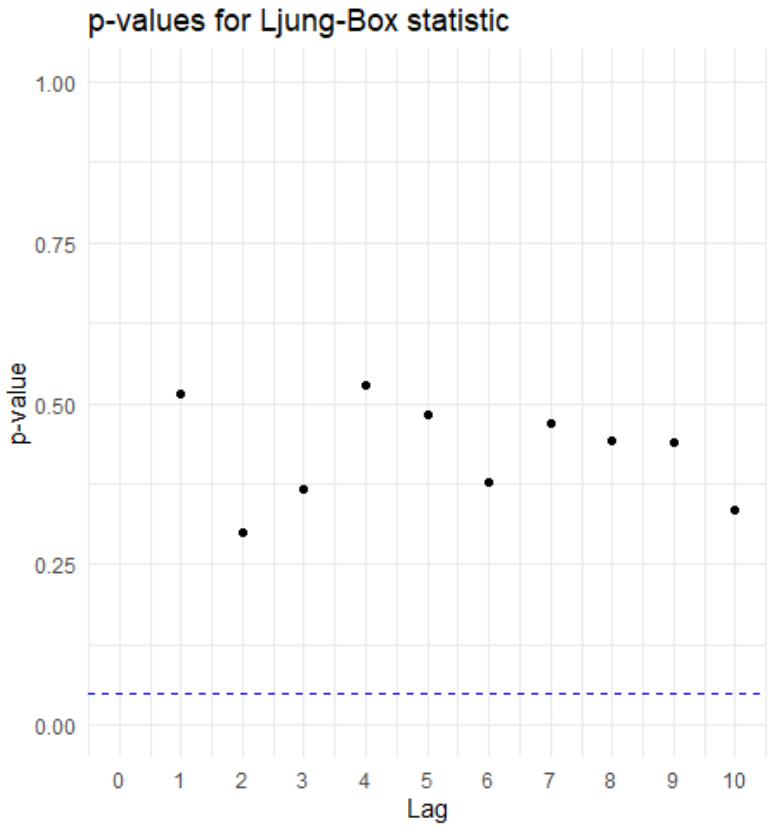

FIGURE B.0.8 – Les p-valeurs pour le test de Ljung-Box pour le noyau d'Erlang pour la catégorie 1 pour la calibration entre 2014 et 2016.

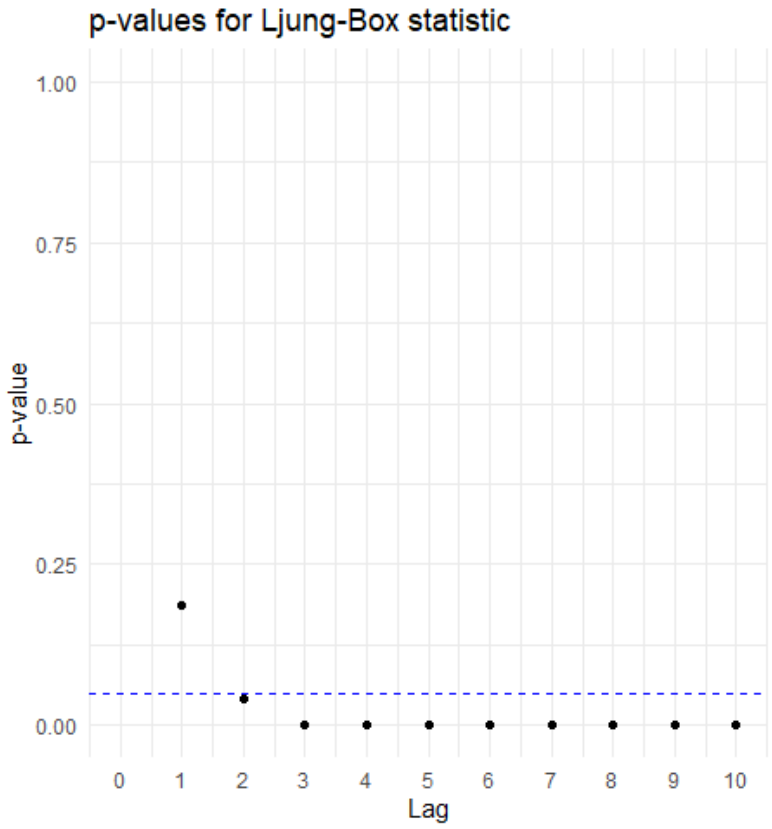

FIGURE B.0.9 – Les p-valeurs pour le test de Ljung-Box pour le noyau exponentiel pour la catégorie 2 pour la calibration entre 2014 et 2016.

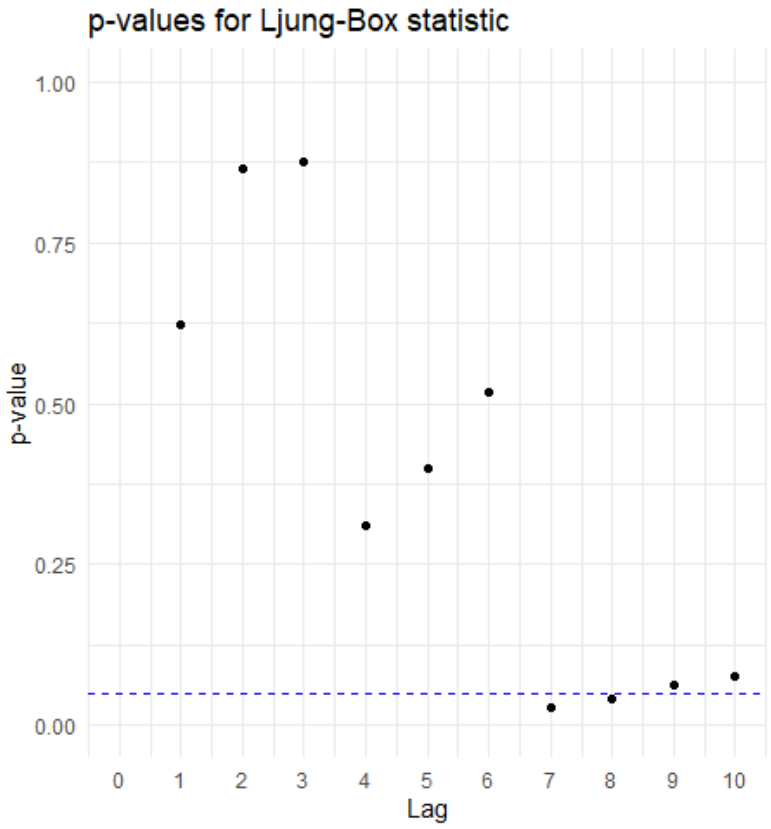

FIGURE B.0.10 – Les p-valeurs pour le test de Ljung-Box pour le noyau de Rayleigh pour la catégorie 3 pour la calibration entre 2014 et 2016.

α_{ij}	1	2	3
1	0.09310	0	0.02117
2	0	10.4168	13.3662
3	0.00325	0	0.05923

FIGURE B.0.11 – Estimation des α pour la calibration entre 2010 et 2016.

β_{ij}	1	2	3
1	0.53844	21.12098277	0.03777
2	0.57461	6.2126	1831.062
3	0.02545	0.56525	0.22514

FIGURE B.0.12 – Estimation des β pour la calibration entre 2010 et 2016.

	μ	γ
1	0.08005	0.03
2	0.01023	0.04638
3	0.08477	-0.01

FIGURE B.0.13 – Estimation des μ et γ pour la calibration entre 2010 et 2016.

Annexe C

Hawkes avec 5 groupes

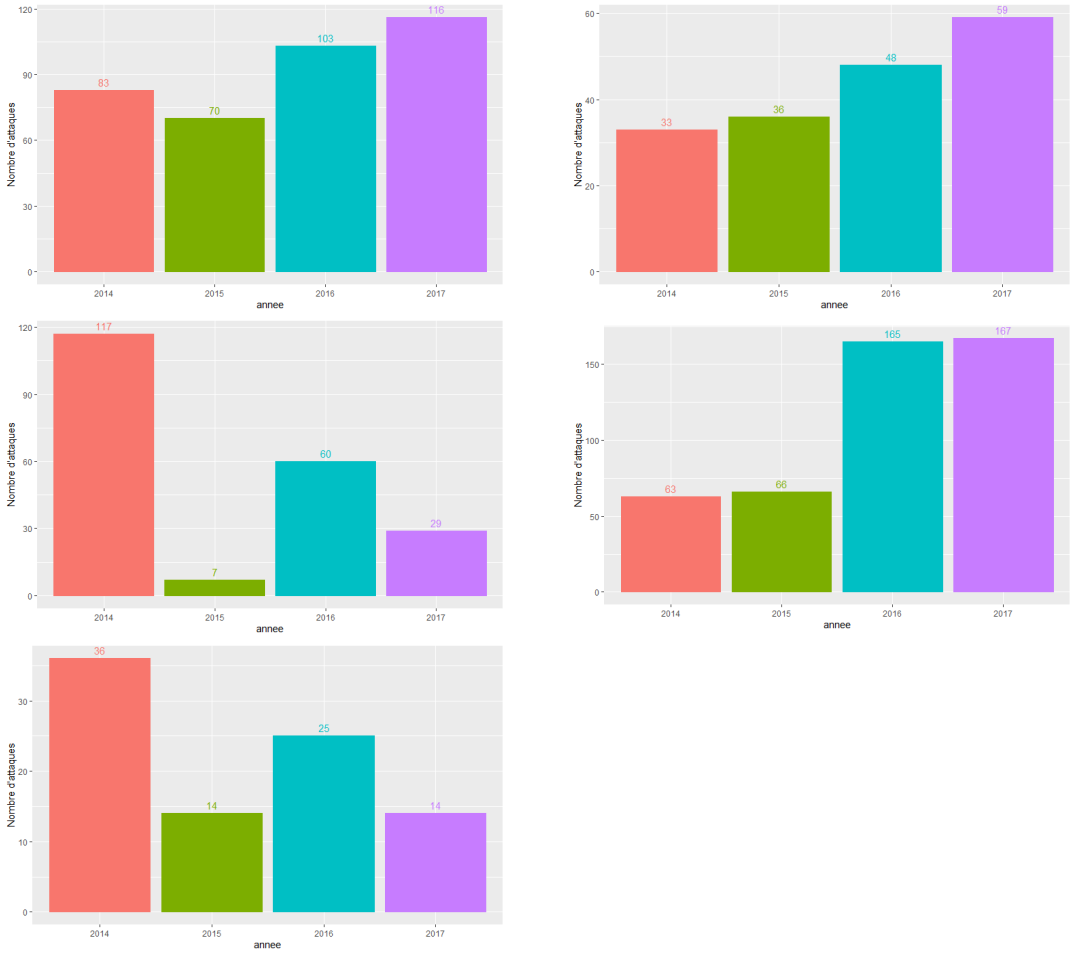

FIGURE C.0.1 – Evolution du nombre d'attaques dans les cinq groupes entre 2014 et 2017.

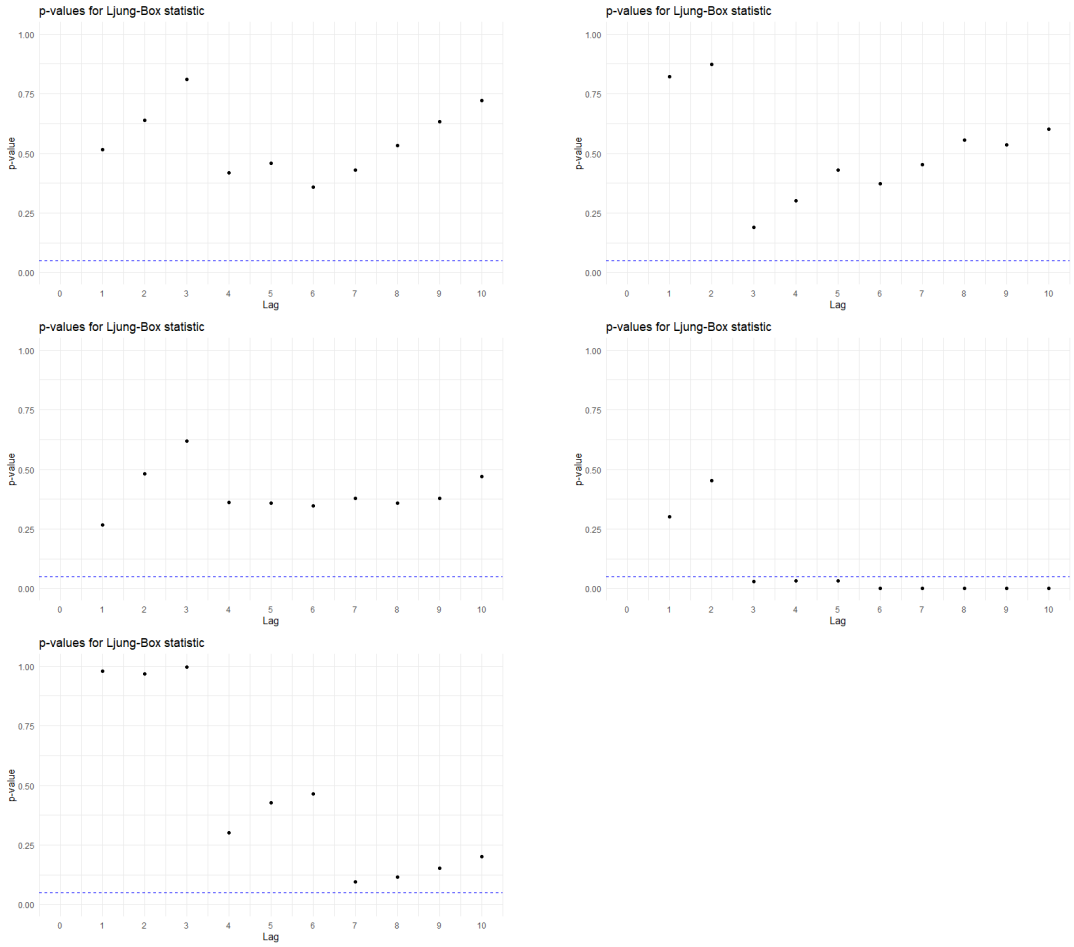

FIGURE C.0.2 – Les p-valeurs pour chaque groupe du test de Ljung-Box.

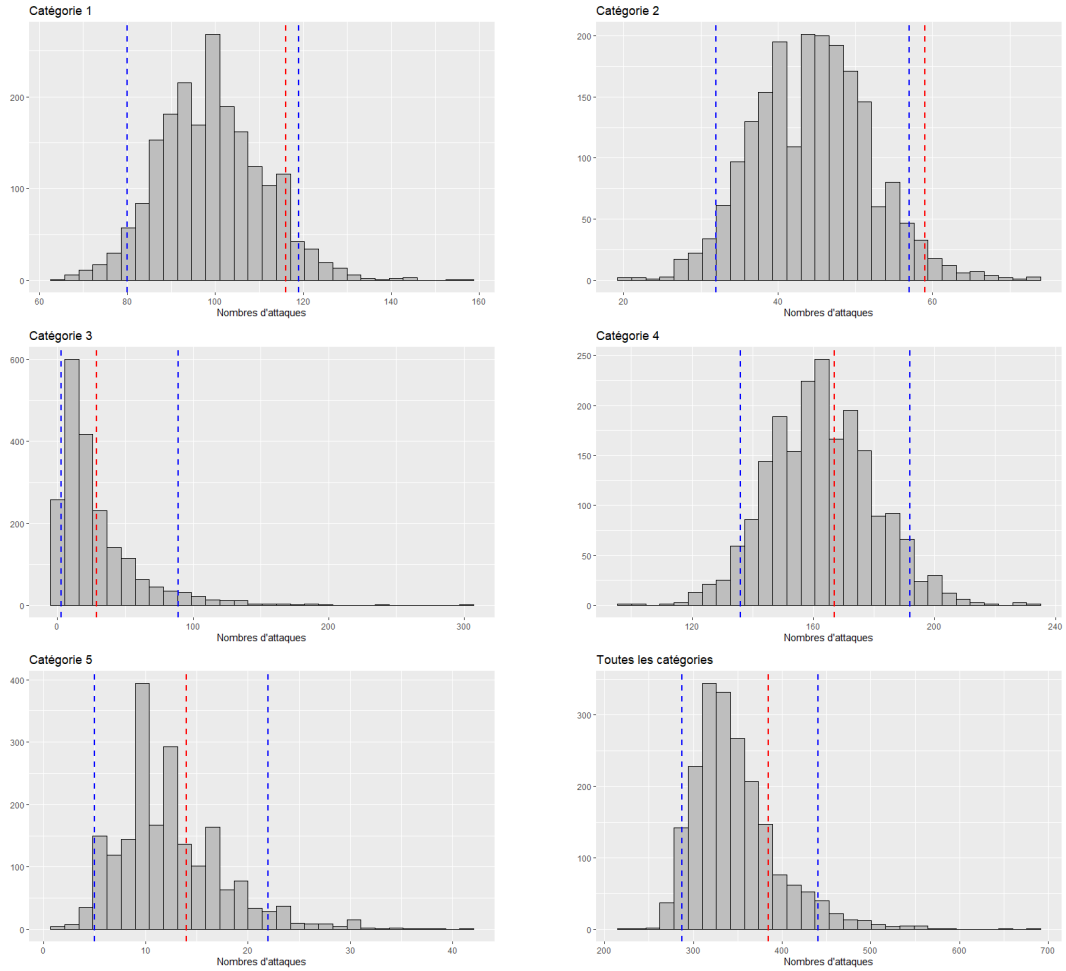

FIGURE C.0.3 – Comparaison entre les simulations du nombres de sinistres en 2017 et le nombre de sinistres observé. en rouge, le nombre de sinistres observé. En bleu, le quantile 5% et 95% des simulations.

α_{ij}	1	2	3	4	5
1	1.029	0.012	0.292	0.168	1.955
2	0.098	0.369	0.203	0.000	0.638
3	0.025	0.002	0.045	0.054	0.096
4	0.000	0.088	0.385	6.356	1.896
5	0.000	0.012	0.052	0.000	21.047

FIGURE C.0.4 – Estimation des α pour la calibration avec 5 groupes.

β_{ij}	1	2	3	4	5
1	7.540	0.045	1.327	0.814	15.732
2	1.801	1.837	13.019	40.748	4.603
3	2.337	0.071	0.051	4.697	7.660
4	16.220	2.746	2.568	4.972	3.345
5	2.496	18.405	0.614	0.162	20.051

FIGURE C.0.5 – Estimation des β pour la calibration avec 5 groupes.

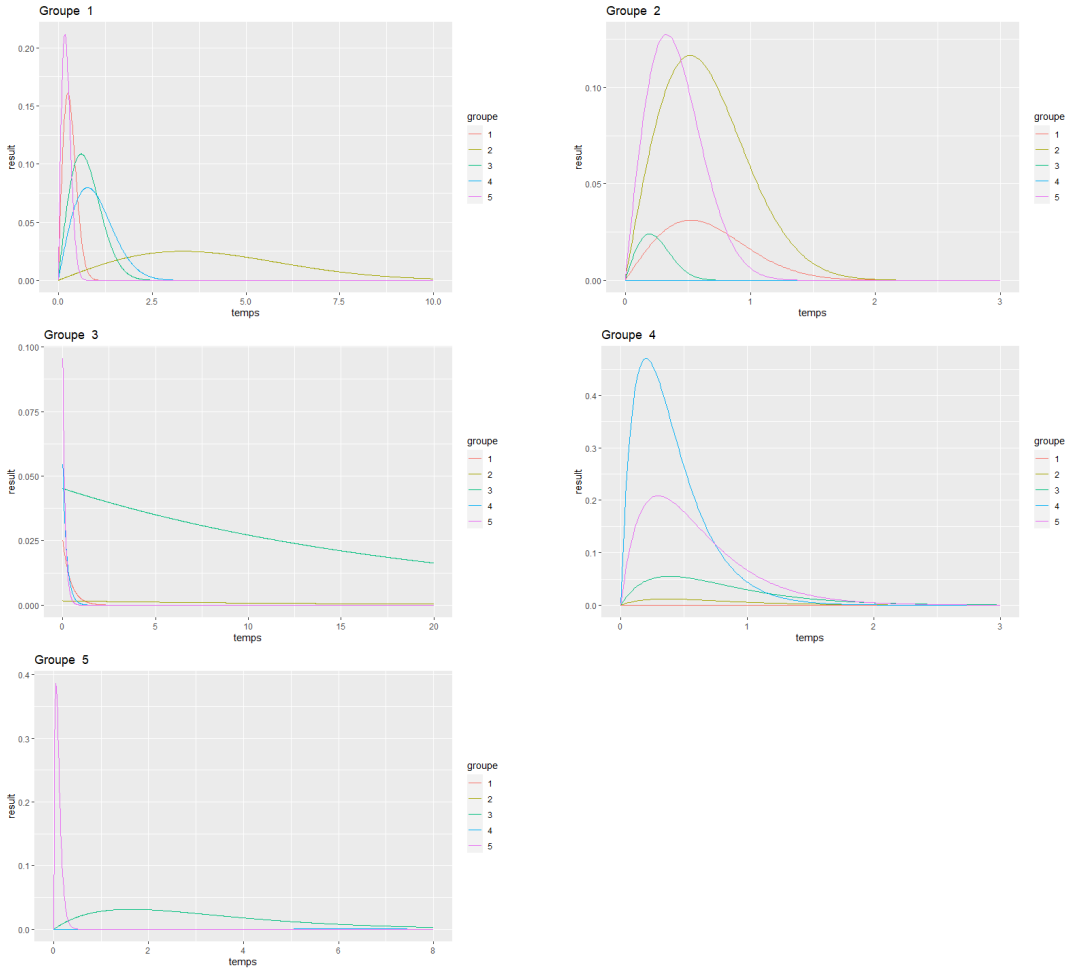

FIGURE C.0.6 – La forme des différents noyaux ϕ_{ij} en utilisant les estimateurs optimaux par maximum de vraisemblance.

Annexe D

Données tronquées

Pour un échantillon $(X_i, T_i)_{n=1,\cdots,n}$ de données tronquées (c'est-à-dire que $X_i > T_i$ pour tout $i \in \{1, \cdots, n\}$), l'estimateur de la fonction de répartition [12] de la variable aléatoire X est donné par

$$F_n(x) = 1 - \prod_{i|X_i \leqslant x} \Big(\frac{nC_n(X_i) - 1}{nC_n(X_i)} \Big)$$

avec

$$C_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{T_i \leqslant x \leqslant X_i\}}.$$

Puisque dans ce cas-ci $T_i = 100000$ pour tout $i \in \{1, \cdots, n\}$, alors $C_n(X_j) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_j \leqslant X_i\}}$ pour tout $j \in \{1, \cdots, n\}$. Dès lors,

$$F_n(x) = 1 - \prod_{i|X_i \leqslant x} \Big(\frac{\sum_{j=1}^{n} 1_{\{X_i \leqslant X_j\}} - 1}{\sum_{j=1}^{n} 1_{\{X_i \leqslant X_j\}}} \Big) = \frac{\sum_{j=1}^{n} 1_{\{X_j \leqslant x\}}}{n}.$$

Bibliographie

[1] R. Böhme et G. Kataria, « Models and measures for correlation in cyber-insurance., » in *WEIS*, t. 2, 2006, p. 3.

[2] J. L. Cebula et L. R. Young, « A taxonomy of operational cyber security risks, » Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, rapp. tech., 2010.

[3] F. Pons, « Etude actuarielle du cyber-risque, » *Mémoire, Institut des actuaires*, 2014.

[4] M. Eling et W. Schnell, « Ten key questions on cyber risk and cyber risk insurance, » *The Geneva Association—'International Association for the Study of Insurance Economics'*, 2016.

[5] M. Eling et W. Schnell, « What do we know about cyber risk and cyber risk insurance ? » *The Journal of Risk Finance*, t. 17, n° 5, p. 474-491, 2016.

[6] Y. Bessy-Roland, « Titre : Modélisation stochastique individuelle de sinistres cyber, » Université de Bretagne Occidental, 2019.

[7] M. Denuit, D. Hainaut et J. Trufin, « Extreme Value Models, » in *Effective Statistical Learning Methods for Actuaries I : GLMs and Extensions*, Cham : Springer International Publishing, 2019, p. 401-441.

[8] M. Denuit, D. Hainaut et J. Trufin, « Regression Tree, » in *Effective Statistical Learning Methods for Actuaries II : Tree-Based Methods and Extensions*, Cham : Springer International Publishing, 2020, p. 107-130, isbn : 978-3-030-57556-4.

[9] S. Farkas, O. Lopez et M. Thomas, « Cyber claim analysis through generalized Pareto regression trees with applications to insurance, » 2020.

[10] Y. Bessy-Roland, A. Boumezoued et C. Hillairet, « Multivariate Hawkes process for cyber insurance, » *Annals of Actuarial Science*, t. 15, n° 1, p. 14-39, 2021.

[11] P. J. Laub, Y. Lee et T. Taimre, *The elements of hawkes processes*. Springer, 2021.

[12] H. Boudada et S. Leulmi, « Estimation non paramétrique pour des données fonctionnelles et tronquées., » thèse de doct., Université Frères Mentouri-Constantine 1, 2022.

[13] « Cyber Risk Toolkit, » American Academy of Actuaries. (février 2022), adresse : https://www.actuary.org/cybertoolkit (visité le 06/04/2023).

[14] E. V. Dubois, O. F. Keskin et U. Tatar. « What do we know about cyber risk and cyber risk insurance ? » SOA Research Institute. (sept. 2022), adresse : https://www.soa.org/4a81c2/globalassets/assets/files/resources/research-report/2022/cyber-risk-modeling.pdf (visité le 06/04/2023).

[15] C. Hillairet, O. Lopez, L. d'Oultremont et B. Spoorenberg, « Cyber-contagion model with network structure applied to insurance, » *Insurance : Mathematics and Economics*, t. 107, p. 88-101, 2022.

[16] O. Lopez et C. Hillairet, « Numerical tools for cyber insurance, » European Actuarial Academy, nov. 2022.

[17] K. Awiszus, T. Knispel, I. Penner, G. Svindland, A. Voß et S. Weber, « Modeling and pricing cyber insurance : Idiosyncratic, systematic, and systemic risks, » *European Actuarial Journal*, p. 1-53, 2023.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**