# Mathématiques appliquées à l'assurance des risques numériques

Sébastien Farkas

# Manuscrit de doctorat

Discipline : Mathématiques appliquées

Université

**Sorbonne Université**

École Doctorale

**Sciences Mathématiques de Paris-Centre**

Unité Mixte de Recherche

**Laboratoire de Probabilités, Statistique et Modélisation**

Thèse défendue par

**Sébastien, Charles-Jean, Etienne Farkas**

---

# Mathématiques appliquées à l'assurance des risques numériques

---

Composition du jury de la soutenance du 4 décembre 2023

| | | |
|---|---|---|
| *Président du jury* | Christian-Yann Robert | professeur à l'université Claude Bernard - Lyon 1 |
| *Examinateur* | Idris Kharroubi | professeur à Sorbonne Université |
| *Rapporteur* | Donatien Hainaut | professeur à l'université catholique de Louvain |
| *Rapporteur* | Gilles Stupfler | professeur à l'université d'Angers |
| *Directeur de thèse* | Olivier Lopez | professeur à Sorbonne Université |
| *Directrice de thèse* | Caroline Hillairet | professeure à l'ENSAE Paris |

**Résumé**

L'émergence des produits d'assurance couvrant les risques numériques s'accompagne d'interrogations relatives à la maîtrise des engagements souscrits par les organismes d'assurance. La volatilité des coûts, la dépendance entre les garanties et les potentielles accumulations de sinistres sont autant de spécificités que nous considérons pour proposer des modèles mathématiques adaptés aux enjeux. Nous introduisons d'abord des arbres de régression permettant de comprendre l'hétérogénéité des queues de distribution des risques. Ensuite, nous étudions l'estimation de copules dans un contexte de données censurées pour préciser l'impact des interactions entre les garanties sur les engagements globaux. Enfin, nous proposons une analyse de la fréquence des sinistres numériques par des processus ponctuels adaptés aux phénomènes d'accumulation. Nos contributions suggèrent des méthodes d'analyse pour la souscription, le provisionnement et la gestion des risques numériques.

Mots clés

**Mathématiques appliquées** ; **Statistique** ; **Science actuarielle** ;
**Assurance des risques numériques** ; **Arbres de régression** ; **Valeurs extrêmes** ;
**Copules** ; **Données censurées** ; **Processus ponctuels** ; **Accumulation**

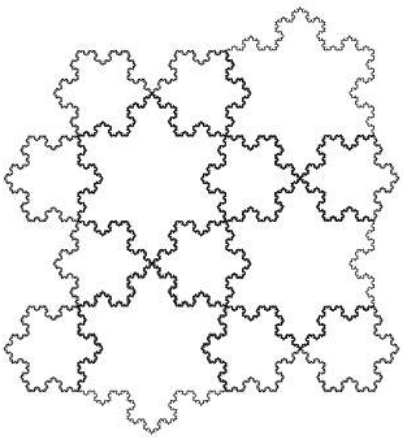
---

**Summary**

The rise of cyber insurance products challenges the quantification of related liabilities for insurers. The volatility of costs, the dependence between covers and the potential concentration of claims are the specific features that we tackle with the aim to propose tailored mathematical models. We first introduce regression trees adapted to extreme values in order to understand the heterogeneity in distribution tails. Next, we study the estimation of copulas in a censored data context to clarify the impact of covers interactions on overall commitments. Finally, we propose an analysis of the frequency of cyber claims using point processes adapted to accumulation phenomena. Our contributions suggest analytical methods for the underwriting, the reserving and the management of numerical risks.

Key words

**Applied mathematics** ; **Statistics** ; **Actuarial science** ;
**Cyber risk insurance** ; **Regression trees** ; **Extreme values** ;
**Copula** ; **Censored data** ; **Point processes** ; **Accumulation**



Laboratoire de Probabilités,
Statistique et Modélisation

Case courrier 158
4, place Jussieu
75 252 Paris cedex 05

École Doctorale Paris centre

Case courrier 158
4, place Jussieu
75 252 Paris cedex 05

# Table des matières

# Introduction

## Les mathématiques appliquées à l'assurance

L'assurance consiste à offrir une garantie financière créditée uniquement lors de l'occurrence d'un événement couvert par le contrat, en échange d'un certain prix appelé prime. Au moment de la vente, la réalisation de tels événements et leurs coûts sont incertains. Autrement dit, le coût de revient d'un contrat d'assurance est aléatoire. Cette spécificité du secteur de l'assurance, appelée cycle inversé de production, est au centre des réflexions tout au long de la chaîne de valeur d'un produit d'assurance (cf. Suru, 2020). L'inversion du cycle de production s'observe partiellement au sein d'autres secteurs d'activités. Les jeux de hasard ou les paris sportifs en sont un exemple. Chaque issue de vente est en effet aléatoire. Cependant, le bookmaker est très informé sur le phénomène aléatoire, les gains associés à chaque issue sont par ailleurs souvent déterministes et, une fois les dés lancés, les clients n'ont plus la possibilité d'influencer le résultat : en France, la quasi-totalité des acteurs participant ou étant en relation avec les compétitions sportives ne peuvent en effet parier. À l'inverse, l'assuré est plus proche de l'objet du contrat que l'assureur, auprès duquel il partage uniquement les informations requises à la souscription du contrat. Or, cette asymétrie d'information complexifie la maîtrise de l'aléa et crée un risque d'anti-sélection pour l'assureur. Par ailleurs, l'assuré reste en interaction avec l'objet du contrat après sa signature et peut ainsi modifier son comportement, créant un risque d'aléa moral pour l'assureur. De plus, l'engagement de l'assureur suite à un événement couvert de manière indemnitaire est aléatoire dans la limite des plafonds des garanties contractuelles et ajoute ainsi au risque de fréquence un risque de sévérité. Enfin, certains événements peuvent affecter au même moment une proportion matérielle d'un portefeuille de clients, créant un risque d'accumulation. Ainsi, si comme pour les jeux de hasard et les paris sportifs la mutualisation des ventes est le premier levier permettant une lisibilité de l'activité, la recherche d'un coût juste du risque se confronte à des problématiques singulières en

assurance. L'analyse attentive des données historiques permet d'apporter des éléments de réponse en supposant une certaine stabilité des risques dans le temps.

La statistique est précisément un domaine des mathématiques dont l'objectif est d'estimer des caractéristiques de phénomènes supposés aléatoires à partir d'observations issues de l'expérience. L'analyse des données peut ainsi préciser le comportement d'aléas présentant un intérêt. Prenons l'exemple de l'engagement associé à un risque assuré, et supposons qu'il définisse une variable aléatoire notée $Y$ suivant la mesure de probabilité $\mathbb{P}_Y^*$. L'espérance de $Y$, notée $\pi^* = \mathbb{E}[Y] = \int_y d\mathbb{P}_Y^*(l)$, caractérise le comportement moyen de l'engagement et représente ainsi une propriété essentielle de la loi $\mathbb{P}_Y^*$ dans le cadre de la définition du prix technique annuel $\tilde{\pi}$ du produit d'assurance associé. Son caractère fini apparaît par exemple comme nécessaire à la souscription du risque. Le prix technique annuel $\tilde{\pi}$ ne peut cependant pas être égal à $\pi^*$, d'une part parce que cette espérance est inconnue et d'autre part parce que la sinistralité est aléatoire et peut dévier de son espérance. La variance de $Y$, notée $v^* = \mathbb{V}[Y] = \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right]$, quantifie précisément la déviation quadratique d'une réalisation de $Y$ par rapport à son espérance, et permet ainsi de maîtriser la déviation de sinistralité des risques souscrits.

Grâce à son expérience et sur des risques homogènes, l'assureur dispose de coûts individuels $(Y_i)_{1\leq i\leq n}$ qui sont généralement supposés indépendants, identiquement distribués et avec une variance finie. Notons $\hat{\pi}$ la moyenne empirique des coûts. Le théorème central limite (TCL) et la méthode delta permettent de délimiter un intervalle de confiance asymptotique de niveau $\alpha \in ]0;1[$, dans lequel la probabilité que $\pi^*$ se situe est supérieure à $\alpha$ lorsque le nombre de données $n$ tend vers l'infini :

$$IC_\pi^\alpha = \left[\hat{\pi} - q_{\frac{1+\alpha}{2}}^{\mathcal{N}(0;1)} \frac{\sqrt{\hat{v}}}{\sqrt{n}}; \hat{\pi} + q_{\frac{1+\alpha}{2}}^{\mathcal{N}(0;1)} \frac{\sqrt{\hat{v}}}{\sqrt{n}}\right],$$

$$\text{avec } \hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i \ , \ \hat{v} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\pi})^2,$$

et où $q_{\frac{1+\alpha}{2}}^{\mathcal{N}(0;1)}$ représente le quantile $\frac{1+\alpha}{2}$ d'une loi normale centrée réduite $\mathcal{N}(0;1)$, vérifiant pour une variable aléatoire $X \sim \mathcal{N}(0;1)$, $\mathbb{P}\left(|X| < q_{\frac{1+\alpha}{2}}^{\mathcal{N}(0;1)}\right) = \alpha$.

A l'échelle d'un portefeuille de $n$ contrats identiques, le volume global des primes doit couvrir une déviation par rapport à l'attendu $n \times \pi^*$. La quantification de ce risque repose sur l'intervalle de fluctuation d'ordre $\alpha$ de la perte empirique du portefeuille $\sum_{i=1}^n Y_i$, exprimé à partir des quantiles d'ordre $\alpha$ de la loi de $\sum_{i=1}^n Y_i$, notés $q_\alpha\left(\sum_{i=1}^n Y_i\right)$ :

$$IF^{\alpha}_{\sum_{i=1}^{n} Y_i} = \left[ q_{\frac{1-\alpha}{2}} \left( \sum_{i=1}^{n} Y_i \right) ; q_{\frac{1+\alpha}{2}} \left( \sum_{i=1}^{n} Y_i \right) \right].$$

Or, la connaissance de ces quantiles nécessite l'entière connaissance du phénomène aléatoire $Y$, à savoir de sa distribution $\mathbb{P}^*_Y$. Par exemple, sous une condition d'indépendance des $Y_i$, les quantiles sont définis à partir de la résultante de $m$ produits de convolution de la densité de $\mathbb{P}^*_Y$. La recherche de propriétés spécifiques d'un phénomène aléatoire comme son espérance n'est donc pas suffisante, l'objectif d'un assureur est plutôt la maîtrise globale des phénomènes aléatoires.

Les mathématiques offrent par ailleurs d'autres applications à l'assurance, par exemple en provisionnement lorsque l'objectif est de valoriser les engagements liés à un portefeuille des sinistres non encore clôturés ou bien en gestion de risques lorsque l'objectif est de quantifier les déviations potentielles des risques par rapport à l'attendu. Sous ce prisme, notre propos se focalise sur les offres d'assurances relatives aux risques numériques proposées aux organisations, dont le développement est relativement récent. Ces solutions reposent sur les garanties classiques de responsabilité civile et de dommages aux biens et prennent en charge les préjudices causés par des attaques numériques. Pour comprendre ce marché émergent, nous rappellerons d'abord quelques faits marquants du développement du numérique dans notre société et introduisons ensuite les risques numériques induits par ces usages. Puis, nous proposerons une analyse du marché de l'assurance des risques numériques et terminerons par l'identification des enjeux mathématiques étudiés dans le cadre de ce manuscrit.

# Les risques numériques

Depuis quelques dizaines d'années, la croissance du numérique au sein de nos modes de vie est une constante qui en fait aujourd'hui un maillon presque systématique. Plusieurs définitions existent et pour préciser notre propos nous citons la définition de Didier (2019). *"Le numérique représente toutes les applications qui utilisent un langage binaire qui classe, trie et diffuse des données [binaires]. Ce terme englobe les interfaces, smartphones, tablettes, ordinateurs, téléviseurs, ainsi que les réseaux qui transportent les données. Il envisage à la fois les outils, les contenus et les usages."* En France, les usages individuels du numérique sont mesurés dès l'année 2000 par le baromètre annuel et gouvernemental du numérique, dont les données sont ouvertes et publiées par l'Autorité de régulation des communications électroniques (2022). On peut ainsi confirmer la croissance

du numérique dans les modes de vie par la dynamique du temps quotidien passé à regarder un écran, culminant en 2022 à plus de quatre heures et demie. Se passer des outils numériques génère en outre une sensation de manque, dès les premières heures sans utilisation pour un tiers des sondés. Au delà de nos frontières, l'accès au réseau internet est encore en pleine croissance et les usages du numérique suivront vraisemblablement cette tendance. Selon l'International Telecommunication Union (2022), cette croissance a été de 17% entre 2019 et 2021, un record. 63% de la population mondiale avait ainsi accès à internet en 2021 et le développement récent d'offres d'accès à internet par des réseaux de satellites en orbite basse devrait à l'avenir soutenir l'accès mondial en connectant des zones peu couvertes par les réseaux terrestres.

Pour les organisations, le numérique représente à la fois un marché, du fait d'un nouvel espace économique, mais également une opportunité d'amélioration des processus de gestion et d'administration. Le numérique s'est ainsi imposé comme l'un des jalons essentiels à l'activité de la plupart des organisations. Comme toute évolution, ce développement s'accompagne d'effets collatéraux, en l'espèce de l'émergence des risques numériques. Notre propos se focalise principalement sur la criminalité numérique dont les principaux objectifs sont selon l'Agence nationale de la sécurité des systèmes d'information (2022), les gains financiers, l'espionnage et la déstabilisation. Les données gouvernementales sur la criminalité numérique sont encore parcellaires mais ce phénomène est en augmentation. Le nombre de dossiers suivis par l'European Union Agency for Criminal Justice Cooperation en lien avec la criminalité numérique a presque doublé entre 2018 et 2021. Par ailleurs, à l'échelle de l'Union Européenne et en 2021, 28% des petites et moyennes entreprises en ont été victimes au moins une fois (cf. Eurobarometer, 2022). Si les profils des individus à l'origine de cette criminalité semblent être stables selon Didier Parsoire (2017), les stratégies mises en oeuvre sont multiples, évolutives et parfois personnalisées. A titre d'illustration, nous présentons ci-dessous cinq types d'attaques.

- L'extorsion au moyen de rançongiciels paralysant tout ou partie de l'activité de la victime et invitant au règlement d'une rançon. Par exemple, en 2021, le rançongiciel Wannacry infecte en l'espace d'une journée au moins 200 000 machines dans plus de 150 pays. Les conséquences pour le système de santé anglais sont notamment évaluées à 100 millions d'euros selon l'Agence nationale de la sécurité des systèmes d'information (2020).
- L'intrusion reposant sur des vulnérabilités informatiques récentes, dites du jour zéro et n'ayant pas encore fait l'objet d'un correctif. Par exemple, en 2021, le service de

messagerie électronique de Microsoft a été perturbé pendant une vingtaine de jours, impactant plus de 60 000 entreprises. Les motivations semblent dans ce cas étatiques (cf. Kanno-Youngs & Sanger, 2021).

- La déstabilisation d'un système informatique par une attaque par déni de service (DDoS) consistant à le saturer de requêtes, le rendant ainsi indisponible aux requêtes légitimes. Par exemple, en 2023, l'entreprise Cloudflare, proposant des services de sécurité, de performance et de fiabilité pour les sites internet, a fait l'objet d'un nombre record de 71 millions de requêtes par seconde. Pour autant, cette organisation est parvenue à les supporter, sensibilisant ainsi d'autres entreprises à l'efficacité de la prévention des risques numériques (cf. Omer Yoachimik, 2021),
- Le vol de données personnelles ou financières, donnant la possibilité de réaliser des actions malveillantes a posteriori. Par exemple, en 2016, les données de plus d'un milliard de comptes de la messagerie électronique Yahoo ! ont été subtilisées, diminuant la valorisation de l'entreprise de 350 millions de dollars. Par la suite, les autorités ont condamné l'entreprise à verser 35 millions de dollars pour avoir trompé le public et omis d'informer les clients de la violation (cf. Reynaud, 2016),
- L'espionnage ou le sabotage de procédés technologiques. Par exemple le ver informatique Stuxnet, découvert en 2010, a été conçu par la National Security Agency pour s'attaquer aux centrifugeuses iraniennes d'enrichissement d'uranium. Le numérique apparaît alors comme un terrain possible de guerre (cf. Sanger, 2012).

En outre, la pression sur la victime peut être accentuée par la combinaison de ces différentes techniques selon Gilles Bénéplanc (2018). Les conséquences de la criminalité numérique ne se limitent toutefois pas à la valeur des informations dérobées : la gestion des dommages collatéraux représente l'essentiel du coût pour les personnes morales impactées. La gestion de crise, les investigations, la sécurisation des systèmes d'information, la perte d'exploitation, les dommages matériels ou corporels, la responsabilité ou encore la réputation sont autant de préjudices potentiels d'actes numériques malveillants (cf. Sébastien Héon, 2013).

Les responsables de la sécurité des systèmes d'information (RSSI) sont garants de la nécessaire gestion des risques numériques au sein de leurs organisations selon l'Agence nationale de la sécurité des systèmes d'information (2018). Leurs actions face aux risques numériques se décomposent en quatre axes principaux : la gouvernance et l'anticipation, la protection, la défense et enfin la résilience. Pour rester aux faits des techniques d'attaques du moment, cette approche exigeante nécessite l'allocation de moyens adéquats de

la part des décideurs (cf. Daniel Zajdenweber, 2020). Afin de renforcer la résilience numérique de l'économie, les pouvoirs publics ont analysé ce sujet et participé à sensibiliser les acteurs économiques de son importance (cf. Meurant & Cardon, 2021). Les réglementations jouent en effet un rôle de catalyseur selon le Club des juristes (2018). A l'échelle de l'Union Européenne, la seconde directive Network and Information Security (NIS 2) élargit notamment le périmètre d'applicabilité de la première directive afin de généraliser les mesures de prévention au sein du tissu économique. De manière complémentaire, un cadre législatif pénalisant les actes criminels numériques semble essentiel pour enquêter et neutraliser les individus malveillants selon le Club des juristes (2021). En France, les peines et les amendes applicables en cas de criminalité numérique ont été renforcées en 2023 et peuvent aller jusqu'à dix ans d'emprisonnement et 500 000 euros d'amende en cas de préjudice corporel ou d'attaque en bande organisée (cf. JO, 2023). A l'échelle d'une entreprise, une gestion rationnelle des risques numériques ne suffit cependant pas à garantir la résilience financière de l'organisation, les conséquences pouvant être considérables. La mutualisation des dommages par l'intermédiaire d'organismes d'assurance peut ainsi contribuer à la résilience des acteurs économiques.

## L'assurance des risques numériques

Les produits couvrant les risques numériques concernent la plupart des garanties classiques, à savoir la responsabilité civile, les pertes pécuniaires et les dommages aux biens, tout en élargissant le périmètre des actifs intangibles assurés (cf. Wrede, Stegen, & Graf von der Schulenburg, 2020). Les méthodes de tarification semblent aujourd'hui principalement basés sur des avis d'experts (cf. Romanosky, Ablon, Kuehn, & Jones, 2019). Le marché de l'assurance des risques numériques est par ailleurs en forte croissance, avec un chiffre d'affaires d'environ 11 milliards de dollars en 2022 qui était moitié moindre en 2018 selon le réassureur MunichRe. Comme pour l'assurance habitation, la garantie responsabilité civile, résultant dans ce cas d'un acte numérique malveillant, constitue le socle des offres. Les garanties dommages et les pertes d'exploitation sont généralement proposées en option. Les statistiques sur ce marché sont cependant très parcellaires, en particulier au sein de l'Union Européenne. De l'autre côté de l'Atlantique, l'autorité compétente de régulation des entreprises d'assurance aux Etats-Unis publie annuellement (cf. NAIC, 2019) le chiffre d'affaires des principales compagnies et un indicateur de profitabilité, à savoir le ratio de sinistralité pour chacune des compagnies. Nous proposons une

analyse de ces rapports sous l'angle du chiffre d'affaires dans un premier temps et sous l'angle de la profitabilité dans un second temps. Les observations ci-dessous concernent donc uniquement les organismes d'assurance aux Etats-Unis.

- Le marché de l'assurance des risques numériques est en très forte croissance. Le chiffre d'affaires atteint près de 5 milliards de dollars en 2021 après une croissance exceptionnelle de 75% en un an. Depuis 2017, le marché était plutôt porté par une croissance d'environ 10%. Le marché semble de surcroît très concurrentiel, les parts de marché évoluant significativement. Par exemple, le plus faible nombre d'acteurs nécessaires pour regrouper plus de 50% des parts de marché était de quatre en 2018 et s'élève désormais à huit en 2021. Si Chubb était clairement leader en 2018 avec une part de marché de 36%, il conserve de justesse sa position en 2021 avec 10%. Les positions des concurrents varient sensiblement année après année.
- Le marché de l'assurance des risques numériques semble plus profitable que le marché global des assurances de biens et de responsabilités. Le ratio de sinistralité est défini par la somme des versements effectués par l'assureur aux parties prenantes des sinistres, des frais de justice et de gestion engagés par l'assureur pour gérer ces sinistres, rapportée au chiffre d'affaires. Il s'additionne au ratio d'administration pour obtenir le ratio combiné, ce dernier étant un indicateur direct de la profitabilité technique : sans prendre en compte la profitabilité financière, une valeur inférieure à 100% caractérise une activité profitable alors qu'une valeur supérieure à 100% indique une activité conduite à perte. Entre 2017 et 2021 sur les activités d'assurance de dommages et de responsabilité, le ratio d'administration fluctue entre 26% et 27% selon l'autorité. Ainsi, la profitabilité technique des produits d'assurance des risques numériques peut-elle être interprétée par la position relative des ratios de sinistralité par rapport au seuil de 73%. Pour l'assurance des risques numériques, le ratio de sinistralité évolue à l'image du chiffre d'affaires, passant de 35% en 2017 et 2018 à 44% en 2019 puis 66% jusqu'en 2021. Globalement, le marché semble donc profitable, avec des marges techniques de 38%, 29% et 7%, certes en décroissance mais très significativement supérieures à la moyenne des produits d'assurance dommages et de responsabilité affichant une marge technique de l'ordre de 3% entre 2017 et 2021.

Ainsi, le marché émergent de l'assurance des risques numériques aux Etats-Unis semble se rapprocher d'un équilibre du fait d'une concurrence dynamique. La nature évolutive des risques numériques peut cependant légitimement questionner la perspective d'un marché comparable aux assurances traditionnelles, l'offre se devant de rester cohérente avec la cri-

minalité numérique. En France, les données sont plus fragmentées. Pour autant, l'Institut national de la statistique et des études économiques (INSEE) a introduit à son enquête sur les technologies de l'information et de la communication dans les entreprises (TIC) des questions relatives à la gestion des risques numériques en 2019. Comme l'illustre le graphique 1, environ 15% des organisations sondées avaient subi un incident numérique en 2018 et 40% étaient assurées au moins en partie sur ces risques. Dans l'ensemble et sans faire de lien de causalité, on observe que les regroupements des sociétés les plus exposées en 2018 à des incidents numériques correspondent aux segments des sociétés les plus souscriptrices d'assurance.

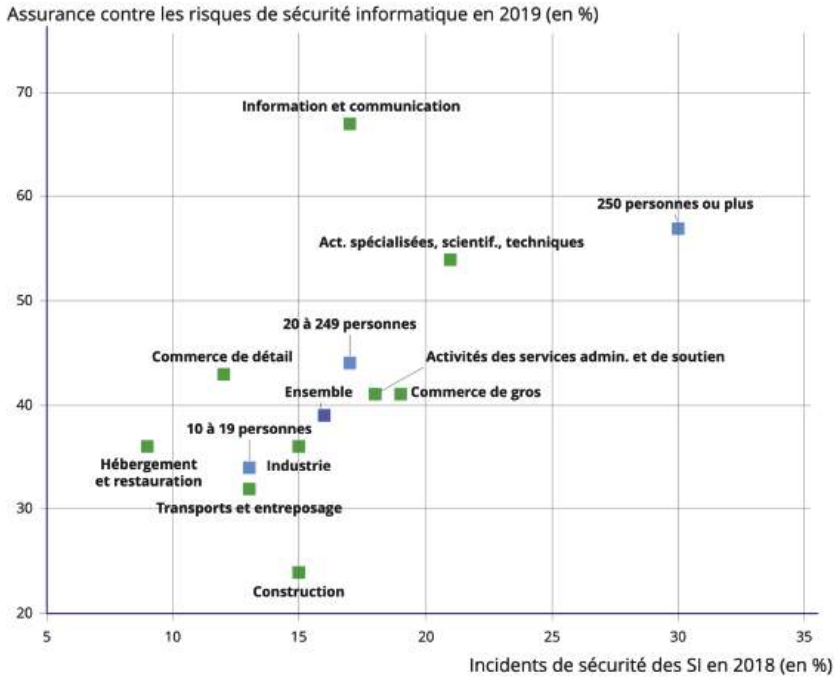



FIGURE 1 – Sécurité numérique des organisations : taux d'incidents et d'assurance d'après l'INSEE (2019).

Les pouvoirs publics ont récemment introduit deux nouvelles catégories dans la classification du code des Assurances : la catégorie ministérielle 32 relative aux dommages aux biens consécutifs aux atteintes aux systèmes d'information et de communication et la catégorie ministérielle 33 relative aux pertes pécuniaires consécutives aux atteintes aux systèmes d'information et de communication (JO, 2023). Le régulateur des assurances pourra ainsi analyser spécifiquement les comptes de ce type d'assurance et enrichir les données de marché. Par ailleurs, cette loi conditionne l'indemnisation d'un professionnel

victime d'une cyberattaque par son assureur au dépôt d'une plainte au plus tard 72 heures après la découverte de la cyberattaque. Les pouvoirs publics souhaitent ainsi renforcer les moyens de lutte contre la cybercriminalité en identifiant en paticulier les cas d'attaques solutionnées de manière négociées et discrètes.

Dans le même temps, les interrogations sur la soutenabilité des produits du marché de l'assurance des risques numériques demeurent (cf. Faure-Muntian, 2021 ; HCJPFP, 2022). En effet, les scénarios d'événements de grande ampleur sont nombreux, à l'image des exemples de cybercriminalité précédemment mentionnés. De plus, les coûts des sinistres semblent particulièrement volatiles. Selon l'International Association of Insurance Supervisors (2023), le coût moyen est estimé à 630 000$, significativement plus élevé que le coût médian de 140 000$. Toujours selon la même étude, les risques numériques sont davantage cédés au marché de la réassurance, à hauteur de 37%, relativement aux risques de dommages aux biens et de responsabilité, réassurés à 12%, illustrant une aversion aux déviations de ce risque vraisemblablement plus importante. Les enjeux mathématiques soulevés par l'analyse des coûts des risques numériques sont ainsi nombreux. Nous proposons de contribuer à l'approfondissement de certaines méthodes spécifiques.

## Les enjeux mathématiques

Dans le sillage du développement de l'assurance des risques numériques, la maîtrise des engagements d'assurance représente un enjeu croissant. Les contributions de recherche se sont ainsi développées. Un cadre de quantification des risques numériques illustrant la pertinence des contrats d'assurance est par exemple introduit (cf. Böhme & Kataria, 2006). La problématique globale de l'assurabilité des risques numériques est notamment posée (cf. Biener, Eling, & Wirfs, 2015). Les risques opérationnels liés au numérique et à la digitalisation sont également étudiés (cf. Eling, 2018 ; Eling & Schnell, 2016). Enfin, un cadre d'analyse de l'équilibre entre menaces criminelles et mesures de prévention est présenté (cf. Insua, Couce-Vieira, & Musaraj, 2018). Les problématiques mathématiques sont nombreuses pour les organismes d'assurance, par exemple la tarification des contrats, le provisionnement des sinistres non encore clôturés et la gestion des risques. Les données et les techniques statistiques utilisées à chacune de ces étapes sont cependant singulières. Ce manuscrit se concentre dans un premier temps sur la classification des risques à partir de valeurs extrêmes. Il développe dans un second temps l'estimation de copules à partir de données censurées et explore enfin l'estimation de processus de Hawkes et de scénarios

d'accumulation. Pour chaque problématique, un résumé des outils mathématiques associés et une synthèse de documents de travail ou de publications sont proposés.

## Classification des risques à partir de valeurs extrêmes

L'objectif de la classification est de comprendre l'hétérogénéité des profils de risques afin de proposer une différenciation pertinente. Généralement, le prix varie suivant les profils de risques pour prendre en compte les différences significatives de coûts associés à chacun des profils. Le principal modèle statistique utilisé à cet effet est le modèle linéaire généralisé, supposant une régularité des distributions de coûts et permettant de les estimer en fonction des variables explicatives considérées. Or, dans le cas des sinistres numériques, nous anticipons une grande volatilité des données de coûts du fait de l'hétérogénéité des situations, mais également de l'occurrence de sinistres très coûteux, à la suite par exemple d'une longue paralysie du fonctionnement d'une organisation.

Dans le cadre des risques numériques, l'hypothèse de régularité supposée dans le modèle linéaire généralisé pourrait ainsi ne pas être vérifiée. Pourtant, l'impact des profils de risque ou de sinistre sur le phénomène aléatoire des coûts n'en demeure pas moins intéressant, si ce n'est davantage dans l'objectif d'ajuster les a priori des expertises métiers par une analyse mathématique a posteriori. En ce sens, nous proposons une synthèse de la théorie des valeurs extrêmes au chapitre 1.1. Puis, nous présentons une adaptation aux valeurs extrêmes de l'algorithme des arbres de régression au chapitre 1.2. Enfin, au chapitre 1.3, nous synthétisons les travaux des publications appliquant cette méthodologie et démontrant des résultats de performance à échantillon fini. Des analyses complémentaires par simulation numérique sont également présentées.

## Estimation de copules à partir de données censurées

L'objectif du processus opérationnel de provisionnement est de valoriser les engagements liés à un portefeuille de sinistres survenus. Généralement, la plupart des sinistres survenus sont déclarés à l'assureur et la majorité de l'aléa réside ainsi dans l'évolution des coûts des dossiers de sinistres encore ouverts. La principale méthode actuarielle utilisée pour répondre à cet enjeu est la méthode dite de Chain Ladder ou de Mack (1993). Cette méthode repose sur des données de coût de sinistralité agrégées en triangle : chaque ligne concerne les sinistres survenus une certaine année et illustre de colonne en colonne les développements des coûts au fur et à mesure des clôtures de sinistres et avec un pas annuel.

L'aspect triangulaire est ainsi obtenu par construction : le recul de développement de la sinistralité est d'autant plus important que l'année d'assurance est ancienne. Cette méthode repose sur l'hypothèse implicite d'un développement en espérance multiplicatif de la sinistralité. Or, dans le cas des sinistres numériques, l'hétérogénéité des sinistres peut vraisemblablement conduire à des évolutions significatives de la typologie des sinistres en fonction des années, et ainsi affecter le développement temporel de la sinistralité. Par ailleurs, un sinistre numérique peut affecter une multiplicité de garanties, relatives aux dommages, aux pertes d'exploitation, aux responsabilités, ou encore à l'assistance. La méthode de Chain Ladder offre ainsi deux possibilités : regrouper les garanties, en supposant l'homogénéité des risques sous-jacents ou bien les considérer séparément avant de sommer les résultats, en supposant l'indépendance des garanties.

De manière alternative, nous proposons une analyse des coûts sinistre par sinistre. Nous disposons de données complètes pour les dossiers de sinistres clos et uniquement de données partielles pour les dossiers de sinistres ouverts. En particulier, les délai de traitement des sinistres ouverts sont inconnus et sont dits censurés à droite. Afin de prendre en compte ce phénomène de censure, nous introduisons des principes généraux théoriques au chapitre 2.1. De la même manière, pour gérer les données multidimensionnelles, nous présentons des résultats de la théorie des copules au chapitre 2.2. Enfin, nous synthétisons au chapitre 2.3 l'approche proposée dans un document de travail introduisant l'estimation d'un phénomène multidimentionnel au travers de ses marginales et de sa copule, dans un contexte de données censurées.

## Estimation de processus de Hawkes et de scénarios d'accumulation

Après avoir analysé les coûts des sinistres numériques, c'est-à-dire la sévérité, notre propos se concentre sur l'étude de la fréquence des sinistres numériques à savoir l'analyse de leur probabilité d'occurrence. Plus précisément, nous étudions les accumulations en fréquence des sinistres numériques, qui représentent un risque pour les organismes d'assurance en termes de réactivité d'assistance, de gestion de sinistres et donc d'aggravation des préjudices à indemniser. Généralement, une approche par scénario est développée pour répondre à cet enjeu en estimant pour chacun sa probabilité d'occurrence et son impact sur le portefeuille de clients. Par exemple, pour les garanties dommages, le scénario d'une inondation sur une zone dense en termes de biens assurés peut être considéré. Dans ce cas, la complémentarité entre des modèles paramétriques d'hydrologie et la cartographie des expositions permet une quantification de l'impact du scénario mais également

la confirmation de la pertinence du scénario retenu par rapport à d'autres alternatives. Or, dans le cas des sinistres numériques, la cartographie, si toutefois elle existait, serait d'une dimension importante, la localisation de chaque organisation contenant a minima l'ensemble des jalons numériques principaux. La modélisation des stratégies de diffusion d'une attaque numérique serait également complexe et l'ensemble de ces thématiques ne sont pas abordées dans ce manuscrit.

De manière alternative, nous proposons une méthodologie complémentaire permettant de quantifier les accumulations en fréquence, en supposant une dynamique de diffusion particulière caractérisée par la propriété suivante : chaque sinistre observé augmente temporairement la probabilité d'occurrence d'un nouveau sinistre, les effets de chacun des sinistres pouvant se cumuler. Notre étude repose plus précisément sur une analyse en temps continu des manifestations des sinistres par l'intermédiaire des processus ponctuels, introduits au chapitre 3.1. Puis, afin de prendre en compte ce mécanisme particulier de diffusion, nous introduisons les processus de Hawkes au chapitre 3.2. Enfin, nous synthétisons au chapitre 3.3 un document de travail analysant la propagation médiatique de rançongiciels et proposant une quantification des besoins simultanés d'assistance.

# Chapitre 1

# Classification des risques à partir de valeurs extrêmes

## Sommaire

## 1.1 Données extrêmes

Les coûts des sinistres numériques étant vraisemblablement volatiles, une modélisation permettant d'estimer les pertes extrêmes et de les expliquer permettrait d'apporter des éléments de réflexion à la souscription des risques. Les quantiles extrêmes de la distribution des coûts représentent ainsi un intérêt certain. La Théorie des Valeurs Extrêmes propose des méthodes d'estimation de quantiles extrêmes, y compris pour des quantiles supérieurs au maximum jamais observé. Nous introduisons un certain nombre de résultats sur lesquels nos contributions reposent et invitons le lecteur intéressé par la Théorie des Valeurs Extrêmes à lire des ouvrages de références (cf. Beirlant, Goegebeur, Segers, & Teugels, 2004 ; Coles, 2001 ; Haan & Ferreira, 2006).

### 1.1.1 Domaines d'attraction des valeurs extrêmes

L'étude des valeurs extrêmement élevées d'une variable aléatoire $Y$ correspond à l'analyse de sa queue de distribution. Par exemple, l'objectif peut être d'estimer la borne supérieure de son support, supposée finie. De manière plus générale, lorsque son support n'est pas supposé borné, l'objectif est plutôt d'estimer ses quantiles particulièrement élevés. Le maximum d'un échantillon, ou plus globalement les quantiles empiriques élevés d'un échantillon, apportent des éléments d'information mais présentent des limites structurelles. Dans le premier cas, la borne supérieure du support est forcément supérieure au maximum observé et dans le second cas, les quantiles empiriques d'ordres supérieurs à l'inverse du nombre d'observations seront tous égaux au maximum, alors même que le support n'est pas supposé borné. L'estimation non-paramétrique des quantiles se confronte ainsi au volume de données structurellement décroissant avec l'ordre considéré. Le théorème de Fisher–Tippett–Gnedenko, un des résultats fondateurs de la théorie des valeurs extrêmes, permet de supposer, sous certaines conditions, un comportement du maximum empirique approximativement paramétrique, simplifiant ainsi son estimation.

**Théorème 1.1.1 (Théorème de Fisher–Tippett–Gnedenko)**
*Soient* $(Y_i)_{1\leq i\leq n}$ *des variables aléatoires i.i.d. et* $F^n$ *la fonction de répartition de* $\max_{1\leq i\leq n}(Y_i)$. *S'il existe deux suites,* $(a_n)$ *de réels strictement positifs et* $(b_n)$ *de réels, telles que la limite* $\lim_{n\to\infty} F^n(a_n y + b_n)$ *existe et converge vers une distribution non dégénérée (cf. théorème 1.1.3 de Haan & Ferreira, 2006), alors cette distribution appartient à la famille paramétrique des distributions extrêmes indexée par* $\gamma$, *et donc chaque membre s'écrit de la*

*manière suivante :*

$$\forall\, \gamma \in \mathbb{R}^*,\, \forall x \text{ tel que } 1+\gamma\, x > 0, G_\gamma(x) = \exp\left(-(1+\gamma\, x)^{-1/\gamma}\right),$$

*avec pour* $\gamma = 0$ *le prolongement* $G_0(x) = \exp\left(-e^{-x}\right)$.

Le paramètre de forme $\gamma$ permet de distinguer trois domaines d'attraction. Le domaine de Weibull est définit pour $\gamma < 0$ et regroupe des distributions dont les supports admettent une borne supérieure finie. Le domaine de Gumbel est définit pour $\gamma = 0$ et regroupe des distributions présentant des queues dont les décroissances sont asymptotiquement équivalentes à des décroissances exponentielles. Enfin, le domaine de Fréchet est définit pour $\gamma > 0$ et regroupe les distributions présentant des queues dont les décroissances sont asymptotiquement équivalentes à des décroissances polynomiales. Globalement, le paramètre de forme, également appelé indice de queue, est un indicateur de la lourdeur de la queue de distribution. Par ailleurs, la famille paramétrique des distributions extrêmes peut être étendue à la famille des distributions extrêmes généralisées en introduisant un paramètre de localisation $\mu$ et un paramètre d'échelle $\sigma$ de la manière suivante : $G_{(\mu,\sigma,\gamma)}(x) = G_\gamma(\frac{x-\mu}{\sigma})$. Cette famille est notamment ajustée dans le cadre de la méthode des maxima par blocs, consistant à extraire des maximums locaux d'un échantillon, cf la section 5.3.1 de Coles (2001). Le théorème de Pickands–Balkema–De Haan est un résultat similaire s'appliquant cette fois aux excès de $Y$ par rapport à un certain seuil et montrant une convergence de la loi des excès vers la famille paramétrique des distributions Pareto généralisées (GPD).

**Distributions Pareto généralisées (GPD)**

La famille paramétrique des distributions Pareto généralisées (GPD), caractérisées par deux paramètres d'échelle et de forme $(\sigma, \gamma)$ et notées $\overline{H}_{(\sigma,\gamma)}$, se définit de la manière suivante :

$$\forall (\sigma,\gamma) \in \mathbb{R}^{+*} \times \mathbb{R}^*,\, \overline{H}_{(\sigma,\gamma)}(z) = \left(1+\gamma\frac{z}{\sigma}\right)^{-1/\gamma},$$

avec pour $\gamma = 0$ le prolongement $\overline{H}_{(\sigma,0)}(z) = e^{-\frac{z}{\sigma}}$.

**Théorème 1.1.2 (Théorème de Pickands–Balkema–De Haan)**

*Soit* $Y$ *une variable aléatoire de fonction de répartition* $F$ *et* $\overline{F}_u$ *la fonction de survie des excès par rapport au seuil* $u$*, définie par* $\overline{F}_u(z) = \mathbb{P}(Y - u > z \mid Y > u)$*. Si la condition du Théorème de Fisher–Tippett–Gnedenko est vérifiée pour des variables* $(Y_i)_{1\leq i\leq n}$ *i.i.d.*

*de même loi que $Y$ (cf. Balkema & de Haan, 1974 ; Pickands, 1975), alors pour un paramètre de forme $\gamma$ identique, la distribution $F_u$ converge vers la famille paramétrique des distributions Pareto généralisées (GPD) lorsque $u$ augmente.*

$$\exists\,(\sigma^*,\gamma^*)\in\mathbb{R}^{+*}\times\mathbb{R} \text{ tels que } \lim_{u\to\infty}\sup_{z>0}|\overline{F}_u(z)-\overline{H}_{(\sigma^*,\gamma^*)}(z)|=0.$$

La famille des distributions Pareto généralisées est plutôt ajustée dans le cadre de la méthode des Peaks Over Threshold, consistant à extraire les excès par rapport à un seuil $u$ au sein d'un échantillon, cf la section 5.3.2 de Coles (2001). De manière empirique, sous les conditions requises, l'erreur d'approximation de la distribution des excès par la famille paramétrique $\overline{H}_{\sigma,\gamma}$ sera d'autant plus faible que $u$ sera grand. Dans le même temps, l'erreur d'estimation de la famille paramétrique $\overline{H}_{\sigma,\gamma}$ sur les excès sera d'autant plus élevée que le nombre d'excès sera réduit. Ce compromis empirique entre biais d'approximation et variance d'estimation a fait l'objet de plusieurs études proposant des méthodes graphiques mais également des seuils optimaux en fonction de critères (cf. Coles, 2001 ; Haan & Ferreira, 2006). Par exemple, le seuil le plus bas permettant une approximation raisonnable peut être retenu. Nous proposons d'étudier plus en détail cette famille GPD, notamment sur le domaine de Fréchet.

### 1.1.2 Domaine de Fréchet et loi Pareto généralisée

Le domaine d'attraction de Fréchet nous intéresse particulièrement dans le cadre de la modélisation de risques sévères pouvant générer des sinistres très coûteux. Pour des valeurs de $\gamma$ strictement positives, il existe une condition équivalente à celle de la convergence non dégénérée que nous utiliserons par la suite.

**Théorème 1.1.3 (Théorème 4 de Gnedenko (1943))**
*Une distribution avec une fonction de survie $\overline{F}$ appartient au domaine d'attraction de Fréchet avec un paramètre de forme $\gamma>0$ si et seulement si $\forall y>0$,*

$$\lim_{t\to\infty}\frac{\overline{F}(ty)}{\overline{F}(t)}=y^{-1/\gamma}.$$

Autrement dit, ce résultat permet de caractériser le domaine de Fréchet par des queues de distributions à décroissances polynomiale. Des conditions équivalentes à celle de la convergence non dégénérée et valides sur l'ensemble des domaines d'attraction existent et sont par exemple précisées aux théorèmes 1.1.6 et 1.2.1 de Haan et Ferreira (2006). Par

ailleurs, au sein de cette classe à décroissances polynomiale, le paramètre de forme de la loi GPD est d'autant plus élevé que l'ordre maximal des moments finis de la distribution est faible, comme le précise le résultat ci-dessous.

**Proposition 1.1.4 (Section 5.3.1 de Haan et Ferreira (2006))**
*Si $Y$ une variable aléatoire appartient au domaine d'attraction de Fréchet, alors ses moments d'ordres strictement inférieurs à $1/\gamma$ sont finis et ceux d'ordres strictement supérieurs à $1/\gamma$ sont infinis,*

$$\forall\, 0 < \alpha < 1/\gamma,\, \mathbb{E}[|Y|^\alpha] < +\infty \text{ et } \forall\, \alpha > 1/\gamma,\, \mathbb{E}[|Y|^\alpha] = +\infty.$$

Par ailleurs, une distribution GPD admet une espérance finie si et seulement si son paramètre de forme $\gamma$ est strictement inférieur à 1 et une variance finie si et seulement si $\gamma$ est strictement inférieur à 0,5. Dans le cadre de notre étude, nous apporterons ainsi une attention particulière à la valeur du paramètre de forme, l'existence de moments finis étant une condition nécessaire à la souscription d'un risque dans le cadre de la maîtrise de la sinistralité espérée et de sa déviation.

L'estimation d'une distribution Pareto généralisée peut se réaliser sur le principe du maximum de vraisemblance dès lors qu'un seuil est fixé et permettant de définir les excès. Dans ce cadre, le processus d'optimisation ne correspond pas tout à fait à un maximum de vraisemblance, étant donné l'erreur d'approximation entre la distribution des excès et celle d'une GPD. La méthode s'apparente ainsi plutôt à un maximum de pseudo-vraisemblance. Supposons disposer de $n$ excès $(Z_i)_{1\leq i\leq n}$ par rapport à un seuil $u$. La pseudo-vraisemblance de l'observation de cet échantillon relativement à une loi GPD de paramètres $\boldsymbol{\theta} = (\sigma, \gamma)$, avec $\gamma \neq 0$, s'exprime de la manière suivante :

$$\forall\, \gamma \in \mathbb{R}^*,\, \mathcal{L}_{(Z_i)_{1\leq i\leq n}}(\sigma, \gamma) = -n\log(\sigma) - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{n} \log\left(1 + \frac{\gamma Z_i}{\sigma}\right).$$

La recherche de l'optimum revient ainsi à résoudre le problème d'annulation du gradient à deux inconnues. Si la solution analytique n'est pas calculable, Grimshaw (1993) remarque cependant que la paramétrisation alternative $\boldsymbol{\theta'} = (\frac{\gamma}{\sigma}, \gamma)$ permet de contenir le problème d'optimisation à une dimension. En effet, soit $(\hat{\sigma}, \hat{\gamma})$ un estimateur de pseudo-vraisemblance annulant les gradients, sa composante $\hat{\gamma}$ s'exprime en fonction de $\frac{\hat{\gamma}}{\hat{\sigma}}$ :

$$\begin{cases} \frac{\partial \mathcal{L}}{\sigma}(\hat{\sigma}) &= 0, \\ \frac{\partial \mathcal{L}}{\gamma}(\hat{\gamma}) &= 0, \end{cases}$$

$$\left\{\begin{array}{rcl} -\frac{n}{\hat{\sigma}} - \left(\frac{1}{\hat{\gamma}}+1\right)\sum_{i=1}^{n}\frac{-\frac{\hat{\gamma}Z_i}{\hat{\sigma}^2}}{1+\frac{\hat{\gamma}Z_i}{\hat{\sigma}}} & = & 0, \\ -\frac{1}{\hat{\gamma}^2}\sum_{i=1}^{n}\log\left(1+\frac{\hat{\gamma}Z_i}{\hat{\sigma}}\right) + \left(\frac{1}{\hat{\gamma}}+1\right)\sum_{i=1}^{n}\frac{\frac{Z_i}{\hat{\sigma}}}{1+\frac{\hat{\gamma}Z_i}{\hat{\sigma}}} & = & 0, \end{array}\right.$$

$$\left\{\begin{array}{rcl} \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\frac{\hat{\gamma}Y_i}{\hat{\sigma}}\right) & = & \hat{\gamma}, \\ \left[1+\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\frac{\hat{\gamma}Z_i}{\hat{\sigma}}\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n}(1-\frac{\hat{\gamma}Z_i}{\hat{\sigma}})^{-1}\right] & = & 1. \end{array}\right.$$

Sous cette formulation, on remarque en effet qu'une fois la connaissance de $\frac{\hat{\gamma}}{\hat{\sigma}}$ acquise, la valeur $\hat{\gamma}$ peut directement être déduite. Ce résultat est strictement relatif aux estimateurs du maximum de vraisemblance, la famille des lois de Pareto généralisées reste en particulier identifiable. Par ailleurs, l'estimateur par maximum de vraisemblance de $\gamma$ est invariant aux effets d'échelle sur les données, voir aussi la section 3.4 de Haan et Ferreira (2006).

### 1.1.3 Analyse conditionnelle et semi-paramétrique des excès

Les caractéristiques de la queue de distribution d'une variable aléatoire $Y$, représentant par exemple le coût unidimensionnel d'un sinistre, peuvent dépendre de variables explicatives $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$, portant par exemple sur le profil de risque assuré ou encore sur les circonstances du sinistre et de sa gestion. L'analyse conditionnelle des excès par rapport aux covariables permet de modéliser ces liens qui nous intéressent. Notre contribution s'appuie ainsi sur une version conditionnelle du résultat d'approximation des excès d'une distribution appartenant au domaine de Fréchet par une GPD. Notons $\overline{F}(\cdot \mid \mathbf{X})$ la fonction de survie de $Y|\mathbf{X}$, définie par $\overline{F}(y \mid \mathbf{X}=\mathbf{x}) = \mathbb{P}(Y > y \mid \mathbf{X}=\mathbf{x})$.

**Hypothèse 1.1.5** *Pour tout $\mathbf{x} \in \mathcal{X}$, la distribution conditionnelle de $Y|\mathbf{X}=\mathbf{x}$ appartient au domaine d'attraction de Fréchet, autrement dit il existe une fonction $\gamma^*$ strictement positive sur $\mathcal{X}$ telle que $\forall\, \mathbf{x} \in \mathcal{X}$, $\forall\, y > 0$,*

$$\lim_{t\to\infty} \frac{\overline{F}(ty \mid \mathbf{x})}{\overline{F}(t \mid \mathbf{x})} = y^{-1/\gamma^*(\mathbf{x})}. \tag{1.1.1}$$

**Corollaire 1.1.6** *Si l'hypothèse 1.1.5 est vérifiée, alors il existe des fonctions $\sigma^*(\mathbf{x})$ et $\gamma^*(\mathbf{x})$, strictement positives sur $\mathcal{X}$ telles que,*

$$\lim_{u\to\infty} \sup_{\mathbf{x}\in\mathcal{X}} \sup_{z>0} |\overline{F}_u(z|\mathbf{X}=\mathbf{x}) - \overline{H}_{(\sigma^*(\mathbf{x}),\gamma^*(\mathbf{x}))}(z)| = 0.$$

Notre contribution repose ainsi sur ce dernier corollaire et fixe donc l'estimation du couple de fonctions $\boldsymbol{\theta}^*(\mathbf{x}) = (\sigma^*(\mathbf{x}), \gamma^*(\mathbf{x}))$ comme l'objectif de l'analyse conditionnelle de la queue de distribution $Y$ sachant $\mathbf{X}$. Cette adaptation de la théorie de valeurs extrêmes au cadre conditionnel, permettant d'identifier l'hétérogénéité des queues de distribution des risques, a déjà été proposée en supposant des structures de dépendance différentes :

- une régularité des fonctions $\sigma^*(\mathbf{x})$ et $\gamma^*(\mathbf{x})$ estimées par la méthode d'estimation non paramétrique des noyaux applicable pour des covariables absolument continues (cf. Beirlant & Goegebeur, 2003),
- une régularité des fonctions $\sigma^*(\mathbf{x})$ et $\gamma^*(\mathbf{x})$ estimées par une méthode d'estimation par polynômes locaux, applicable pour des covariables absolument continues (cf. Beirlant & Goegebeur, 2004),
- une régularité des fonctions $\sigma^*(\mathbf{x})$ et $\gamma^*(\mathbf{x})$ estimées par une technique d'estimation reposant sur des polynômes locaux pour les covariables absolument continues, applicable pour tout type de covariables (cf. Chavez-Demoulin, Embrechts, & Hofert, 2015).

Nous développons de manière alternative une approche reposant sur les arbres de régression et adaptée au cadre des valeurs extrêmes.

## 1.2 Arbres de régression

Dans le cadre de la souscription des risques numériques, la compréhension du lien entre des caractéristiques du risque et la sinistralité contribuent aux stratégies de souscription.

### 1.2.1 Données hétérogènes et M-estimation

En statistique, le principe de la régression consiste à estimer un phénomène aléatoire d'intérêt, par exemple $L$, en fonction de comportements aléatoires notés $\mathbf{X}$ et potentiellement informatifs. Notre propos se focalise plus précisément sur les M-estimateurs.

**M-estimateurs** [Huber (1964)]
Les M-estimateurs reposent sur l'hypothèse d'existence d'une fonction de régression $m^*$ s'exprimant comme l'argument minimum d'un critère déterministe au sein d'un espace de fonctions $\mathcal{M}$,

$$m^* = \arg\min_{m \in \mathcal{M}} E[\phi(L, m(\mathbf{X}))],$$

où $\phi$ est une fonction de perte entre la variable d'intérêt $L$ et sa modélisation $m(\mathbf{X})$.

Pour un espace de fonctions $\mathcal{M}$ et une fonction de perte $\phi$, l'objectif d'une analyse de régression est alors de proposer un estimateur $\hat{m}$ de la fonction de régression $m^*$. Le choix préalable de $\mathcal{M}$ et de $\phi$ doit s'adapter au contexte aléatoire $(L, \mathbf{X})$, et aux objectifs d'apprentissage sur $L$. En voici quelques exemples :

- sous le contexte d'une espérance finie de la variable d'intérêt $\mathbb{E}[L] < +\infty$, et pour l'espace de fonction $\mathcal{M} = \{m : \mathbb{R}^d \to \mathbb{R} \ / \ \mathbb{E}[m(\mathbf{X})^2] < +\infty\}$, le choix de la perte quadratique $\phi(l, m(\mathbf{x})) = (l - m(\mathbf{x}))^2$ implique que la fonction de régression corresponde à l'espérance conditionnelle $m^*(\mathbf{x}) = \mathbb{E}[L|\mathbf{X} = \mathbf{x}]$,
- dans un contexte général et pour l'espace de fonction $\mathcal{M} = \{m : \mathbb{R}^d \to \mathbb{R}\}$, le choix de la perte quadratique $\phi(l, m(\mathbf{x})) = |y - m(\mathbf{x})|$ implique que la fonction de régression corresponde à la médiane conditionnelle $m^*(\mathbf{x}) = q_{0.5}[Y|\mathbf{X} = \mathbf{x}]$,
- sous le contexte d'un modèle linéaire homoscédastique et pour l'espace paramétrique des fonctions linéaires $\mathcal{M}_\beta = \{m_\beta : \mathbb{R}^d \to \mathbb{R} \ / \ m_\beta(\mathbf{X}) = \mathbf{X}\beta\}$, le choix de la perte quadratique $\phi(y, m(\mathbf{x})) = (l - m(\mathbf{x}))^2$ implique que la fonction de régression corresponde au modèle : $m^* = m_{\beta^*}$.

Dans le contexte d'analyse conditionnelle des queues de distribution, modélisées au sein de la famille paramétrique des lois Pareto généralisées par l'objectif $\boldsymbol{\theta}^*(\mathbf{x})$, le choix de la

perte opposée à la pseudo log-vraisemblance $\mathcal{L}(\sigma, \gamma)$, définie à l'équation 1.1.2, implique que la fonction de régression approche le modèle $m^* \approx \boldsymbol{\theta}^*(\mathbf{x})$. Nous allons nous intéresser au M-estimateur de $\boldsymbol{\theta}^*(\mathbf{x})$ construit à partir d'arbres de régression.

### 1.2.2 Algorithme CART

Une solution de régression est l'algorithme d'arbre de régression introduit sous l'acronyme CART pour *"Classification and Regression Trees"* (cf. Breiman, Friedman, Stone, & Olshen, 1984). Cet algorithme, codé sous licence publique générale (cf. Terry M. Therneau & Mayo, 2022), est une nouvelle méthode de construction d'un estimateur de fonction de régression $\hat{m}$. Pour une certaine perte $\phi$, un arbre de régression est un algorithme de construction d'un estimateur $\hat{m}$. Cet estimateur est considéré au sein de l'espace général des fonctions constantes par hypercubes de $\mathcal{X} \subset \mathbb{R}^d$. Ces hypercubes partitionnant l'espace $\mathcal{X}$ peuvent s'interpréter comme des règles de classification de profils, sur chacune desquelles une prédiction constante est associée. Pour un ensemble d'indicatrices d'hypercubes ou de règles indéxé par $\mathcal{S}$, que nous notons $\{R_l\}_{l \in \mathcal{S}}$, la fonction de régression $\hat{m}$ associée est de la forme :

$$\hat{m}^{\mathcal{S}}(\mathbf{x}) = \sum_{l \in \mathcal{S}} \hat{m}(R_l) R_l(\mathbf{x}),$$

avec pour tout $l \in \mathcal{S}, R_l \in \{R : \mathcal{X} \to \{0;1\} \ / \ R(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_W \leq \mathbf{x} < \mathbf{x}_M}\}$ pour $\mathbf{x}_W$ et $\mathbf{x}_M$ différents éléments de $\mathcal{X}$, $R_l(\mathbf{x}) R_{l'}(\mathbf{x}) = 0$ pour $l \neq l'$, $\sum_{l \in \mathcal{S}} R_l(\mathbf{x}) = 1$, et $\hat{m}(R_l) = \arg\min_{m \in \mathcal{M}} \sum_{i=1}^{n} \phi(Y_i, m(\mathbf{x}_i)) R_l(\mathbf{x}_i)$. Nous allons précisément utiliser la perte $\phi$ égale à l'opposée de la pseudo log-vraisemblance liée à une loi de Pareto généralisée, proposant une variante à la fonction de perte quadratique. Les variantes des arbres de régression sont multiples comme le montrent les revues de littérature Loh (2011) et Loh (2014). Par exemple, une perte liée à la vraisemblance des échantillons par rapport à une famille paramétrique a été considérée (cf. Su, Wang, & Fan, 2004), et l'algorithme est adaptable (cf. Therneau & Mayo, 2022).

Le processus d'obtention des hypercubes ou des règles partitionnant l'espace $\mathcal{X}$ caractéristique des arbres de régression se décompose en deux étapes : une phase de *"croissance"* de l'arbre aboutissant à une partition de $\mathcal{X}$ par $K_{max}$ hypercubes, et une étape d'*"élagage"* consistant à extraire un sous-arbre de $\hat{K}$ hypercubes.

#### 1.2.2.1 Phase de *"croissance"*

Le partitionnement de l'espace $\mathcal{X}$ en hypercubes est effectué au moyen d'une division itérative. Chaque étape $k$ consiste à faire évoluer un ensemble de règles $\mathcal{R}^k$ indexé par $\mathcal{S}^k$ vers un nouvel ensemble de règles $\mathcal{R}^{k+1}$ indexé par $\mathcal{S}^{k+1}$ de la manière suivante :
**Initialisation :** l'espace $\mathcal{X}$ n'est pas encore partitionné et nous disposons d'une seule règle, $\mathcal{S}^1 = \{1\}$, et $R_1(\mathbf{x}) = 1$ pour tout $\mathbf{x} \in \mathcal{X}$. La prédiction non conditionnelle est alors $\hat{m}(R_1) = \arg\min_{m\in\mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i))$. Nous notons cette estimation d'un arbre à 1 feuille $\hat{\boldsymbol{\theta}}^1(\mathbf{x})$.
**Étape** $K+1$ **:** à partir de l'ensemble des règles obtenues à l'étape $K$, noté $\mathcal{R}^k = \{R_l\}_{l\in\mathcal{S}^k}$ et correspondant à l'estimateur à $K$ feuilles $\hat{\boldsymbol{\theta}}^K(\mathbf{x})$, nous recherchons une division optimale sur tous les hypercubes. Pour tout $l \in \mathcal{S}^k$ nous réalisons les opérations suivantes :

- si toutes les observations telles que $R_l(\mathbf{X}_i) = 1$ ont le même profil $\mathbf{x}$, nous ne créons pas de nouvelles règles.
- sinon, nous créons deux nouvelles règles $R_{l1}$ et $R_{l2}$ de la manière suivante :
  - pour toute composante $X^{(j)}$ de $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, nous calculons la meilleure position de division $\hat{x}^{(j)} = \arg\max_{x^{(j)}} \Phi(R_l, x^{(j)})$, avec :

$$\begin{aligned}
\Phi(R_l, x^{(j)}) &= \sum_{i=1}^n \phi(Y_i, \widehat{m}(R_l)) R_l(\mathbf{X_i}) \\
&- \sum_{i=1}^n \phi(Y_i, \hat{m}_{j-}(\mathbf{X}_i, R_l)) \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_l(\mathbf{X_i}) \\
&- \sum_{i=1}^n \phi(Y_i, \hat{m}_{j+}(\mathbf{X}_i, R_l)) \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_l(\mathbf{X_i}),
\end{aligned}$$

et où

$$\begin{aligned}
\widehat{m}(R_l) &= \arg\min_{m\in\mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i)) R_l(\mathbf{X}_i), \\
\hat{m}_{j-}(R_l) &= \arg\min_{m\in\mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i)) \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_l(\mathbf{X}_i), \\
\hat{m}_{j+}(R_l) &= \arg\min_{m\in\mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i)) \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_l(\mathbf{X}_i).
\end{aligned}$$

- la division optimale $\hat{x}^{(\hat{j})}$ est sélectionnée parmi les divisions optimales de chacune des composantes $\widehat{j} = \arg\min_j \Phi(R_l, \hat{x}^{(j)})$ ,

- les nouvelles règles correspondantes sont introduites $R_{l1}(\mathbf{x}) = R_l(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(\hat{j})} \leq \hat{x}^{(\hat{j})}}$, et $R_{l2}(\mathbf{x}) = R_l(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(\hat{j})} > \hat{x}^{(\hat{j})}}$, et $S^{k+1} = S^k \cup \{l1, l2\}$. Nous notons alors à cette étape l'estimateur obtenu à $K+1$ feuilles $\hat{\boldsymbol{\theta}}^{K+1}(\mathbf{x})$

**Arrêt de l'itération** si aucune nouvelle règle n'a été créée, c'est-à-dire si $|S^{k+1}| = |S^k|$, où $|.|$ est l'opérateur cardinal. Notons alors $\mathcal{R}^{K_{max}}$ la dernière partition ainsi obtenue.

Dans cette version de l'algorithme CART, toutes les covariables sont continues ou binaires $\{0, 1\}$. Ainsi, pour chaque variable qualitative ayant $p$ modalités avec $p > 2$, deux solutions sont envisageables : soit la transformer en $p-1$ variables binaires, soit compléter l'algorithme au niveau des positions de division de $R_l$ en introduisant un ordre parmi les modalités. Par exemple en ordonnant les modalités par rapport aux valeurs moyennes - ou médianes - des observations $Y_i$ associées à chaque modalité. Enfin, l'algorithme peut être personnalisé par des hyperparamètres limitant les divisions considérées ou encore la règle d'arrêt. A titre d'illustration, un nombre minimal d'observations dans chaque feuille peut être fixé.

#### 1.2.2.2 Phase d'*"élagage"*

L'étape d'élagage consiste à sélectionner parmi l'ensemble des partitions imbriquées $\left(\mathcal{R}^K\right)_{1 \leq K \leq K_{max}}$ obtenues lors de l'algorithme celle qui minimise un critère $C_\alpha$ pénalisant la performance de chaque sous-arbre par son nombre de feuilles :

$$\hat{K}_\alpha = \arg \min_{1 \leq K \leq K_{max}} C_\alpha(\mathcal{S}^K) = \sum_{i=1}^{n} \phi(Y_i, m^{\mathcal{S}^K}(\mathbf{X}_i)) + \alpha |\mathcal{S}^K|,$$

où la constante de pénalisation $\alpha$ est choisie à l'aide d'un échantillon de test ou d'une validation croisée en $p-$parties. Dans le premier cas, les données sont divisées en deux parties avant la croissance de l'arbre (un ensemble de données d'apprentissage et un échantillon de test qui n'est pas utilisé pour le calcul de l'arbre). Dans le second cas, l'ensemble de données est partitionné aléatoirement en $p$ parties qui servent successivement d'échantillon d'apprentissage ou de test. Une fois $\widehat{\alpha}$ calibré, nous obtenons la partition définitive $\mathcal{S}^{\hat{K}_{\hat{\alpha}}}$, et l'estimation du modèle de régression associé $m^{\hat{K}_{\hat{\alpha}}}$, noté $\hat{m}^{\hat{K}}$ ou $\hat{m}$. Nous notons alors l'estimateur associé à $\hat{K}$ feuilles $\hat{\boldsymbol{\theta}}^{\hat{K}}(\mathbf{x})$.

### 1.2.3 Données manquantes et importance des variables

La retranscription de l'expérience en une base de données peut générer des données manquantes ou incomplètes. Le questionnaire de souscription d'un organisme d'assurance peut par exemple évoluer avec le temps et modifier les informations enregistrées dans son système d'information à chaque souscription. Dans cette situation, la nature ou la précision des informations de chaque contrat varie sur l'ensemble de la base de données disponible et on peut parler de données manquantes. Grâce à une culture de la donnée, les porteurs de risques disposent généralement d'un socle de données de qualité issu des questionnaires de souscription et des formulaires de déclarations de sinistres. Conformément à la réglementation du secteur, ces données doivent permettre une compréhension suffisante des risques assurés. La gestion des données manquantes peut toutefois présenter des intérêts.

L'algorithme CART permet de gérer la présence de données manquantes dans les variables explicatives $\mathbf{X}$ grâce au concept de règles de substitution (cf. Breiman et al., 1984 ; Nicholas J Tierney & Mengersen, 2015). Ces règles de substitution sont par ailleurs utilisées pour évaluer l'importance des variables, dont l'objectif est de manière informelle de quantifier les contributions de chaque variable explicative à l'apprentissage statistique.

Pour obtenir les règles de substitution, le processus de croissance d'un arbre de régression est adapté de la manière suivante :

- à partir de $R_l$, les meilleures positions de division $\hat{x}^{(j)}$ pour chacune des composantes $j = 1, \ldots, d$, sont calculées uniquement sur la sous population des observations pour laquelle les valeurs de la composante $X^{(j)}$ sont complètes. Autrement dit la fonction $\Phi$ est adaptée par $\tilde{\Phi}^{(j)}$ parcourant uniquement cette sous population. La division optimale $\hat{x}^{(\hat{j})}$ est ensuite sélectionnée parmi les divisions optimales de chacune des composantes $\widehat{j} = \arg\min_j \tilde{\Phi}^{(j)}(R_l, \hat{x}^{(j)})$, définissant les nouvelles règles optimales $(R_{l1}, R_{l2})$.
- les règles de substitution sont ensuite introduites et sélectionnées à partir de règles alternatives $(\tilde{R}^s_{l1}, \tilde{R}^s_{l2})$, chacune résultante d'une division sur la composante $s \in \{1, \ldots, d\} \setminus \widehat{j}$. Ces règles visent à imiter autant que faire ce peu les règles précédemment établies $(R_{l1}, R_{l2})$. Autrement dit, les divisions optimales ne maximisent plus $\tilde{\Phi}^{(j)}$ mais minimisent plutôt la fonction d'erreur de classification $\Delta$ définie comme suit :

$$\Delta_l^s : \left((R_{l1}, R_{l2}), (\tilde{R}_{l1}^s, \tilde{R}_{l2}^s)\right) \longmapsto \frac{\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} \left(R_{l1}(\mathbf{X}_i)\tilde{R}_{l2}^s(\mathbf{X}_i) + R_{l2}(\mathbf{X}_i)\tilde{R}_{l1}^s(\mathbf{X}_i)\right)}{\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} \left(R_{l1}(\mathbf{X}_i) + R_{l2}(\mathbf{X}_i)\right)}.$$

Ensuite, parmi ces couples de règles candidates, ne sont sélectionnés que ceux dont la performance excède celle du couple des règles majoritaires dirigeant toutes les observations manquantes vers la règle la plus abondante entre $R_{l1}$ et $R_{l2}$. Plus précisément, les couples sélectionnés vérifient $\Delta_l^s\left((R_{l1}, R_{l2}), (\tilde{R}_{l1}^s, \tilde{R}_{l2}^s)\right) < \Delta_l^s\left((R_{l1}, R_{l2}), (\tilde{R}_{l1}^{maj}, \tilde{R}_{l2}^{maj})\right)$, avec :

$$\begin{cases} (\tilde{R}_{l1}^{maj}, \tilde{R}_{l2}^{maj}) = (R_l(\mathbf{x}), 0) \text{ si } \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} R_{l1}(\mathbf{X}_i) \geq \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} R_{l2}(\mathbf{X}_i), \\ \text{ou} \\ (\tilde{R}_{l1}^{maj}, \tilde{R}_{l2}^{maj}) = (0, R_l(\mathbf{x})) \text{ si } \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} R_{l1}(\mathbf{X}_i) < \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{\widehat{(j)}} \neq \mathrm{NA}} R_{l2}(\mathbf{X}_i), \end{cases}$$

Les observations partielles sont alors allouées aux différentes feuilles $R_{l1}$ et $R_{l2}$ grâce aux règles alternatives. Soit $\Lambda_l$ l'ensemble des $s \in \{1, \ldots, d\} \setminus \widehat{j}$ à l'origine d'une règle de substitution et notons $\Delta_l^s$ les erreurs associées pour $s \in \Lambda_l$, les règles de substitutions sont utilisées par ordre décroissant de $\Delta_l^s$ pour allouer les observations partielles. S'il en reste et à défaut, les règles majoritaires $(\tilde{R}_{l1}^{maj}, \tilde{R}_{l2}^{maj})$ sont utilisées.

Enfin, l'importance des variables est une mesure empirique de la contribution de chaque composante à la réduction de la perte permise par l'arbre de régression. Cette mesure empirique fait l'objet de plusieurs définitions. Elle peut par exemple s'appuyer sur l'amélioration de la performance de l'arbre à chaque étape de sa croissance et jusqu'à la profondeur sélectionnée $\hat{k}$. À chaque étape $k+1$, cette amélioration correspond à $\Phi(R_l, \hat{x}^{\widehat{(j)}})$ pour la division principale. La mesure de l'importance des variables peut également se baser sur les règles de substitution. En particulier, l'importance de chaque composante $s$ impliquée dans des règles de substitution à l'étape $k+1$ est définie comme une part de l'amélioration de la division principale $\Phi(R_l, \hat{x}^{\widehat{(j)}}$ définie par le poids $w_l^s = (\Delta_l^s - \Delta_l^{\mathrm{maj}})/(1 - \Delta_l^{\mathrm{maj}})$. Alors, l'importance des variables est obtenue en normalisant le vecteur $I = (i_j)_{j \in 1, \ldots, d}$ défini par :

$$i_j = \sum_{l=1}^{\hat{K}} \left(\Phi(R_l, \hat{x}^{\widehat{(j)}}) \mathbf{1}_{j=\widehat{j}} + \Phi(R_l, \hat{x}^{\widehat{(j)}}) w_l^s \mathbf{1}_{j \in \Lambda_j}\right).$$

**Remarque 1.2.1** *Pour calculer l'importance des variables, la recherche de toutes les règles de substitution doit être effectuée indépendamment de la présence de valeurs manquantes. Cet indicateur n'est cependant pas exempt de limites, ne reposant pas sur un*

*objectif mathématique précis et pouvant induire des biais dans le cas de la présence de valeurs manquantes (cf. Kim & Loh, 2001).*

- *Le premier biais concerne la phase de "croissance" de l'arbre de régression et se présente lorsqu'au sein d'une feuille, les proportions de valeurs manquantes de chaque variables sont différentes. Comme la performances des règles candidates sont uniquement évaluées sur les données complètes de chaque variable, alors le nombre de points sur lesquels la performance est calculée est d'autant plus faible que la proportion de données manquantes pour chaque variable est importante au sein de la feuille.*
- *Le second biais concerne les règles de substitution et donc a fortiori l'importance des variables. Lorsque pour les variables alternatives, les proportions de valeurs manquantes en commun avec la variable utilisée par la règle sélectionnée sont différentes, alors les variables alternatives avec la plus forte proportion de valeurs manquantes sont davantage pénalisées. A l'inverse de la performance, l'erreur de classification est calculée sur un même périmètre pour chaque variable : sur les données complètes de la variable utilisée par la règle sélectionnée uniquement. Or, plus une variable a des valeurs manquantes sur ce périmètre, moins ses chances seront élevées puisque limitées au périmètre des valeurs complètes en commun avec la variable utilisée pour la règle sélectionnée.*

## 1.3 Analyse conditionnelle de la sinistralité numérique extrême

### 1.3.1 Résultats mathématiques

Nous supposons observer des réalisations i.i.d. $(Y_i, \mathbf{X}_i)_{1\leq i\leq n}$ de même loi que $(Y, \mathbf{X})$, $Y$ représentant les coûts des sinistres et $\mathbf{X}$ les covariables. Notre objectif est d'estimer l'hétérogénéité des queues de distribution des coûts des sinistres $Y$ en fonction des covariables $\mathbf{X}$, autrement dit de réaliser l'estimation semi-paramétrique de $\boldsymbol{\theta}^*(\mathbf{x})$. A cet effet, la méthode que nous proposons consiste en une classification des queues de distribution par un arbre de régression maximisant la vraisemblance d'une loi Pareto généralisée à chaque raffinement et dont la profondeur est pénalisée. Nous supposons l'estimation préalable d'un seuil $\hat{u}$ au-delà duquel l'approximation de la partie droite de la distribution par une loi de Pareto généralisée et la variance d'estimation sont considérées comme satisfaisantes. Nous supposons que ce seuil $\hat{u}$ appartienne à un compact $[u_{min}; u_{max}]$, défini à partir de la distribution de $Y$ et d'une suite $k_n$, tendant vers l'infini et asymptotiquement négligeable devant $n$ :

$$\begin{cases} \mathbb{P}(Y \geq u_{\min}) &= \frac{k_n}{n}, \\ \mathbb{P}(Y \geq u_{\max}) &= \frac{\tau k_n}{n} \text{ pour } 0 < \tau < 1. \end{cases}$$

L'étude porte alors sur les observations au-delà de ce seuil $\hat{u}$, dont le nombre est asymptotiquement dominé et approximativement égal à $k_n$. Les éléments de contexte exhaustif sont précisés en partie B. L'objectif global de la méthode est d'estimer les paramètres conditionnels de la GPD théorique $\boldsymbol{\theta}_0(\mathbf{x})$ par $\widehat{\boldsymbol{\theta}}(\mathbf{x})$. Nous notons $\boldsymbol{\theta}^*(\mathbf{x})$ l'estimateur optimal sur pseudo données Pareto généralisées. L'objectif de cette partie est de réaliser une synthèse de la performance stochastique de cette procédure, c'est-à-dire de la distance entre $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ et $\boldsymbol{\theta}^*(\mathbf{x})$ en trois axes. Nous présenterons dans un premier temps nos objectifs d'évaluation de la performance, en cohérence avec les contraintes d'application. Nous expliciterons ensuite les principaux résultats obtenus. Nous terminerons cette synthèse par un résumé des étapes clés des démonstrations. Nous n'aborderons pas l'erreur d'approximation entre $\boldsymbol{\theta}_0(\mathbf{x})$ et $\boldsymbol{\theta}^*(\mathbf{x})$.

#### 1.3.1.1 Contraintes d'application et définition des objectifs

Notre méthode repose sur approximativement $k_n$ données extrêmes dont l'identification dépend du seuil $\hat{u}$ retenu. Empiriquement, $k_n$ est ainsi limité par la taille $n$ de

l'échantillon initial et par la proportion d'excès $\frac{k_n}{n}$, contrainte par un compromis entre biais et variance. Nous souhaitons donc contrôler les déviations du gradient de la pseudo log-vraisemblance à distance finie, afin d'en déduire des propriétés sur la qualité de l'estimation sur un échantillon fini. Nous souhaitons de plus que ce contrôle soit uniforme sur le compact des seuils considérés $[u_{min}; u_{max}]$ pour prendre en compte l'estimation de $\hat{u}$ à partir des données, $\hat{u}$ n'étant pas fixé a priori. Nous pourrons ainsi quantifier la performance de l'estimation et de la pénalisation à échantillon fini, permettant de discuter l'impact du nombre de classes de risques extrêmes. Cependant, l'application des inégalités de concentration fournissant des bornes exponentielles n'est pas immédiate, ces dernières reposant sur des hypothèses de support borné. En effet, les moments exponentiels de la vraisemblance d'une loi Pareto généralisée et de ses dérivées, utilisés dans chacune des feuilles, ne sont généralement pas bornés. Or les variables Pareto généralisées ne sont pas bornées pour $\gamma$ strictement positif et leur queue de distribution potentiellement lourde, ne permettant d'obtenir des moments finis uniquement jusqu'à l'ordre $1/\gamma$ exclu, ne permet pas de considérer des versions tronquées de ces variables comme des approximations satisfaisantes. L'objectif du résultat souhaité est ainsi d'ajuster des inégalités de concentration performantes sur l'erreur stochastique, qui plus est uniformément par rapport au seuil $u$.

#### 1.3.1.2 Performances stochastiques

Nous définissons l'erreur stochastique de l'estimation semi-paramétrique de la loi Pareto généralisée à partir d'une métrique entre deux fonctionnelles $\boldsymbol{\theta}(\mathbf{x})$ et $\tilde{\boldsymbol{\theta}}(\mathbf{x})$ :

$$\|\boldsymbol{\theta}(\mathbf{x}) - \tilde{\boldsymbol{\theta}}(\mathbf{x})\|_2 = \left( \int \|\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}'(\mathbf{x})\|_\infty^2 \mathrm{d}P_{\mathbf{X}}(\mathbf{x}) \right)^{1/2},$$

avec $\|\boldsymbol{\theta}(\mathbf{x}) - \tilde{\boldsymbol{\theta}}(\mathbf{x})\|_\infty = \max\left(|\sigma(\mathbf{x}) - \tilde{\sigma}(\mathbf{x})|, |\gamma(\mathbf{x}) - \tilde{\gamma}(\mathbf{x})|\right)$ et $P_{\mathbf{X}}$ la distribution des covariables $\mathbf{X}$. Pour un certain seuil $u$, l'erreur stochastique est étudiée sur deux plans : d'abord par le théorème 1.3.1 pour une partition à $K$ feuilles donnée entre l'estimateur ${}^u\hat{\boldsymbol{\theta}}^K(\mathbf{x})$ et son objectif ${}^u\boldsymbol{\theta}^{*K}(\mathbf{x})$ puis par le théorème 1.3.2 pour un ensemble de sous-arbres entre l'estimateur sélectionné suite à la pénalisation sur $K$, noté ${}^u\hat{\boldsymbol{\theta}}^{\hat{K}}(\mathbf{x})$ et son objectif ${}^u\hat{\boldsymbol{\theta}}^{K^*}(\mathbf{x})$.

**Théorème 1.3.1** *Sous les conditions précisées au théorème 1 de l'article B, nous obtenons l'existence de constantes positives $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_3$ telles que $\forall\ t > K(\log k_n)k_n^{-1}$,*

$$\begin{aligned}\mathbb{P}&\left(\sup_{u_{\min}\leq u\leq u_{\max}} \|{}^u\hat{\boldsymbol{\theta}}^K(\mathbf{x}) - {}^u\boldsymbol{\theta}^{*K}(\mathbf{x})\|_2^2 \geq t\right)\\ &\leq\ 2\left(\exp\left(-\frac{\mathcal{C}_1 k_n t}{K(\log k_n)^2}\right) + \exp\left(-\frac{\mathcal{C}_2 k_n t^{1/2}}{K^{1/2}\log k_n}\right)\right) + \frac{\mathcal{C}_3 K}{k_n^{5/2} t^{3/2}}, \qquad (1.3.1)\end{aligned}$$

*Nous pouvons notamment majorer l'espérance de l'erreur stochastique de la manière suivante pour une constante positive $\mathcal{C}_4$.*

$$\mathbb{E}\left[\sup_{u_{\min}\leq u\leq u_{\max}} \|{}^u\hat{\boldsymbol{\theta}}^K(\mathbf{x}) - {}^u\boldsymbol{\theta}^{*K}(\mathbf{x})\|_2^2\right]^{1/2} \leq {\mathcal{C}_4}^{1/2}\frac{K^{1/2}\log k_n}{k_n^{1/2}}. \qquad (1.3.2)$$

Nous remarquons dans la première partie du théorème l'obtention d'une inégalité de concentration uniforme en $u$ dont la borne est en partie exponentielle et en partie polynomiale, chacune étant croissante en fonction du nombre de feuilles $K$ et décroissante en fonction du nombre approximatif de données $k_n$. Ce premier résultat peut avoir plusieurs applications, par exemple l'estimation d'intervalles de confiance. En particulier, la seconde partie du théorème met en avant une borne proportionnelle à $K^{1/2}$ et asymptotiquement équivalente à celle d'un estimateur paramétrique ${k_n}^{-1/2}$.

**Théorème 1.3.2** *Parmi les conditions précisées au théorème 3 de l'article B, nous supposons en particulier que la constante de pénalisation du nombre de feuilles $\lambda$, vérifie l'inégalité suivante :*

$$0 < c_2\{\log k_n\}^{1/2}k_n^{-1/2} \leq \lambda \leq \mathfrak{D} - 2c_2\{\log(k_n)\}^{1/2}k_n^{-1/2},$$

*avec $c_2 > 0$ et $\mathfrak{D} = \inf_u \inf_{K<K^*(u)} \Delta L({}^u\boldsymbol{\theta}^{*K}(\mathbf{x}), {}^u\hat{\boldsymbol{\theta}}^{K^*}(\mathbf{x}))$.*

*Sous l'ensemble des conditions, nous obtenons pour les constantes positives $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_3$ du théorème 1.3.1 les majorations suivantes des probabilités de retenir un sous-arbre avec un nombre de feuilles erroné,*

$$\begin{aligned}\forall K > K^*(u),\, \mathbb{P}(\widehat{K}(u) = K) \;\leq\; & 2\exp\left(-\frac{C_1 k_n \lambda^2 (K - K^*(u))^2}{\log k_n^2}\right) \\ & +2\exp\left(-\frac{C_2 k_n \lambda (K - K^*(u)))}{\log k_n}\right) \\ & +\frac{C_3}{k_n^{5/2}\lambda^3 (K - K^*(u))^3},\end{aligned}$$

$$\begin{aligned}\forall K < K^*(u),\, \mathbb{P}(\widehat{K}(u) = K) \;\leq\; & 4\exp\left(-\frac{C_1 k_n \{\mathfrak{D} - \lambda(K^*(u) - K)\}^2}{(\log k_n)^2}\right) \\ & +4\exp\left(-\frac{C_2 k_n \{\mathfrak{D} - \lambda(K^*(u) - K)\}}{\log k_n}\right) \\ & +\frac{2C_3}{k_n^{5/2}\{\mathfrak{D} - \lambda(K^*(u) - K)\}^3},\end{aligned}$$

*Nous pouvons notamment majorer l'espérance de l'erreur stochastique de la manière suivante pour une constante positive $C_5$ dépendante de ${}^u\hat{\boldsymbol{\theta}}^{K^*}(\mathbf{x})$.*

$$\mathbb{E}\left[\|{}^u\hat{\boldsymbol{\theta}}^{\hat{K}}(\mathbf{x}) - {}^u\hat{\boldsymbol{\theta}}^{K^*}(\mathbf{x})\|_2^2\right]^{1/2} \leq \frac{\mathcal{C}_5^{1/2} K^*(u)^{1/2} \log k_n}{k_n^{1/2}}.$$

Nous remarquons dans la première partie du théorème l'obtention d'inégalités de concentration sur les probabilités de sélectionner un sous-arbre erroné, dont les bornes sont en partie exponentielle et en partie polynomiale, comme pour le théorème précédent. Les bornes sont décroissantes en fonction du nombre approximatif de données $k_n$ et également de l'écart par rapport au nombre de feuilles cible $|K - K^*(u)|$. La seconde partie du théorème est un résultat oracle montrant que la sélection du sous-arbre à $\hat{K}$ feuilles, sans la connaissance du nombre de feuilles optimal $K^*(u)$, ne dégrade pas l'ordre de grandeur de la performance.

**Remarque 1.3.3** *Les constantes mentionnées reprennent les notations de la contribution mais ne correspondent pas exactement, afin de proposer un résultat simplifié tout en maintenant une interprétation cohérente.*

#### 1.3.1.3 Étapes principales de la démonstration

Les résultats reposent globalement sur la proximité entre des sommes empiriques de fonctions d'excès i.i.d. et de leurs espérances. Notons ainsi $\mathcal{D}_u$ la fonction qui a toute fonction $\psi$ associe cette différence :

$$\mathcal{D}_u(\psi) = \frac{1}{k_n}\sum_{i=1}^{n}\psi(Y_i)\mathbf{1}_{Y_i>u} - \mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{n}\psi(Y_i)\mathbf{1}_{Y_i>u}\right].$$

Pour un nombre de feuilles $K$ fixé, le résultat repose plus précisément sur la proximité des dérivées empiriques et espérées de la fonction de vraisemblance, par rapport aux paramètres d'échelle et de forme, afin de garantir la qualité des estimations locales par optimisation. Nous cherchons ainsi des inégalités de concentration sur les quantités suivantes :

$$\mathcal{D}_u\left(\frac{\partial\phi(\cdot-u,\boldsymbol{\theta})}{\partial\sigma}\right) \text{ et } \mathcal{D}_u\left(\frac{\partial\phi(\cdot-u,\boldsymbol{\theta})}{\partial\gamma}\right).$$

Pour la sélection du sous-arbre à $\hat{K}$ feuilles, le résultat repose plutôt sur la proximité des variations empiriques et espérées de la vraisemblance globale par rapport au nombre de feuilles. Cela permet de garantir la sélection du nombre de feuilles attendu $K^*(u)$. Nous cherchons ainsi des inégalités de concentration sur les quantités suivantes :

$$\mathcal{D}_u\left(\phi(\cdot-u,\boldsymbol{\theta})\right).$$

Or, nous pouvons majorer uniformément les fonctions $\partial_\sigma\phi(\cdot-u,\boldsymbol{\theta})$ et $\partial_\gamma\phi(\cdot-u,\boldsymbol{\theta})$ par $\Phi(\cdot)=\log(1+w\cdot)$, à une constante multiplicative près et avec $w=\gamma_{\max}/\sigma_{\min}$. Par ailleurs $\phi(\cdot-u,\boldsymbol{\theta})$ est bornée par $\log\sigma_n+\Phi(\cdot)=O(\log(k_n))+\Phi(\cdot)$. Nous allons ainsi nous concentrer sur $\mathcal{D}_u(\phi)$, que nous analysons en deux parties :

$$\mathcal{D}_u(\phi)=\mathcal{D}_u(\phi(\cdot)\mathbf{1}_{\Phi(\cdot)\le M_n})+\mathcal{D}_u(\phi(\cdot)\mathbf{1}_{\Phi(\cdot)>M_n}) \text{ avec } M_n=\beta\log k_n \text{ pour } \beta>0.$$

- pour le premier terme, nous obtenons une inégalité exponentielle uniforme sur $u$ en combinant deux inégalités (Talagrand, 1994) et (Einmahl, Mason, et al., 2005),
- pour le second terme, nous obtenons une inégalité polynomiale en supposant l'existence d'un moment exponentiel pour $\Phi(Y)$ et en appliquant l'inégalité de Markov.

Cette dernière hypothèse est vérifiée pour des variables dans le domaine de Fréchet, $\Phi(Y)$ étant asymptotiquement équivalent à $\log(Y)$. Or, $Y$ admet des moments finis pour

des ordres strictement inférieurs à $1/\gamma$. En revanche, les constantes intervenant dans les majorations seront d'autant plus élevées que l'ordre de ce moment exponentiel fini est faible. Il est à noter que, outre des hypothèses classiques sur la convergence du maximum de vraisemblance, et des hypothèses naturelles pour contrôler le nombre minimal d'observations par feuille, les hypothèses d'obtention de ce résultat sont relativement minimales.

### 1.3.2 Analyse empirique des performances

Les résultats théoriques sur les arbres de régression étant conditionnés à la qualité des partitions successives des sous-arbres, nous analysons la pertinence des partitions de manière empirique. Pour ce faire, nous considérons 3 covariables qualitatives $\mathbf{X} = \{X_1, X_2, X_3\}$ avec chacune 4 valeurs $\{a, b, c, d\}$. Nous supposons de plus un modèle d'arbre GPD à 3 feuilles résultant de divisions effectuées sur les deux premières covariables $\{X_1, X_2\}$, la troisième covariable n'influençant pas le comportement de la variable d'intérêt $Y$. Cet arbre est illustré dans la figure 1.1 et se caractérise de la manière suivante :

$$
\begin{gathered}
Y \sim GPD\left(\sigma(\mathbf{X}); \gamma(\mathbf{X})\right), \text{avec} \\
X \subset \mathcal{X} \subset \mathbb{R}^d, \text{ avec } d = 3 \text{ , } \mathcal{X} = \{a, b, c, d\}^3 \text{ et} \\
\sigma(\mathbf{X}) = \sigma_1 \mathbb{1}_{X_1 \in \{c,d\}} + \sigma_2 \mathbb{1}_{X_1 \in \{a,b\}} \mathbb{1}_{X_2 \in \{c,d\}} + \sigma_3 \mathbb{1}_{X_1 \in \{a,b\}} \mathbb{1}_{X_2 \in \{a,b\}} \text{ et} \\
\gamma(\mathbf{X}) = \gamma_1 \mathbb{1}_{X_1 \in \{c,d\}} + \gamma_2 \mathbb{1}_{X_1 \in \{a,b\}} \mathbb{1}_{X_2 \in \{c,d\}} + \gamma_3 \mathbb{1}_{X_1 \in \{a,b\}} \mathbb{1}_{X_2 \in \{a,b\}} \text{ avec} \\
(\sigma_1, \sigma_2, \sigma_3) \in ]0; +\infty[^3 \text{ et } (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{R}^3.
\end{gathered}
$$

Nous simulons un échantillon de données distribuées suivant ce modèle d'arbre GPD. Plus précisément, nous réalisons un tirage déterministe de $\mathbf{X}$ en imposant un nombre égal d'observations dans chaque cube $\{i, j, k\}$ avec $(i, j, k) \in \{a, b, c, d\}^3$ et un nombre égal d'observations dans chaque feuille. Cette condition impose que le nombre de simulations $n$ soit un multiple de 96. Nous simulons ensuite le tirage conditionnel de $Y$.

La qualité du processus d'estimation du modèle théorique est mesuré par les erreurs d'estimation des paramètres mais également par la pertinence de l'estimation de la structure de l'arbre. Nous considérons plus précisément les indicateurs suivants :

- l'erreur quadratique moyenne de l'estimation de $\sigma$ par $\hat{\sigma}$,
- l'erreur quadratique moyenne de l'estimation de $\gamma$ par $\hat{\gamma}$,
- la distance $L_2$ moyenne entre l'importance des variables estimée et celle qui serait issue de l'arbre modèle,

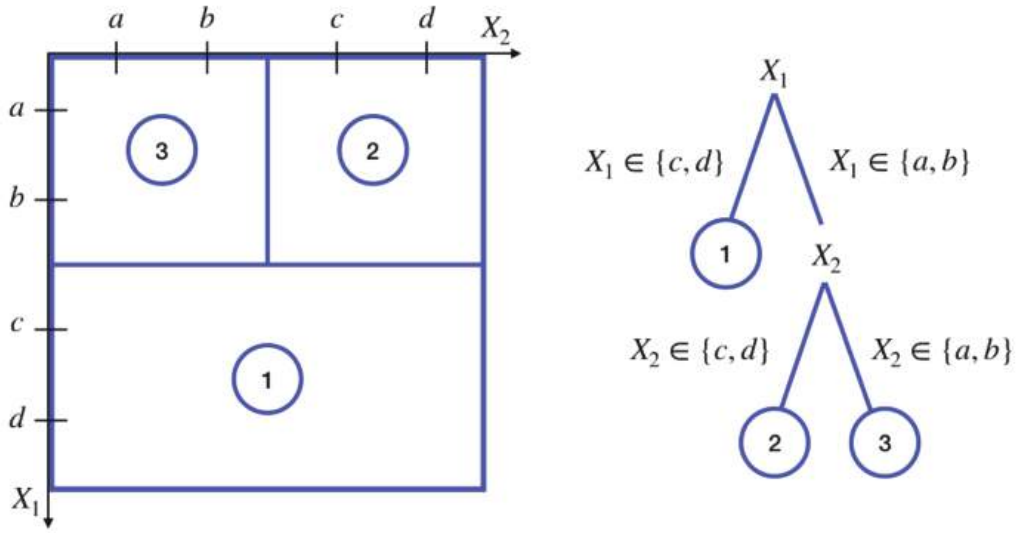



FIGURE 1.1 – Illustration du modèle d'arbre GPD

- le nombre de feuilles de l'arbre élagué,
- le nombre de séparations parfaitement estimées.

Nous proposons trois expériences afin d'analyser les propriétés asymptotiques des arbres GPD, l'influence de l'hétérogénéité des paramètres de forme $(\gamma_1, \gamma_2, \gamma_3)$, et enfin l'influence de leur moyenne. L'étape d'élagage de l'arbre GPD est réalisée au moyen d'une validation croisée et les configurations considérées pour chacune des trois expériences sont les suivantes :

- Pour un nombre croissant de simulations $n \in (96, 384, 768, 1536, 3072, 6144)$, nous fixons les paramètres des GPD à $(\sigma_1, \sigma_2, \sigma_3) = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 1, 1.5)$.
- Pour un nombre de simulations égal à 3072, nous fixons les paramètres d'échelle à $(\sigma_1, \sigma_2, \sigma_3) = (0.5, 1, 1.5)$ et considérons six triplés de paramètres de forme $(\gamma_1, \gamma_2, \gamma_3)$ centrés autour de 1 et aux étendues croissantes :
  - $(0.9375, 1, 1.0625)$, soit une étendue de 0.125,
  - $(0.875, 1, 1.125)$, soit une étendue de 0.25,
  - $(0.75, 1, 1.25)$, soit une étendue de 0.5,
  - $(0.5, 1, 1.5)$, soit une étendue de 1,
  - $(0.25, 1, 1.75)$, soit une étendue de 1.5,
  - $(0, 1, 2)$, soit une étendue de 2.

- Pour un nombre de simulations égal à 3072, nous fixons les paramètres d'échelle à $(\sigma_1, \sigma_2, \sigma_3) = (0.5, 1, 1.5)$ et considérons six triplés de paramètres de forme $(\gamma_1, \gamma_2, \gamma_3)$ aux étendues égales à 1 et centrés autour d'une moyenne croissante :
  - $(-1, -0.5, 0)$, soit une moyenne de 0.125,
  - $(-0.5, 0, 0.5)$, soit une moyenne de 0.25,
  - $(0, 0.5, 1)$, soit une moyenne de 0.5,
  - $(0.5, 1, 1.5)$, soit une moyenne de 1,
  - $(1, 1.5, 2)$, soit une moyenne de 1.5,
  - $(1.5, 2, 2.5)$, soit une moyenne de 2.

Pour chacune de ces expériences, nous réalisons 200 tirages sur lesquels les indicateurs de performance sont évalués. L'amélioration de la précision des paramètres et des partitions estimés au fur et à mesure de l'augmentation du nombre de simulations est illustrée en figure 1.2. Pour $n \in (96, 384)$, la majorité des arbres estimés ne sont pas suffisamment ramifiés mais captent globalement une partie de la structure. Lorsque le nombre de simulations augmente, les structures des arbres estimés tendent vers la structure réelle et les performances s'améliorent. Notons cependant une tendance asymptotique à sélectionner une partition trop granulaire, invitant à un rehaussement de la constante de pénalisation. Les résultats de la seconde expérience illustrent en figure 1.3 l'amélioration de la précision des paramètres et des partitions estimées avec l'augmentation de l'hétérogénéité des comportements extrêmes, tout chose égale par ailleurs. Enfin, les résultats de la troisième expérience mettent en avant en figure 1.4 des performances optimales pour des paramètres de forme compris entre 0 et 1. Les paramètres de forme strictement inférieurs à $-\frac{1}{2}$ sont associés à des performances dégradées et instables. Cette observation est cohérente avec les contraintes de l'estimation d'une GPD par maximum de vraisemblance, dont les propriétés asymptotiques ne sont vérifiées que pour des valeurs strictement supérieures à $-\frac{1}{2}$ (Haan & Ferreira, 2006). Enfin, dans une moindre proportion, l'augmentation des paramètres à partir de 1 semblent légèrement dégrader les prédictions des paramètres et de structure.

### 1.3.3 Application aux événements de failles de données

#### 1.3.3.1 Base de données Privacy Right Clearinghouse

La Privacy Rights Clearinghouse (PRC) est une organisation américaine à but non lucratif dont l'objectif est de contribuer à la protection des données personnelles des amé-

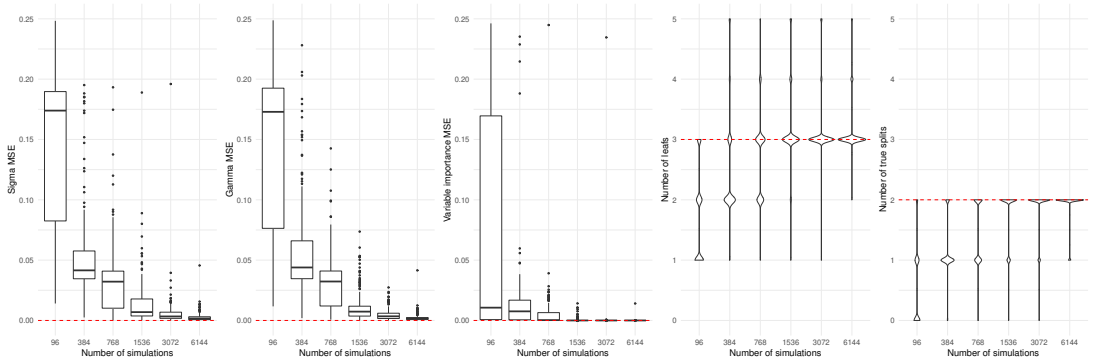



FIGURE 1.2 – Illustration de l'évolution des performances en fonction du nombre de données.

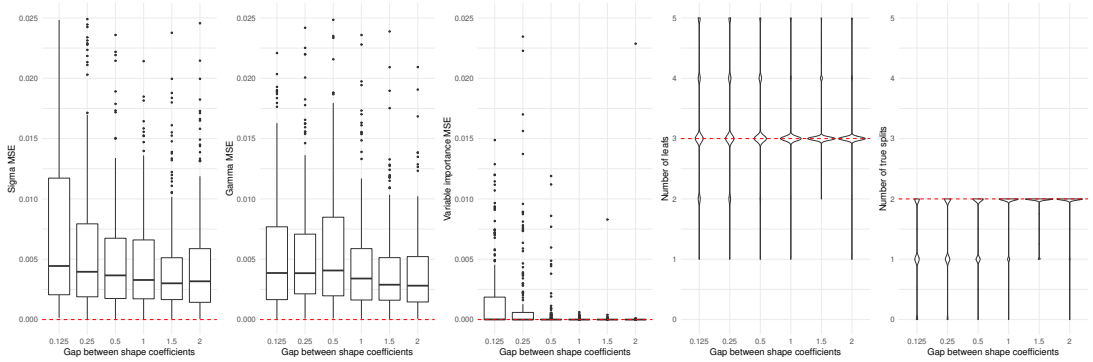



FIGURE 1.3 – Illustration de l'évolution des performances en fonction de l'hétérogénéité des paramètres de forme.

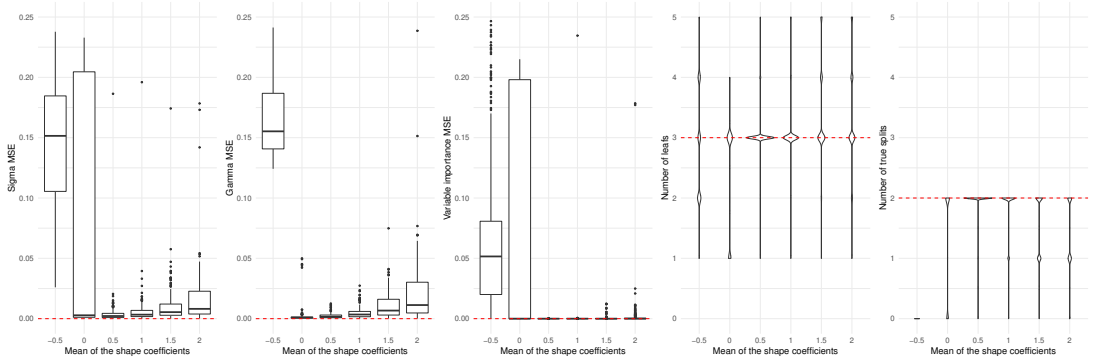



FIGURE 1.4 – Illustration de l'évolution des performances en fonction de la moyenne des paramètres de forme.

ricains, notamment par des actions de sensibilisation. Cette association mène en particulier une veille des failles de données, sur la base des déclarations aux institutions gouvernementales, des révélations médiatiques ou des identifications ciblées. De cette façon, elle maintient à jour une base libre d'accès. Les variables disponibles sont explicitées en Table 1.1. La sévérité de chaque faille peut être estimée à partir du nombre de profils dont les données ont fuité. Les étendues des impacts en fonction des covariables sont significativement différentes. La figure 1.5 souligne par exemple que les événements relayés par les médias sont de plus grande ampleur.

TABLE 1.1 – Liste des variables de la base de données Privacy Right Clearinghouse

| Type de donnée | Variable |
|---|---|
| Données d'exposition | Secteur d'activité |
| | Secteur géographique |
| Données d'événement | Source d'information |
| | Date de l'événement |
| | Type de la faille |
| | Nombre de profils impactés |

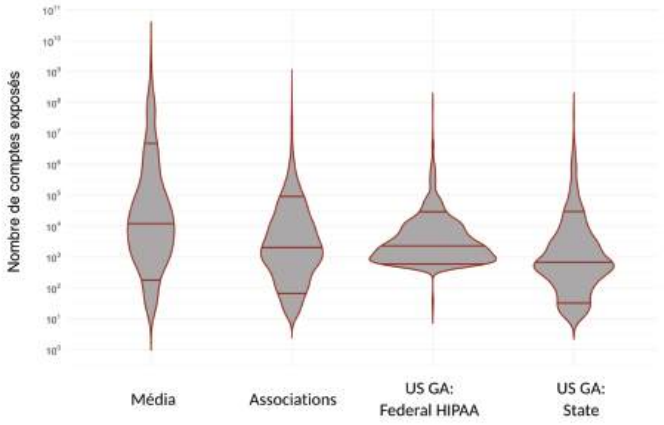



FIGURE 1.5 – Hétérogénéité des ampleurs des failles de données en fonction de la source d'information.

Nous avons d'abord estimé le coût d'une faille de données à partir de son ampleur en suivant des approximations suggérées par Jacobs (2014). Puis, sur ces coûts et ces

covariables, nous avons d'une part estimé une loi Pareto généralisée et d'autre part implémenté la méthode GP CART. Sur l'ensemble des données, nous obtenons l'intervalle de confiance à 95% suivant pour le paramètre de forme : $[0,94; 1,13]$. Globalement, le phénomène semble donc très volatile, et la caractéristique d'une espérance finie ne semble pas garantie. La méthode GP CART permet de distinguer trois types d'événements comme le montre la Figure 1.6. Les secteurs d'activités de la santé, de l'éducation et des organisations à but non lucratif sont ainsi associés à la queue de distribution la plus faible, avec pour le paramètre de forme l'intervalle de confiance à 95% suivant : $[0,66; 0,98]$. Pour autant, la variance du phénomène sous-jacent semble ainsi infinie. Les événements relayés par les médias sont a contrario associés à la queue de distribution la plus lourde avec un paramètre de forme compris à 95% dans l'intervalle $[1,16; 1,98]$ et une espérance infinie. Cette application sur des données publiques motive ainsi la pertinence de la méthode GP CART dans le cadre des stratégies de souscription des risques numériques.

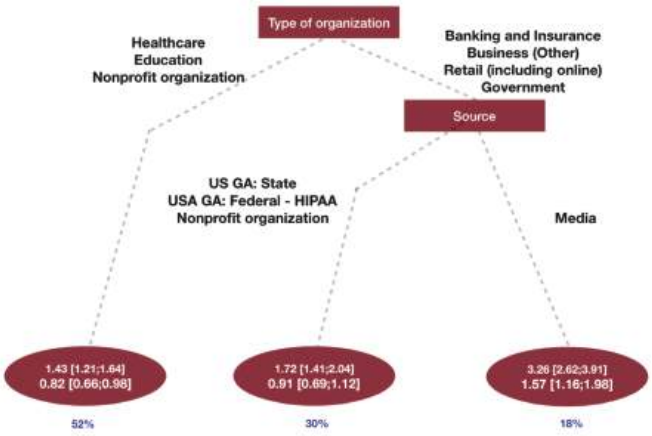



FIGURE 1.6 – Arbre GPD estimé et élagué sur les données d'ampleur des failles de données.

Afin de compléter ces développements relatifs à un portefeuille de prospects ou de clients de produits d'assurance des risques numériques, nous proposons des contributions relatives à la valorisation d'un portefeuille de sinistres numériques, objectif principal du processus opérationnel de provisionnement. Nous considérons deux contraintes à cet exercice : la présence de sinistres ouverts au sein du portefeuille générant des données parcellaires d'une part, et la multiplicité des garanties potentiellement impactées par un sinistre d'autre part. Avant de présenter nos contributions au chapitre 2.3, nous introduisons la

gestion statistique des données censurées au chapitre 2.1 et la gestion statistique des données multidimensionnelles au chapitre 2.2.

# Chapitre 2

# Estimation de copules à partir de données censurées

## Sommaire

## 2.1 Données censurées

### 2.1.1 Introduction

Les organismes d'assurances sont tenus de disposer de données d'une qualité suffisante pour apprécier les risques couverts. Cependant, certaines données quantitatives peuvent être incomplètes lorsque la fenêtre d'observation est limitée. Plus précisément, nous parlons de donnée censurée si la donnée enregistrée correspond par défaut à l'une des bornes de la fenêtre d'observation, et de donnée tronquée si la donnée n'est pas enregistrée en dehors de la fenêtre d'observation. Voici quelques exemples d'informations incomplètes et généralement présentes dans les bases de données des organismes d'assurance.

- le délai de traitement d'un sinistre, défini comme la durée entre la date d'ouverture et de clôture de son dossier, est connu si son dossier est clos. Sinon, le délai n'est que minoré. Dans ce dernier cas, le délai de traitement est dit censuré à droite,
- le coût brut d'un sinistre clos peut être une donnée incomplète. Par exemple, lorsque le préjudice est inférieur à la franchise, le sinistre n'est pas forcément déclaré ou encore indemnisé. Cette donnée est dite tronquée à gauche. A l'inverse, lorsque le dossier a dépassé une limite de garantie, le coût brut est minoré par le montant des paiements à ce titre, et en particulier censuré à droite.
- le coût brut d'un sinistre ouvert est minoré par le montant des paiements déjà effectués à ce titre, et en particulier censuré à droite.

Nous focalisons notre propos sur la problématique du phénomène de censure à droite. Notre contribution concerne en particulier les exemples cités en premier et en dernier. Dans ces deux cas de figure, la censure est liée à une durée, celle du délai de traitement d'un sinistre. L'observation d'une durée nécessitant par définition une certaine attente, son étude a par conséquent contribué au développement de la compréhension mathématique du phénomène de censure à droite.

**Censure à droite : notations**

Introduisons les notations mathématiques liées à la gestion de la censure à droite dans l'application au provisionnement des sinistres numériques.

- $T$ : le délai de traitement d'un sinistre, inconnu pour les dossiers ouverts,
- $C$ : la censure de $T$, supposée connue dans notre application puisque générée par la date de dernière actualisation des données. La censure minore $T$ pour les

sinistres ouverts et la majore pour les sinistres clos,

- $Y$ : le délai observé de traitement d'un sinistre, complet pour les sinistres clos et censuré à droite pour les sinistres ouverts, plus précisément :

$$Y = \min(T, C),$$

- $\delta$ : l'indicatrice de clôture d'un sinistre, illustrant l'observation de la durée $T$, à savoir :

$$\delta = \mathbb{1}_{\{Y=T\}}.$$

Notons $F^X$ les fonctions de répartition de chaque variable aléatoire notée $X$, par exemple $F^T$ pour la fonction de répartition de $T$. Dans le cadre d'une censure à droite, nous étudierons en particulier les fonctions de survie de chaque variable aléatoire, notées $S^X$ et définies par $S^X : x \mapsto \mathbb{P}(X > x) = 1 - F^X(x)$. A partir des données disponibles $(T_i, C_i, Y_i, \delta_i)_{1 \leq i \leq n}$, la prise en compte du phénomène de censure permet de palier les éventuels biais, tels que :

- l'estimation de $S^T$ uniquement à partir des données observées $(Y_i)_{1 \leq i \leq n}$. Dans ce cas, on s'appuie par nature sur des données sous-estimant le phénomène et induisant un biais en ce sens,
- l'estimation de $S^T$ uniquement à partir des données complètes $(T_i)_{\{1 \leq i \leq n / \delta_i = 1\}}$. Dans ce cas, on néglige les données censurées, dont la distribution peut être différente de celle des données complètes, induisant un biais relatif à ces différences.

Afin de modéliser la distribution de $T$, nous introduisons l'estimateur de Kaplan-Meier au paragraphe 2.1.2 en présentant quelques résultats fondamentaux. Une vision davantage élaborée est proposée par Lopez (2007).

### 2.1.2 Estimateur de Kaplan-Meier

Edward Lynn Kaplan et Paul Meier introduisent un estimateur pertinent de la fonction de survie $S^T$ prenant en compte l'ensemble des informations disponibles $(T_i, C_i, Y_i, \delta_i)_{1 \leq i \leq n}$. Nous rappelons ci-dessous sa définition initiale (cf. Kaplan & Meier, 1958) et dans un second temps sa représentation en somme de variables pondérées, introduite par Stute (1995) :

$$\hat{S}^T(y) = \prod_{Y_i \leq y} \left(1 - \frac{\delta_i}{\sum_{k=1}^n \mathbb{1}_{Y_j \geq Y_i}}\right),$$

$$\hat{S}^T(y) = \sum_{i=1}^n \hat{W}_{in} \mathbb{1}_{Y_i \leq y} \text{ avec } \hat{W}_{in} = \frac{\delta_i}{n\hat{S}^C(Y_i^-)},$$

avec $\hat{S}^C$ l'estimateur empirique de la fonction de survie de $C$. La seconde représentation met en lumière un estimateur $\hat{S}^T$ constant par morceaux, dont les sauts sont localisés aux données complètes $(T_i)_{\{1 \leq i \leq n/\delta_i = 1\}}$. Toutefois, à la différence de la fonction de répartition empirique des données complètes, les sauts ne sont pas unitaires mais sont d'autant plus faibles que les censures supérieures à la position du saut sont nombreuses. Cette procédure de pondération est présentée par Van der Laan et Robins (2003) et les poids sont notamment discutés par Satten et Datta (2001). L'estimateur de Kaplan-Meier rééquilibre donc l'importance des données complètes en accentuant d'autant plus leur poids qu'elles ont réussi à dépasser la censure. Autrement dit, l'estimateur de Kaplan Meier peut s'obtenir en calculant la fonction de survie empirique sur une base de données contenant les données complètes et habilement dupliquées. Notre propos s'appuie sur la seconde représentation de l'estimateur de Kaplan-Meier mettant en avant cette pondération. Dans le cas général, pour lequel la censure n'est pas forcément observée, cette représentation est en réalité indirecte puisque les poids $\hat{W}_{in}$ reposent sur une estimation de $\hat{S}^C$ par l'estimateur de Kaplan-Meier, (Stute & Wang, 1993) proposant une expression équivalente. Dans le cas particulier considéré d'une censure complètement observée, les poids $\hat{W}_{in}$ sont plus simples. Pour autant, les résultats théoriques synthétisés ci-dessous sont vérifiés dans le cas général. Pour ce faire, nous émettons tout d'abord trois hypothèses.

**Hypothèse 2.1.1**

*La limite droite du support de $T$ est inférieure à celle du support de $C$.*

$$\inf\{y \in \mathbb{R} / S^C(y) = 0\} \leq \inf\{y \in \mathbb{R} / S^T(y) = 0\}.$$

**Hypothèse 2.1.2**

*La probabilité que $T$ soit égale à $C$ est nulle, $\mathbb{P}(T = C) = 0$.*

**Hypothèse 2.1.3**

*La durée $T$ et la censure $C$ sont des variables aléatoires indépendantes.*

Les hypothèses ci-dessous peuvent être interprétées comme suit. La première hypothèse est nécessaire pour garantir une information parcellaire mais possible de $T$ sur tout son

support. La seconde, plus technique, implique que la permutation entre $T$ et $C$, la première considérée comme la censure et la seconde comme la durée, induit des censures $\tilde{\delta}_i$ inversés, c'est-à-dire $\tilde{\delta}_i = 1 - \delta_i$. Cette flexibilité est essentielle à la construction de l'estimateur $\hat{S}^T$, comme expliqué par (Fleming & Harrington, 2011). La dernière, également technique, est nécessaire à l'obtention des propriétés attendues d'un estimateur.

**Théorème 2.1.4 (Théorème 1.1 de (Stute & Wang, 1993))**
*Soit $\phi$ une fonction. Si les hypothèses 2.1.1, 2.1.2 et 2.1.3 sont vérifiées, alors*

$$\lim_{n\to\infty} \sum_{i=1}^{n} n\hat{W}_{in}\phi(Y_i) = \sum_{i=1}^{n} nW^*_{in}\phi(Y_i) = \mathbb{E}[\phi(Y)], \text{ avec } W^*_{in} = \frac{\delta_i}{n\hat{S}^C(Y_i)}.$$

Ce résultat, assimilable à un résultat de type loi des grands nombres, permet de valider la pertinence asymptotique des estimateurs empiriques reposant sur les poids estimés $\hat{W}_{in}$.

**Théorème 2.1.5 (Théorème de (Akritas, 2000))**
*Soit $\phi$ une fonction. Si les hypothèses 2.1.1, 2.1.2 et 2.1.3 sont vérifiées et si $\phi$ vérifie la condition suivante :*

$$E\left[\frac{\phi(T)^2}{\hat{S}^C(T)}\right] < \infty,$$

*alors nous obtenons une décomposition en sommes i.i.d de l'expression pondérée :*

$$\sum_{i=1}^{n} \hat{W}_{i,n}\phi(Y_i) = \sum_{i=1}^{n} W^*_{in}\phi(Y_i) + \frac{1}{n}\sum_{i=1}^{n} \xi_i(\phi) + o_P(n^{-1/2}), \text{ avec } \mathbb{E}[\xi_i(\phi)] = 0.$$

Ce résultat permet notamment d'obtenir des résultats de type central limite (Stute, 1995) et peut également être étendu à un cadre de régression (Stute, 1999). Dans le cadre de données censurées, nous estimerons ainsi les espérances fonctionnelles ou les intégrales par rapport à la mesure de $T$ grâce à des moyennes empiriques des observations $Y_i$ pondérées selon l'estimateur de Kaplan-Meier $\sum_{i=1}^{n} \hat{W}_{in}\phi(Y_i)$. Ces fonctionnelles vont dépendre des objets statistiques introduits par la suite et permette la modélisation de coûts multidimensionnels : les copules.

## 2.2 Introduction aux copules

### 2.2.1 Introduction

Les produits d'assurance relatifs aux risques numériques regroupent une pluralité de garanties afin de palier les multiples conséquences générées par un sinistre. L'assistance, la perte d'exploitation, les dommages aux biens et la responsabilité civile en sont quelques exemples et peuvent intervenir de concert à la suite d'un même sinistre. Au sein d'un organisme d'assurance, l'observation d'un sinistre peut être synthétisée par les pertes $L^{(1)}, ..., L^{(d)}$ de chacune des $d$ garanties relatives aux risques numériques, créant ainsi des données multidimensionnelles. De manière alternative, la perte globale $L_{sin} = \sum_{i=1}^{d} L^{(i)}$ peut synthétiser l'information de manière unidimensionnelle.

La première approche permet d'analyser les garanties une à une, par exemple pour déterminer des conditions contractuelles adaptées, ou encore pour étudier les interactions entre chaque garantie afin d'anticiper les typologies des dossiers de sinistres encore ouverts. Dans le cadre du provisionnement des sinistres numériques, l'utilisation de données multidimensionnelles peut servir à comprendre le coût d'un sinistre pouvant s'exprimer comme la résultante de plusieurs coûts, associés à chacune des garanties impactées.

**Copule**

Une copule en dimension $d$ est une fonction de répartition sur $[0,1]^d$ avec des marginales uniformes sur $[0,1]$.

**Inverse généralisée**

Si $F$ une fonction de répartition, nous pouvons définir son inverse généralisée à gauche, notée $F^{(-1)}$, de la manière suivante :

$$F^{(-1)}(u) = \inf \{y : F(y) \geq u\}.$$

Les théorèmes de (Sklar, 1959) permettent de décomposer l'étude de la distribution d'une variable multidimensionnelle, en montrant que la loi jointe $F$ du vecteur aléatoire $L^{(1)}, ..., L^{(d)}$ se caractérise par deux composantes complémentaires :

- les lois marginales $F^{(1)}, ..., F^{(d)}$,
- une structure de dépendance entre les marginales, appelée copule. Indépendante des lois marginales, elle repose notamment sur la loi jointe de $F(\mathbf{L})$ dont chaque composante est une variable aléatoire uniforme.

**Théorème 2.2.1 (Proposition 5.2 de (Embrechts, Rüdiger, & Mcneil, 2015))**
*Soient $U$ et $Y$ deux variables aléatoires réelles.*

- *Si $U$ suit une loi uniforme sur $[0;1]$, c'est-à-dire $U \sim \mathcal{U}_{[0;1]}$, alors $F^{(-1)}(U)$ suit la loi définie par $F$, c'est-à-dire que $\mathbb{P}(F^{(-1)}(U) \leq y) = F(y)$*
- *Si $Y$ suit la loi définie par $F$ et que $F$ est continue, alors $F(Y)$ suit une loi uniforme sur $[0;1]$, c'est-à-dire $F(Y) \sim \mathcal{U}_{[0;1]}$.*

**Hypothèse 2.2.2 (Condition d'unicité de la copule)**
*Soit $\mathbf{L}$ un vecteur aléatoire de dimension $d$ avec des marginales associées $L^{(1)}, ..., L^{(d)}$. Supposons chacune des variables aléatoires marginales $L^{(1)}, ..., L^{(d)}$ absolument continue sur son support.*

**Théorème 2.2.3 (Théorème de Sklar (1959))**
*Soit $F$ une fonction de répartition d'un vecteur aléatoire de dimension $d$ avec des marginales associées $F^{(1)}, \ldots, F^{(d)}$. Alors, (cf. Embrechts et al., 2015 ; Schweizer & Sklar, 1983) il existe une copule telle que :*

$$\forall \mathbf{l} = (l^{(1)}, ..., l^{(d)}) \in \mathbb{R}^d, F(l^{(1)}, \ldots, l^{(d)}) = \mathfrak{C}(F^{(1)}(l^{(1)}), \ldots, F^{(d)}(l^{(d)})),$$

*où $F(l^{(1)}, \ldots, l^{(d)}) = \mathbb{P}(L^{(1)} \leq l^{(1)}, \ldots, L^{(d)} \leq l^{(d)})$, $F^{(k)}(l^{(k)}) = \mathbb{P}(L^{(k)} \leq l^{(k)})$, et où la fonction $\mathfrak{C}$, appelée copule, est une fonction de répartition d'un vecteur aléatoire de dimension $d$, dont les marginales sont uniformes sur $[0,1]$.*

*Par ailleurs, lorsque l'hypothèse 2.2.2 est vérifiée, la copule $\mathfrak{C}$ est unique et vérifie $\forall \mathbf{u} = (u_1, \ldots, u_d) \in [0;1]^d$ :*

$$\mathfrak{C}(u_1, \ldots, u_d) = F\left(F_1^{(-1)}(u), \ldots, F_d^{(-1)}(u)\right).$$

Autrement dit, la connaissance des lois des marginales $L^{(1)}, ..., L^{(d)}$ et d'une copule associée $\mathfrak{C}$ permet de déterminer la loi du vecteur $\mathbf{L}$. Par ailleurs, la connaissance de la loi de $\mathbf{L}$ et des marginales permet de déterminer une copule associée $\mathfrak{C}$.

Le théorème de Sklar (1959) met ainsi en évidence l'existence d'une structure de dépendance au sein de toute loi jointe, indépendante des lois marginales. La copule permet par exemple de quantifier la probabilité d'observer simultanément des marginales avec un niveau de quantile différent, ou au contraire, proche d'un même niveau de quantile. La dépendance est d'autant plus forte que l'une de ces probabilités est forte, avec des complémentarités dans le premier cas et des analogies dans le second cas. Le théorème de Fréchet (1960) permet de délimiter ces comportements singuliers.

**Théorème 2.2.4 (Théorème de Fréchet (1960))**
*Une copule à d dimensions* $\mathfrak{C} : [0,1]^d \mapsto [0,1]$ *satisfait les inégalités suivantes (cf. Nelsen, 2007) :*

$$\forall \mathbf{u} = (u_1, \ldots, u_d) \in [0;1]^d,\ W(\mathbf{u}) \leq \mathfrak{C}(\mathbf{u}) \leq M(\mathbf{u}),$$

$$\textit{avec } W(\mathbf{u}) = \max\left(\sum_{i=1}^{d} u_i - d + 1; 0\right) \textit{ et } M(\mathbf{u}) = \min(u_1; \ldots; u_d).$$

Ce théorème précise ainsi les limites de l'espace des copules. La borne inférieure $W$, ne pouvant être atteinte uniquement pour $d = 2$, caractérise une structure de dépendance complémentaire appelée anti-comonotone. A l'inverse, la borne supérieure $M$, toujours atteignable, caractérise une structure de dépendance analogue, appelée comonotone. Au sein de cet espace, un cas particulier de copule est celui de la copule $\Pi$ caractérisant l'indépendance, définie de la manière suivante :

$$\Pi : \mathbf{u} = (u_1, \ldots, u_d) \in [0;1]^d \mapsto \prod_{i=1}^{d} u_i.$$

### 2.2.2 Indicateurs de dépendances bidimensionnelles

Considérons un couple de variables aléatoires $\mathbf{L} = (L^{(1)}, L^{(2)})$, chacune étant respectivement définie par les fonctions de répartition $F^{(1)}$ et $F^{(2)}$. Nous introduisons deux indicateurs permettant de synthétiser l'information de la copule $\mathfrak{C}$ associée à la structure de dépendance de $\mathbf{L}$ : le tau de Kendall et le rho de Spearman. Ces deux indicateurs s'appuient sur les probabilités contraires de concordance et de discordance que nous présentons au préalable.

Soient $\mathbf{L}_1 = (L_1^{(1)}, L_1^{(2)})$ et $\mathbf{L}_2 = (L_2^{(1)}, L_2^{(2)})$ deux vecteurs aléatoires bidimensionnels. La notion de concordance entre deux variables bidimensionnelles est définie comme la possibilité d'ordonner strictement $\mathbf{L}_1$ et $\mathbf{L}_2$. Dans ce cas, on observe deux réalisations différentes mais dont les écarts relatifs de chacune des composantes sont non nuls et du même signe, permettant de les ordonner strictement. La concordance est ainsi vérifiée lorsque $\forall j \in \{1;2\}$, nous observons soit $(L_1^{(j)} > L_2^{(j)})$ soit $(L_1^{(j)} < L_2^{(j)})$. Dès lors, la probabilité de concordance se définit comme la probabilité que $\forall j \in \{1;2\}$, les $(L_1^{(j)} - L_2^{(j)})$ soient non nuls et du même signe, ou de manière équivalente que leur produit soit strictement positif. De même, la probabilité de discordance entre deux variables bidimensionnelles

se définit comme la probabilité que $\forall j \in \{1;2\}$, les $(L_1^{(j)} - L_2^{(j)})$ soient non nuls et de signes opposés, ou de manière équivalente que leur produit soit strictement négatif. Dans ce cas, on observe deux réalisations différentes mais dont les écarts relatifs de chacune des composantes sont non nuls et de signes opposés, aux antipodes de la concordance. Dans le cas de marginales continues, la concordance est l'événement complémentaire de la discordance, l'égalité étant presque sûrement impossible. Ce concept étant introduit, nous présentons successivement le tau de Kendall et le rho de Spearman.

**Tau de Kendall**

Le tau de Kendall de $\mathbf{L} = (L^{(1)}, L^{(2)})$, noté $\tau(L^{(1)}, L^{(2)})$, repose sur deux vecteurs aléatoires $\mathbf{L}_1$ et $\mathbf{L}_2$ i.i.d de même loi que $\mathbf{L}$. Dès lors, il se définit comme la différence entre la probabilité de concordance et celle de discordance des vecteurs aléatoires $\mathbf{L}_1$ et $\mathbf{L}_2$, comme suit :

$$\tau(\mathbf{L}) = \mathbb{P}\left(\left(\prod_{j=1}^{2} L_1^{(j)} - L_2^{(j)}\right) > 0\right) - \mathbb{P}\left(\left(\prod_{j=1}^{2} L_1^{(j)} - L_2^{(j)}\right) < 0\right).$$

**Rho de Spearman**

Le rho de Spearman de $\mathbf{L}$, noté $\rho(L_1^{(1)}, L_1^{(2)})$, repose sur $\mathbf{L}_1$ de même loi que $\mathbf{L}$ et sur le vecteur indépendant $\mathbf{L}_{2\Pi}$ dont les marginales sont identiques à celles de $\mathbf{L}$ mais dont la structure de dépendance est une structure d'indépendance. Autrement dit, $\mathbf{L}_{2\Pi}$ est caractérisé par la fonction de répartition $F_{2\Pi}$ suivante :

$$F_{2\Pi}(l^{(1)}, l^{(2)}) = \mathfrak{C}_{\Pi}\left(F^{(1)}(l^{(1)}), F^{(2)}(l^{(2)})\right) = F^{(1)}(l^{(1)}) \times F^{(2)}(l^{(2)}).$$

Le rho de Spearman est alors défini comme le triple de la différence entre la probabilité de concordance et celle de discordance des vecteurs aléatoires $\mathbf{L}_1$ et $\mathbf{L}_{2\Pi}$, comme suit :

$$\rho(\mathbf{L}) = 3\left[\mathbb{P}\left(\left(\prod_{j=1}^{2} L_1^{(j)} - L_{2\Pi}^{(j)}\right) > 0\right) - \mathbb{P}\left(\left(\prod_{j=1}^{2} L_1^{(j)} - L_{2\Pi}^{(j)}\right) < 0\right)\right].$$

Par définition, le tau de Kendall et le rho de Spearman sont compris entre -1 et 1. Le tau de Kendall et le rho de Spearman étant des indicateurs de la structure de dépendance de $\mathbf{L}$, ils sont caractérisés par la copule $\mathfrak{C}_{\mathbf{L}}$. Les théorèmes suivants précisent leurs formulations analytiques en fonction de la copule.

**Théorème 2.2.5 (Yanagimoto et Okamoto (1969))**
*Soient $(U, V)$ deux variables aléatoires uniformes associées par la copule $\mathfrak{C}_{\mathbf{L}}$. Supposons l'hypothèse 2.2.2 vérifiée pour $F$ la fonction de répartition du vecteur aléatoire $\mathbf{L}$. Alors :*

$$\tau(\mathbf{L}) = 4 \int_{[0;1]^2} \mathfrak{C}_{\mathbf{L}}(u, v) d\mathfrak{C}_{\mathbf{L}}(u, v) - 1 = 4\mathbb{E}[\mathfrak{C}_{\mathbf{L}}(U, V)] - 1$$

**Théorème 2.2.6 (Yanagimoto et Okamoto (1969))**
*Soient $(U, V)$ deux variables aléatoires uniformes associées par la copule $\mathfrak{C}_{\mathbf{L}}$. Supposons l'hypothèse 2.2.2 vérifiée pour $F$ la fonction de répartition du vecteur aléatoire $\mathbf{L}$. Alors :*

$$\rho(\mathbf{L}) = 12 \int_{[0;1]^2} uv d\mathfrak{C}_{\mathbf{L}}(u, v) - 3 = 12\mathbb{E}[UV] - 3 = \frac{\mathbb{E}[UV] - 1/4}{1/12} = \frac{\mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V]}{\sqrt{\mathbb{V}[U]}\sqrt{\mathbb{V}[V]}}$$

Nous remarquons ainsi que, lorsque l'hypothèse 2.2.2 est vérifiée, le rho de Spearman de $\mathbf{L}$ correspond au coefficient de corrélation entre les marginales $F^{(1)-1}(L^{(1)})$ et $F^{(2)-1}(L^{(2)})$. Autrement dit, pour deux variables aléatoires uniformes, le rho de Spearman est égal au coefficient de corrélation. En revanche, de manière générale, le coefficient de corrélation ne peut s'exprimer uniquement en fonction de la structure de dépendance mais également en fonction des marginales. Par ailleurs, dans le cas particulier d'une structure d'indépendance, le tau de Kendall et le rho de Spearman sont égaux au coefficient de corrélation et nuls. La réciproque n'est toutefois pas vérifiée. Par contre, pour les bornes de Fréchet uniquement, ces coefficients caractérisent la structure de dépendance (cf. Embrechts, McNeil, & Straumann, 2002) :

- la structure de dépendance de $\mathbf{L}$ est anti-comonotone si et seulement si $\tau(\mathbf{L}) = -1$, ou de manière équivalente si $\rho(\mathbf{L}) = -1$,
- la structure de dépendance de $\mathbf{L}$ est comonotone si et seulement si $\tau(\mathbf{L}) = 1$, ou de manière équivalente si $\rho(\mathbf{L}) = 1$.

Enfin, dans le cadre de l'étude d'événements extrêmes, la dépendance bi-dimensionnelle aux limites, inférieures et supérieures, des supports des marginales permet de synthétiser la structure de dépendance lors d'événements extrêmes.

**Indicateurs de dépendance extrême**

Les indicateurs de dépendance extrême $\lambda_L$ et $\lambda_U$ se définissent plus précisément comme les limites suivantes, sous condition d'existence, et peuvent par ailleurs s'exprimer en fonction de la structure de dépendance.

$$\lambda_L(\mathbf{L}) = \lim_{\alpha \overset{>}{\to} 0} \mathbb{P}\left(L^{(2)} \leq F_2^{(-1)}(\alpha) | L^{(1)} \leq F_1^{(-1)}(\alpha)\right)$$

$$\lambda_U(\mathbf{L}) = \lim_{\alpha \overset{<}{\to} 1} \mathbb{P}\left(L^{(2)} > F_2^{(-1)}(\alpha) | L^{(1)} > F_1^{(-1)}(\alpha)\right)$$

**Théorème 2.2.7 ( Théorème 5.4.2 de Nelsen (2007))**
*Supposons l'hypothèse 2.2.2 vérifiée pour $F$ la fonction de répartition du vecteur aléatoire* $\mathbf{L}$. *Supposons la première (respectivement la seconde) limite ci-dessus existe, alors nous avons respectivement :*

$$\lambda_L(\mathbf{L}) = \lim_{\alpha \overset{>}{\to} 0} \frac{\mathfrak{C}_{\mathbf{L}}(\alpha, \alpha)}{\alpha}$$

$$\lambda_U(\mathbf{L}) = 2 - \lim_{\alpha \overset{<}{\to} 1} \frac{1 - \mathfrak{C}_{\mathbf{L}}(\alpha, \alpha)}{1 - \alpha}$$

### 2.2.3 Exemples de familles de copules

Les structures de dépendance forment donc un très vaste espace, d'autant plus vaste que le nombre de marginales est élevé. En pratique, il est ainsi intéressant de se restreindre à des classes de structures de dépendances plus singulières. La classe des copules Archimédienne que nous introduisons en est un exemple.

**Théorème 2.2.8 (Théorème 2.2 de McNeil et Nešlehová (2009))**
*Soit $\mathfrak{C}_\Phi$ une fonction définie de la manière suivante par une fonction $\Phi$ appelée générateur.*

$$\mathfrak{C}_\Phi : \mathbf{u} \mapsto \Phi^{-1}\left(\sum_{i=1}^{d} \Phi(u_i)\right).$$

$\mathfrak{C}_\Phi$ *est une copule de dimension $d$ si et seulement si son générateur $\Phi$ est une fonction $d-$monotone sur $[0; +\infty[$, c'est-à-dire dérivable $(d-2)$ fois et vérifiant $\forall x \in [0; +\infty[$ :*

$$(-1)^k \Phi^{(k)}(x) \geq 0, \text{ pour } 0 \leq k \leq d-2,$$

$$\text{et } (-1)^{d-2}\Phi^{(d-2)} \text{ décroissante et convexe sur } [0; +\infty[.$$

*L'espace des copules Archimédiennes est ainsi défini par l'ensemble des générateurs $\Phi$ $d-$monotone sur $[0; +\infty[$.*

La classe des copules Archimédiennes regroupe notamment trois sous familles paramétriques de copules particulières qui illustrent des dépendances localisées dans $[0;1]^d$ :

$$\text{Clayton (1978)} : \Phi(x) = \frac{x^{-\theta} - 1}{\theta}, \text{ pour } \theta \in [-1; 0[\cup]0; +\infty[,$$

$$\text{Frank (1979)} : \Phi(x) = -\log \frac{\mathrm{e}^{-\theta x} - 1}{\mathrm{e}^{-\theta} - 1} \text{ pour } \theta \in \mathbb{R}^*,$$

$$\text{Gumbel (1960)} : \Phi(x) = (-\log(x))^{\alpha} \text{ pour } \theta \geq 1.$$

Les indicateurs de dépendance bidimensionnelle précédemment introduits et dont on connaît une écriture analytique pour ces trois familles de copules sont détaillés ci-dessous. La famille de Clayton se démarque par une dépendance marquée au voisinage de $\mathbf{0} = (0, \ldots, 0)$, la famille de Frank par une dépendance centrale au voisinage de $\mathbf{0.5} = (0.5, \ldots, 0.5)$, et Gumbel, avec une dépendance marquée au voisinage de $\mathbf{1} = (1, \ldots, 1)$.

| Famille | $\tau$ | $\rho$ | $\lambda_L$ | $\lambda_U$ |
|---|---|---|---|---|
| Clayton | $\frac{\theta}{\theta+2}$ [a] | - | $2^{-1/\theta}$ [a] | $0$ [a] |
| Frank | $1 - \frac{4}{\theta}(1 - D_1(\theta))$ [b] | $1 - \frac{12}{\theta}(D_1(\theta) - D_2(\theta))$ [c] | $0$ [a] | $0$ [a] |
| Gumbel | $1 - \frac{1}{\theta}$ [a] | - | $0$ [a] | $2 - 2^{1/\theta}$ [a] |

- [a] Embrechts, Lindskog, et Mcneil (2001),
- [b] Genest et MacKay (1986),
- [c] R. B. Nelsen (1986), avec $D_k(x) = \frac{k}{x^k} \int_0^x \frac{t^k}{e^t - 1} dt$.

### 2.2.4 Techniques d'estimation d'une copule

Nous présentons deux techniques d'estimation d'une copule associée à $n$ réalisations i.i.d d'un vecteur aléatoire $\mathbf{L} = (L^{(1)}, ..., L^{(d)})$. La première méthode suppose l'appartenance de la copule à une famille paramétrique tandis que la seconde est non-paramétrique. Toutes deux reposent sur les pseudo-observations $\left(\hat{\mathbf{U}}_i\right)_{1 \leq i \leq n}$, définies composante par composante à partir des observations $(\mathbf{L}_i)_{1 \leq i \leq n}$ et des fonctions de répartitions empiriques des marginales, de la manière suivante :

$$\forall j \in \{1, \ldots, d\}, \forall i \in \{1, \ldots, n\}, \hat{U}_i^{(j)} = \hat{F}^{(j)}(L_i^{(j)}), \text{ avec } \hat{F}^{(j)}(l) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{L_i^{(j)} \leq l}.$$

Autrement dit, chaque composante $j$ des pseudo-observations se compose de l'ensemble des transformations $\left(\hat{U}_i^{(j)}\right)_{1\leq i\leq n}$ suivant une distribution d'autant plus proche d'une loi uniforme que l'estimation $\hat{F}^{(j)}(l)$ est juste. Le théorème 2.2.1 montre en effet qu'en cas d'estimation parfaite, les marginales des pseudo-observations sont uniformes. Les pseudo-observations sont ainsi des estimations des réalisations de $\mathbf{U}$, concentrant l'information nécessaire à l'estimation de la structure de dépendance $\mathfrak{C}$.

L'estimation paramétrique suppose que la copule caractérisant la structure de dépendance, notée $\mathfrak{C}_{\theta_0}$, appartienne à une famille paramétrique de copules $\mathfrak{C}_\Theta = \{\mathfrak{C}_\theta, \theta \in \Theta \subset \mathbb{R}^k\}$ de dimensions finies. L'estimation de $\theta_0$ par maximum de vraisemblance canonique, introduite par Heyde (1997), est définie de la manière suivante :

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log\left(\mathfrak{c}_\theta(\hat{\mathbf{U}}_i)\right).$$

Les propriétés de convergence de cet estimateur semi-paramétrique sont notamment étudiées (cf. Fermanian & Wegkamp, 2012 ; Genest & Rivest, 1993 ; Tsukahara, 2005).

De manière alternative, Deheuvels (1979) introduit une technique d'estimation non-paramétrique par la fonction de répartition empirique des pseudo-observations $\left(\hat{\mathbf{U}}_i\right)_{1\leq i\leq n}$ :

$$\hat{\mathfrak{C}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{\mathbf{U}}_i\leq\mathbf{u}}.$$

Les propriétés de convergence de cet estimateur non-paramétrique sont notamment étudiées (cf. Fermanian, Radulovic, & Wegkamp, 2004 ; Segers, 2012).

Enfin, l'estimation de copules dans un contexte de censure à droite a été étudiée, dans le cadre semi-paramétrique (cf. Shih & Louis, 1995) comme dans le cadre non-paramétrique (cf. Geerdens, Janssen, & Veraverbeke, 2016 ; Gribkova & Lopez, 2015).

## 2.3 Provisionnement des sinistres numériques

### 2.3.1 Motivation

De manière générale, la valorisation des engagements à venir des organismes d'assurance au titre des sinistres, comptabilisés en tant que provisions techniques au sein des bilans comptables, est un exercice de prédiction dont les variations impactent directement le compte de résultat et la solvabilité de l'entreprise. Concrètement, l'exercice consiste généralement à estimer le montant ultime du coût global des sinistres rattachés à chaque année de couverture assurantielle, puis à déduire les flux de paiements correspondants et déjà comptabilisés afin d'obtenir la valorisation des provisions techniques. La valorisation de la sinistralité d'une année de survenance donnée est ainsi principalement estimée en début d'année, les flux comptables étant encore parcellaires. Ces derniers se développent au fur et à mesure pour atteindre une valorisation finale appelée ultime, observée après un certain nombre d'années.

La principale méthode utilisée à cette fin est la méthode dite de Chain Ladder, reposant sur une représentation agrégée des données de sinistralité en triangle. Chaque ligne correspond à une année de couverture assurantielle et renseigne l'évolution des flux de paiements cumulés observés au fil du développement de la sinistralité. Par construction, les données ainsi agrégées forment un triangle : plus l'année est ancienne et plus le recul est important, permettant l'observation d'un nombre plus important de développements. Pour chaque couple de colonnes associées à des développements successifs, la méthode Chain Ladder estime un facteur de développement, supposant une succession d'évolutions multiplicatives. Cette méthodologie permet donc d'estimer les montants ultimes liés aux sinistres en répliquant les développements historiques sur les dernières informations encore partielles à disposition, contenues dans la dernière diagonale du triangle.

De manière complémentaire, il peut être intéressant d'estimer les montants de sinistres ultimes sur les données sous-jacentes, sinistre par sinistre. Il l'est d'autant plus pour les risques émergents comme les risques numériques pour lesquels la sinistralité est davantage mouvante, tout comme ses développements successifs. Ces techniques étant statistiquement plus complexes, la pertinence de leurs applications dépend des informations disponibles. Dans le cas des risques numériques, le nombre de sinistres est encore faible mais pourrait par exemple être compensé par une information plus précise des circonstances et des conséquences de chaque sinistre.

Dans cette situation hypothétique, une analyse des coûts ultimes de chacun des si-

nistres permettrait d'expliquer l'hétérogénéité globale des coûts par des comportements discriminants. Les prédictions de la sinistralité ultime de chacune des années de couverture seraient ainsi adaptées en fonction des différentes typologies de sinistres. En plus d'anticiper une éventuelle déviation des développements futurs par rapport aux développements historiques, ces techniques permettraient une vision complémentaire et détaillée de la rentabilité. Or, dans un marché en développement, ces résultats constitueraient des éléments d'informations souhaitables pour les prises de décision.

Afin de contribuer à la quantification des montants ultimes des sinistres liés aux risques numériques, nous proposons un cadre d'analyse des données sinistres en associant la prise en compte de la censure des dossiers de sinistres encore ouverts, des circonstances de chaque sinistre et enfin des natures multiples des coûts observés.

### 2.3.2 Cadre d'apprentissage d'un modèle de provisionnement

Dans cette optique, nous supposons disposer des informations suivantes sur les sinistres clôturés :

- le délai de traitement $T$ du dossier de sinistre.
- la décomposition des coûts par type de poste $\mathbf{L} = (L^{(1)}, ..., L^{(d)})$. A titre d'exemple, les garanties de responsabilité, de dommages, de perte d'exploitation, ou encore de réputation peuvent être dissociées. De même, nous pouvons distinguer les frais annexes liés par exemple aux expertises, à l'assistance ou aux suites juridiques. Le coût total s'obtient alors en sommant les coûts de chacue composante $L_{tot} = \sum_{k=1}^{d} L^{(k)}$,
- les variables explicatives $\mathbf{X} \in \mathbb{R}^p$, portant par exemple sur le type de risque assuré ou encore sur le type de sinistre,

Cependant, pour les dossiers de sinistres encore ouverts, toutes les informations ne sont pas encore disponibles. Nous supposons en particulier disposer d'informations incomplètes sur les coûts, supposés censurés à droite par 0, à savoir $\mathbf{M} = (M^{(1)}, ..., M^{(d)}) = \mathbf{0}$ et observer les données suivantes :

- le délai de traitement censuré du dossier de sinistre, noté $C$,
- l'ensemble des variables explicatives $\mathbf{X} \in \mathbb{R}^p$,

Ainsi globalement, pour l'ensemble des $n$ sinistres ouverts ou clôturés, nous supposons observer des réalisations i.i.d. $(\delta_i, Y_i, \mathbf{M}_i, \mathbf{X}_i)_{1 \leq i \leq n}$ de $(\delta, Y, \mathbf{M}, \mathbf{X})$ avec :

$$\begin{cases} \delta & = & \mathbf{1}_{T \leq C}, \\ Y & = & \inf(T, C), \\ \mathbf{M} & = & (\delta L^{(1)}, ..., \delta L^{(d)}) \\ \mathbf{X} & = & \mathbf{X}, \end{cases}$$

Par ailleurs, nous supposons une flexibilité globale de la loi de $(\delta, Y, \mathbf{M}, \mathbf{X})$ contrainte par le cadre suivant :

- Concernant la censure $C$, nous la supposons indépendante de $(T, \mathbf{L}, \mathbf{X})$ et avec un support identique à celui de $T$. Dans ces conditions, la méthode des poids IPCW peut en effet être appliquée,
- A propos des fonctions de répartition des coûts marginaux $L^{(1)}, ..., L^{(d)}$, nous les supposons conditionnelles à $\mathbf{Z} = (\mathbf{X}, T)$, selon un modèle paramétrique, semi-paramétrique ou non paramétrique :

  $$\forall k \leq d, F^{(k)}(l_k|\mathbf{Z} = \mathbf{z}) = m^{(k)}(\mathbf{z}) + \varepsilon^{(k)}, \text{ avec}$$

  $m^{(k)}$ une fonction appartenant à une classe potentiellement infinie et $\varepsilon^{(k)}$ un résidu.
- A propos de la structure de dépendances des coûts marginaux, nous la supposons paramétrique, appartenant à une famille $\mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, avec $\Theta \subset \mathbb{R}^m$, et potentiellement conditionnelle à $\mathbf{Z} = (\mathbf{X}, T)$ :

  $$F(l^{(1)}, ..., l^{(d)}|\mathbf{Z} = \mathbf{z}) = \mathfrak{C}_{\theta(\mathbf{z})}(F^{(1)}(l^{(1)}|\mathbf{Z} = \mathbf{z}), ..., F^{(d)}(l^{(d)}|\mathbf{Z} = \mathbf{z})), \text{ avec :}$$

  - soit une hypothèse de dépendance simplifiée : $\theta(\mathbf{z}) = \theta_0 \in \Theta$,
  - soit une dépendance conditionnelle d'une certaine régularité, précisée dans les hypothèses.

L'utilisation de l'estimateur de Kaplan-Meier permet de corriger les biais liés à la censure, le principal étant la vraisemblable sous-estimation de $L_{tot}$ si les observations sont équitablement pondérées. En effet, les dossiers de sinistres longs à clôturer sont généralement plus coûteux et moins représentés dans les sinistres clos. Ce cadre étant defini, nous cherchons par la suite à estimer la structure de dépendance à travers $\theta_0$ dans le cas simplifié et $\theta(\mathbf{z})$ sinon.

### 2.3.3 Architecture de la démonstration

#### 2.3.3.1 Contexte et objectif

L'objectif de la méthodologie est ainsi d'estimer $\theta_0$ par $\hat{\theta}$ dans le modèle de copule simplifié, ou respectivement $\theta_0(\mathbf{z})$ par $\hat{\theta}(\mathbf{z})$ dans le modèle de copule semi-paramétrique. Nous distinguons pour chacun des cas l'erreur stochastique et l'erreur d'approximation. Selon les notations introduites ci-dessous, l'erreur stochastique correspond à la différence entre $\hat{\theta}$ (resp. $\hat{\theta}(\mathbf{z})$) et $\theta^*$ (resp. $\theta_h^*(\mathbf{z})$). L'erreur d'approximation est égale à la différence entre $\theta^*$ (resp. $\theta_h^*(\mathbf{z})$) et $\theta_0$ (resp. $\theta_0(\mathbf{z})$). Le tableau ci-dessous précise les différentes quantités d'intérêts liées aux M-estimateurs de $\theta_0$ et leurs définitions.

| | | | | | | |
|---|---|---|---|---|---|---|
| $N(\theta)$ | $=$ | $E\left[\log \mathfrak{c}_\theta(\mathbf{U})\right]$ | , et | $\theta_0$ | $=$ | $\arg\max_\theta N(\theta)$, |
| $N_n^*(\theta)$ | $=$ | $\sum_{i=1}^n W_{i,n}^* \log \mathfrak{c}_\theta(\mathbf{U}_i^*)$ | , et | $\theta^*$ | $=$ | $\arg\max_\theta N_n^*(\theta)$, |
| $\hat{N}_n(\theta)$ | $=$ | $\sum_{i=1}^n \hat{W}_{i,n} \log \mathfrak{c}_\theta(\hat{\mathbf{U}}_i)$ | , et | $\hat{\theta}$ | $=$ | $\arg\max_\theta \hat{N}_n(\theta)$, |
| $M(\theta, \mathbf{z})$ | $=$ | $E\left[\log \mathfrak{c}_\theta(\mathbf{U})\|\mathbf{Z}=\mathbf{z}\right] f_{\mathbf{Z}}(\mathbf{z})$ | , et | $\theta(\mathbf{z})$ | $=$ | $\arg\max_\theta M(\theta, \mathbf{z})$, |
| $M_n^*(\theta, \mathbf{z})$ | $=$ | $\frac{1}{h^{p+1}} \sum_{i=1}^n W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_{(\theta,\mathbf{z})}(\mathbf{U}_i^*)$ | , et | $\theta_h^*(\mathbf{z})$ | $=$ | $\arg\max_\theta M_n^*(\theta, \mathbf{z})$, |
| $\hat{M}_n(\theta, \mathbf{z})$ | $=$ | $\frac{1}{h^{p+1}} \sum_{i=1}^n \hat{W}_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_{(\theta,\mathbf{z})}(\hat{\mathbf{U}}_i)$ | , et | $\hat{\theta}(\mathbf{z})$ | $=$ | $\arg\max_\theta \hat{M}_n^*(\theta, \mathbf{z})$. |

Nous proposons dans cette partie une analyse des propriétés de l'erreur globale dans le cas simplifié et de l'erreur stochastique dans le cas semi-paramétrique. Ces erreurs peuvent se décomposer en plusieurs axes :

- les écarts entre les poids optimaux $W_{i,n}^*$ et leurs estimations $\hat{W}_{i,n}$,
- les écarts entre les pseudo-observations $U_i^*$ et leurs estimations $\hat{U}_i$,
- et la pertinence du lissage dans le cadre de la copule semi-paramétrique.

#### 2.3.3.2 Résultats

**Théorème 2.3.1** *Pour le modèle simplifié et dans les conditions précisées dans notre contribution, y compris la définition d'un certain $\xi_i^\Phi$, nous obtenons la décomposition i.i.d suivante de l'erreur globale.*

$$(\hat{\theta} - \theta_0) = -\Sigma^{-1} \left[ \sum_{i=1}^n W_{in}^* \phi(\mathbf{U}_i, Y_i, \mathbf{X}_i) + \frac{1}{n} \sum_{i=1}^n \xi_i^\phi + \sum_{k=1}^d \Lambda_i^{(k)} \right] + o_P(n^{-1/2}), \text{ avec}$$

$$\begin{cases} \Sigma &= E\left[\nabla^2_\theta \log \mathfrak{c}_{\theta_0}(\mathbf{U})\right] \text{ supposée inversible,} \\ \mathbb{E}[\xi_i^\phi] &= 0, \\ \mathbb{E}[\Lambda_i^{(k)}] &= 0. \end{cases}$$

*Nous pouvons en particulier en déduire un théorème central limite pour une certaine matrice de covariance $V$.*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, \Sigma^{-1} V \Sigma\right),$$

*où $\Longrightarrow$ illustre la convergence en loi.*

Nous observons que la vitesse de convergence de l'erreur d'estimation est de l'ordre de $n^{1/2}$. Par ailleurs, le comportement asymptotique dépend à la fois de l'estimation des pseudo-observations, par l'intermédiaire des $\Lambda_i^{(k)}$, de la censure par l'intermédiaire des $\mu_i^\Phi$ et de l'estimation de la copule à travers $\Sigma$. Les définitions exactes sont précisées dans notre contribution.

**Théorème 2.3.2** *Pour le modèle semi-paramétrique, avec $K$ pour noyau, et dans les conditions précisées dans notre contribution, nous obtenons la décomposition i.i.d suivante de l'erreur stochastique.*

$$\sqrt{n}\left(\hat{\theta}(\mathbf{z}) - \theta_h^*(\mathbf{z})\right) = -\Sigma(\mathbf{z})^{-1}\left\{\frac{1}{h^{p+1}}\sum_{i=1}^n W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right)\Phi(\mathbf{U}_i)\right\} + o_P(n^{-1/2}h^{(p+1)/2}),$$

*avec $\Sigma(\mathbf{z}) = \left(E\left[\Phi^{(j,k)}_{\theta(\mathbf{z})}(\mathbf{U_i})|\mathbf{Z} = \mathbf{z}\right]\right)_{j,k}$ supposée inversible. Nous pouvons en particulier en déduire un théorème central limite en notant $S(\mathbf{z}) = E\left[\Phi(\mathbf{U})|\mathbf{Z} = \mathbf{z}\right]$ et $S(\mathbf{z})'$ sa transposée.*

$$n^{1/2}h^{(p+1)/2}\{\hat{\theta}(\mathbf{z}) - \theta_h^*(\mathbf{z})\} \Longrightarrow \mathcal{N}\left(\mathbf{0}, \Sigma(\mathbf{z})^{-1} S(\mathbf{z}) S(\mathbf{z})' \Sigma(\mathbf{z})^{-1}\right).$$

Dans le cadre semi-paramétrique, la vitesse de convergence de l'erreur d'estimation est sensible à la dimension $p$ des covariables, de l'ordre de $n^{-1/2}h^{(p+1)/2}$. Il s'agit de la vitesse de convergence de référence des estimateurs non-paramétriques à noyau. De même, pour

le terme de biais, $\theta_h^*(\mathbf{z})$ converge vers $\theta(\mathbf{z})$ à la vitesse $O(h^2)$. Par ailleurs, le comportement asymptotique ne dépend ni de la censure ni de l'estimation des lois marginales. Ces deux premières erreurs sont en effet asymptotiquement négligeables par rapport à l'estimation semi-paramétrique de la copule, qui influence le comportement asymptotique de l'erreur stochastique.

#### 2.3.3.3 Résultats sous-jacents et hypothèses

La démonstration repose principalement sur la maîtrise des trois axes de l'erreur, à savoir l'estimation des poids liés à la censure, l'estimation des pseuso-observations et enfin le lissage pour l'estimation semi-paramétrique. La performance de l'estimation des poids repose sur les résultats de convergences (cf. Van der Laan & Robins, 2003) et (cf. Gill, 1983). Nous supposons alors l'existence d'espérances finies. Le contrôle de la vraisemblance de la copule aux bornes du support repose sur une démarche proche de celle développée par Tsukahara (2005). Enfin, la performance de l'estimation semi-paramétrique repose sur l'inégalité de concentration de Einmahl et al. (2005) et sur des hypothèses de régularité. Globalement, notre résultat est indépendant de la méthode d'estimation des marginales, englobant l'estimation paramétrique comme non-paramétrique. La seule condition est d'avoir un développement asymptotique i.i.d. des pseudo-observations.

# Chapitre 3

# Estimation de phénomènes d'accumulation auto-excités

## Sommaire

## 3.1 Processus de comptage

L'enregistrement de dates d'occurrence d'événements au fil du temps peut constituer une donnée d'intérêt, par exemple pour un organisme d'assurance les dates de survenance de sinistres. Au chapitre précédent, nous avons étudié les données multidimensionnelles générées par chacun de ces sinistres. Nous proposons ici de nous concentrer sur l'étude de la fréquence de la sinistralité, c'est-à-dire sur la suite des dates de survenance ou de déclaration des sinistres. Cette donnée peut se modéliser par un processus de comptage, un cas particulier de processus ponctuel, et une filtration. Rappelons quelques définitions fondamentales de la théorie des processus ponctuels. Un processus ponctuel $N = (N_t)_{t\in\mathbb{R}^+}$ est représenté par une famille de variables aléatoires réelles sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, et indexées par $t \in \mathbb{R}^+$. Une filtration exprime l'information croissante détenue à chaque instant par l'observateur et se représente par une suite croissante de sous-tribus $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{R}^+}$. On dit qu'un processus ponctuel $N$ est $\mathcal{F}$-adapté si pour tout $t \in \mathbb{R}^+$ $N_t$ est $\mathcal{F}_t$ mesurable. Nous introduisons le concept de processus de comptage ci-dessous.

**Processus de comptage** (Section 3.1 de D. J. Daley (2008a))
Un processus ponctuel $N$ défini sur $\mathbb{R}^+$ est un processus de comptage si presque sûrement ses trajectoires sont croissantes par saut d'amplitude 1, continues à droites et nulles à l'instant 0.

Autrement dit et comme son nom l'indique, le processus de comptage $N$ saute unitairement aux temps de réalisations, disjoints, d'événements d'intérêts. Pour un horizon $T > 0$, l'observation de $N$ sur $[0; T]$ se caractérise ainsi par ses temps de sauts $(\tau_i)_{i\geq 1}$ inférieurs à $T$ et le temps $T$ correspondant à la fin de l'observation : $\left((\tau_i)_{i\geq 1}, T\right)$.

$$\forall t \leq T, N(t) = \sum_{i\geq 1} \mathbb{1}_{\tau_i \leq t}.$$

### 3.1.1 Notions de compensateur et d'intensité

Nous proposons de rappeler les notions de compensateur et d'intensité d'un processus ponctuel, sur lesquelles notre analyse repose. Toutefois, les concepts fondamentaux de la théorie des processus ponctuels ne sont pas introduits, nous invitons le lecteur à consulter des ouvrages de référence D. J. Daley (2008a). Les notions de processus ponctuel, de filtration, de filtration naturellement adaptée à un processus ponctuel, de martingale, d'in-

tégrale stochastique, de prévisibilité et de localité sont notamment utilisées sans définition préalable.

**Compensateur** (Section 3.3 de D. J. Daley (2008a))
Soit $N$ un processus ponctuel. On appelle compensateur de $N$ un processus croissant, continu à droite et prévisible $\Lambda$ tel que $N - \Lambda$ soit une martingale locale.

**Intensité** (Section 3.3 de D. J. Daley (2008a))
Soit $N$ un processus ponctuel et $\Lambda$ son compensateur. Si presque sûrement toute trajectoire de $\Lambda$ est absolument continue, alors on note $\lambda(t) = \frac{d\Lambda(t)}{dt}$ l'intensité de $N$.

Sous réserve d'existence, l'intensité est ainsi un processus prévisible, presque sûrement localement intégrable et à valeurs positives. L'intensité s'interprète comme la probabilité infinitésimale d'observer un saut conditionnellement aux informations passées.

$$\mathbb{P}(dN_t = 1 | \mathcal{F}_{t^-}) = \lambda(t)dt.$$

Plus l'aire sous la courbe de l'intensité est élevée entre $t_W$ et $t_M$, plus la probabilité d'observer des événements sur cette fenêtre est élevée.

### 3.1.2 Exemples de processus de comptage

Nous introduisons trois familles de processus de comptage : les processus de Poisson homogènes, les processus de Poisson inhomogènes et les processus de Cox. Le processus de Poisson homogène, reprenant le nom de Siméon Denis Poisson (1837), peut se définir de la manière suivante.

**Processus de Poisson homogène** (Section 2.1 de D. J. Daley (2008a))
Soit $N$ un processus de comptage adapté à la filtration $\mathcal{F}$. $N$ est un processus de Poisson homogène de paramètre $\lambda > 0$ si pour tout couple de réels $(t, s)$ vérifiant $t > s > 0$, les assertions suivantes sont vraies :

- $N(t) - N(s)$ suit une loi de Poisson de paramètre $\lambda(t - s)$,
- $N(t) - N(s)$ est indépendant de $\mathcal{F}_s$.

Le processus de Poisson homogène est ainsi toujours en régime stationnaire, au sens où la probabilité d'occurrence d'événements entre $s$ et $t$ ne dépend ni de $s$ ni de $t$, mais plutôt de la différence $t-s$. Par ailleurs, plusieurs caractérisations du processus de Poisson homogène existent, nous en présentons deux d'entre-elles.

**Théorème 3.1.1 (Section 2.1 de D. J. Daley (2008a))**
*Soit $N$ un $\mathcal{F}$-processus de comptage. $N$ est un $\mathcal{F}$-processus de Poisson homogène si et seulement si les temps inter-arrivés $(\tau_{i+1} - \tau_i)_{i\geq 1}$ sont des variables aléatoires indépendantes et identiquement distribuées suivant une loi exponentielle de paramètre $\lambda$.*

**Théorème 3.1.2 (Théorème 2.3 de (Watanabe, 1964))**
*Soit $N$ un $\mathcal{F}$-processus de comptage vérifiant $\mathbb{E}[N(t)] = \lambda t$, c'est-à-dire ayant une intensité constante égale à $\lambda$, alors $N$ est un processus de Poisson homogène adapté à la filtration $\mathcal{F}$. La réciproque est également vraie (cf. Meyer, 1971).*

La première caractérisation permet par exemple de simuler un processus de Poisson homogène à partir de simulations d'une loi exponentielle. La seconde souligne que la probabilité d'occurrence infinitésimale des événements issus d'un processus de Poisson homogène est uniforme dans le temps. Réciproquement, si l'intensité d'un processus de comptage est constant, il s'agit d'un processus de Poisson homogène.

**Remarque 3.1.3** *L'observation d'un processus de Poisson homogène $N$ peut se limiter à sa valeur en $T$ lorsque la trajectoire sur $[0; T($ n'est pas connue. Cette observation se limite alors à la réalisation d'une variable aléatoire suivant une loi de Poisson de paramètre $\lambda T$.*

Plus globalement, la famille des processus de Poisson homogènes appartient à la famille plus large des processus de Poisson inhomogènes, caractérisés par une intensité déterministe.

**Processus de Poisson inhomogène** [Section 2.1 de (D. J. Daley, 2008a)]
Soit $N$ un processus de comptage adapté à la filtration $\mathcal{F}$. $N$ est un processus de Poisson inhomogène d'intensité déterministe $\lambda : s \mapsto \lambda(s)$ si pour tout couple de réels $(t, s)$ avec $t > s > 0$, les conditions suivantes sont vérifiées :

- $N(t) - N(s)$ suit une loi de Poisson de paramètre $\int_s^t \lambda(u)du$,
- $N(t) - N(s)$ est indépendant de $\mathcal{F}_s$.

**Remarque 3.1.4** *L'observation d'un processus de Poisson inhomogène $N$ peut se limiter à sa valeur en $T$ lorsque sa trajectoire sur $[0; T($ n'est pas connue. L'observation se limite alors à la réalisation d'une variable aléatoire suivant une loi de Poisson de paramètre $\int_0^T \lambda(u)du$. Plusieurs réalisations pour un même $T$ permettent de modéliser le nombre total d'événements entre $[0; T]$ mais la compréhension du phénomène aléatoire sous-jacent*

*nécessite des observations au fil du temps, des réalisations de processus de comptage. Le processus de Poisson inhomogène permet par exemple de modéliser des événements dont la probabilité d'occurrence est saisonnière, comme les orages de grêle.*

Nous présentons enfin la famille des processus Cox, également appelés processus de Poisson doublement stochastique. Ces processus généralisent les processus de Poisson inhomogènes et sont caractérisés par une intensité aléatoire.

**Processus de Cox** (Section 6.2 de D. J. Daley (2008a))
Soit $N$ un processus de comptage adapté à la filtration $\mathcal{F}$ dont l'intensité $\lambda : s \mapsto \lambda(s)$ est prévisible par rapport à la filtration $\mathcal{F}$. $N$ est un processus de Cox si et seulement si $N(t) - \int_0^t \lambda(s)ds$ est une $\mathcal{F}$-martingale (cf. Cox, 1955).

Nous remarquons qu'un processus de Cox ne vérifie plus nécessairement l'indépendance de $N(t) - N(s)$ par rapport à $\mathcal{F}_s$. L'intensité d'un processus de Cox est un processus ponctuel qui peut en effet dépendre des informations passées, c'est-à-dire de la filtration.

### 3.1.3 Quelques résultats préliminaires

Nous présentons successivement trois résultats relatifs à l'estimation par maximum de vraisemblance des processus de comptage, au changement d'échelle d'un processus ponctuel et enfin à des résultats de concentration du supremum d'une intégrale stochastique.

**Théorème 3.1.5 (Vraisemblance Rubin (1972))**
*Soit $N$ un processus de comptage associé à une intensité $\lambda$ et un compensateur $\Lambda = \int \lambda$. Nous supposons observer une réalisation de ce processus sur l'intervalle $[0;T]$. La vraisemblance d'observer une suite d'événements aux temps $(\tau_i)_{1\leq i\leq N(T)}$, notée $\mathcal{L}\left(\left((\tau_i)_{i\geq 1}, T\right), N\right)$, s'exprime de manière analytique :*

$$\mathcal{L}\left(\left((\tau_i)_{i\geq 1}, T\right), N\right) = \prod_{i=1}^{N(T)} \left( \lim_{t \overset{<}{\to} \tau_i} \lambda(t) \right) e^{-\Lambda(T)}.$$

Ce résultat permet notamment d'estimer de manière paramétrique tout processus de comptage appartenant à une famille de processus paramétriques $(N_{\lambda_\theta})_{\theta\in\Theta}$. L'estimateur du maximum de vraisemblance $N_{\lambda_{\hat{\theta}}}$ peut en effet être obtenu en maximisant la vraisemblance empirique, analytiquement ou numériquement.

**Remarque 3.1.6** *L'estimateur du maximum de vraisemblance réalise ainsi un compromis entre les deux termes de l'expression de la vraisemblance, d'un côté la maximisation de l'intensité aux limites à gauche des temps* $(\tau_j)_{1\leq j\leq N(T)}$ *et de l'autre sa minimisation entre les événements :*

- *le premier terme est une fonction croissante des limites à gauche de l'intensité aux temps* $(\tau_j)_{1\leq j\leq N(T)}$ *. Plus ces dernières sont élevées et plus il est en effet vraisemblable que l'observation des événements aux temps* $(\tau_j)_{1\leq j\leq N(T)}$ *découle du modèle* $N$*,*
- *le second terme est une fonction décroissante du compensateur en* $T$*, à savoir de l'intégrale de l'intensité entre* $0$ *et* $T$*.*

*Le compromis peut se résumer ainsi : plus l'intensité candidate est élevée aux événements et plus il est difficile d'obtenir un second terme élevé sous les contraintes de régularité de la famille paramétrique. Inversement, plus l'évaluation du compensateur candidat est faible en* $T$ *et plus il est difficile d'obtenir des évaluations élevées de l'intensité aux événements sous les contraintes de régularité de la famille paramétrique. Dans le domaine des variables aléatoires, le maximum de vraisemblance vise également la maximisation de la densité aux points observés sous la contrainte d'une famille de densité paramétrique dont les intégrales sont par définition unitaires.*

Nous présentons maintenant le théorème de remise à l'échelle, positionnant le processus de Poisson homogène d'intensité constante égale à 1 comme une référence.

**Théorème 3.1.7 (Papangelou (1974))**
*Soit* $M$ *un processus ponctuel adapté à une filtration* $\mathcal{G}$ *avec des valeurs finies et croissantes et un* $\mathcal{G}$*-processus ponctuel de Poisson homogène d'intensité unitaire. Soit* $\mathcal{F}$ *la filtration générée par les* $\sigma-$*algèbres* $\mathcal{F}_t = \mathcal{G}_{M(t)}$*. Alors,* $N(t) = N_0(M(t))$ *est un processus ponctuel adapté à la filtration* $\mathcal{F}$ *et admettant* $M(t)$ *comme compensateur adapté à* $\mathcal{F}$*.*
*Par ailleurs, soit* $N$ *un processus ponctuel adapté à une filtration* $\mathcal{F}$ *avec une* $\mathcal{F}$*-intensité* $\lambda$ *strictement positive et un* $\mathcal{F}$*-compensateur* $\Lambda(t) = \int_0^t \lambda(u)du$ *presque sûrement non borné. Alors, sous la transformation temporelle potentiellement aléatoire* $\Lambda$*, le processus transformé défini par* $\tilde{N} = N(\Lambda^{-1}(t))$ *est un processus de Poisson d'intensité unitaire et adapté à la filtration générée par les* $\sigma-$*algèbres* $\mathcal{G}_t = \mathcal{F}_{\Lambda^{-1}(t)}$*.*

Une interprétation possible de ce résultat est de noter que le processus de Poisson homogène d'intensité unitaire est aux processus ponctuels ce que la loi uniforme est aux

variables aléatoires, en faisant l'analogie avec le théorème 2.2.1. Ce théorème motive l'algorithme de "Thinning" de Lewis et Shedler permettant de simuler des processus ponctuels dont l'intensité est déterministe à partir de simulations de processus de Poisson homogènes unitaires (Lewis & Shedler, 1976), algorithme 7.5.II de (D. J. Daley, 2008a). Lorsque l'intensité est aléatoire, l'algorithme peut être adapté en réalisant des simulations successives et conditionnelles (cf. algorithme 7.5.IV de D. J. Daley, 2008a ; Ogata, 1981). Par ailleurs, le théorème de remise à l'échelle permet d'appliquer les tests statistiques relatifs aux processus de Poisson homogènes unitaires à des processus ponctuels quelconques (cf. Cox & Lewis, 1966 ; Ogata, 1988). Une analyse statistique de cette généralisation est proposée par Reynaud-Bouret, Rivoirard, Grammont, et Tuleau-Malot (2014).

Nous présentons enfin des résultats de concentration relatifs à des intégrales stochastiques par rapport à la martingale locale associée à chaque processus ponctuel $N - \Lambda$. Le premier résultat permet d'identifier une condition suffisante pour que $N - \Lambda$ soit une martingale. Puis, l'inégalité maximale ou sous-martingale de Doob est un résultat de concentration relatif au supremum d'une sous-martingale.

**Théorème 3.1.8 (Lemme 7.2.V. de D. J. Daley (2008a))**
*Soit $N$ un processus ponctuel sur $[0; +\infty[$ adapté à une filtration $\mathcal{F}$. S'il admet une $\mathcal{F}$-intensité continue à gauche $\lambda(t)$, alors le processus $M(t) = N(t) - \Lambda(t)$ est une $\mathcal{F}$-martingale : pour tout $s > t > 0$, $\mathbb{E}[M(s)|\mathcal{F}_t] = M(t)$, avec $\Lambda(t) = \int_0^t \lambda(u)du$.*

**Théorème 3.1.9 (Inégalité maximale ou sous-martingale Doob (1953))**
*Si $M$ est une sous-martingale positive alors pour tout $T > 0$, l'inégalité sur son supremum ci-dessous est vérifiée.*

$$\mathbb{P}(\sup_{0 \leq t \leq T} M(t)) \geq m) \leq \frac{\mathbb{E}[M(T)]}{m}.$$

Cette inégalité de concentration intervient de manière sous-jacente au résultat de concentration utilisé dans notre contribution. D'autres inégalités de concentration relatives aux processus ponctuels existent (cf. Bremaud, 1981).

## 3.2 Processus de Hawkes

### 3.2.1 Motivation et définition

Dans le cadre de l'assurance des risques numériques, nous étudions la fréquence des sinistres, autrement dit leurs dates de survenance successives. A priori, la probabilité d'occurrence infinitésimale ne semble pas constante. En effet, la menace malveillante et les mesures de préventions évoluent. Le processus de Poisson homogène ne semble donc pas être un modèle adapté. L'intensité pourrait par exemple dépendre d'un indicateur de cette résultante, qui s'observerait en continu et constituerait un processus stochastique conjoint à celui du processus de comptage des sinistres. De la même manière le processus de Poisson inhomogène ne permet pas à son intensité d'être dépendante d'observations en lien avec le processus d'intérêt. Par ailleurs, les informations sous-jacentes à l'activité malveillante, aux mesures de préventions et aux dépendances informatiques sont plutôt étrangères à la culture sectorielle des organismes d'assurances et nécessite des expertises informatiques spécifiques. Les assureurs observent partiellement les conséquences sur leurs portefeuilles. Enfin, partant du constat de l'interdépendance, de la centralisation des systèmes informatiques et de l'irrégularité de la menace malveillante, les accumulations de sinistres semblent attendues, comme elles le sont sur le principe - sans réaliser de comparaison quantitative - pour les risques de crédit.

Ainsi, semble-t-il vraisemblable de supposer que le risque d'occurrence des sinistres à venir est plus élevé lorsque la concentration de sinistres survenus récemment dépasse significativement le niveau habituel. Dans le cadre d'un processus de comptage, cette observation correspond à une intensité dépendante des événements passés, caractéristique des processus de Cox. Cependant, un processus de Cox est difficilement calibrable lorsque les données sont limitées. Les processsus de Hawkes, plus facilement calibrables, formalisent de surcroît les phénomènes d'auto-excitations (cf. Hawkes, 1971a, 1971b). Nous étudions plus précisément les processus de Hawkes dont l'intensité augmente à la suite de la survenance d'événements passés, pour traduire des phénomènes d'auto-excitation de la sinistralité numérique.

Les domaines d'application des processus de Hawkes sont nombreux, de la finance (cf. Bacry, Mastromatteo, & Muzy, 2015 ; Hawkes, 2018), à la neurologie (Bonnet, Dion, Gindraud, & Lemler, 2021), en passant par la quantification du risque (cf. Chavez-Demoulin & McGill, 2012 ; Errais, Giesecke, & Goldberg, 2010). Pour certaines applications, le phénomène d'auto-excitation a une explication spécifique ou scientifique. Par exemple,

l'accumulation de tremblements de terre est liée aux répétitions de répliques, dont la loi d'Omori fait l'objet en sismologie. De même, l'accumulation de bâtis inondés s'explique par la propagation des eaux de pluie étudiée en hydrologie. En ce qui concerne les risques numériques, plusieurs causes pourraient induire une accumulation de sinistres, par exemple :

- le phénomène d'auto-réplication d'un programme malveillant,
- la concentration temporelle d'attaques informatiques liées à des opportunités spécifiques,
- la neutralisation au moins partielle d'un jalon central des traitements informatiques (hébergement centralisé, élément répandu).

L'observation d'accumulation de sinistres numériques étant probable, nous proposons d'en étudier les dynamiques de manière statistique, afin d'apporter des éléments complémentaires à une analyse informatique. Nous considérons ainsi un type de processus Hawkes : les processus de Hawkes linéaires auto-excitants.

**Processus de Hawkes linéaire auto-excitant** (Hawkes (1971a))
Un processus de Hawkes linéaire auto-excitant, noté $N^{(\mu,h)}$, est défini par une intensité $\lambda^{(\mu,h)}$ vérifiant :

$$\lambda^{(\mu,h)}(t) = \mu + \int_0^t h(t-s) dN_s^{(\mu,h)}, \text{ avec}$$

- $\mu > 0$ une constante, assimilable à l'intensité d'un processus de Poisson homogène,
- $h$ une fonction dite noyau, de $\mathbb{R}^+$ dans $\mathbb{R}^+$, traduisant le phénomène d'auto excitation et nulle en 0 pour garantir la prédictibilité du processus, comme le montre l'exemple 14.3(c) de D. J. Daley (2008b).

Un premier indicateur intéressant de l'auto-excitation est la norme $L^1$ du noyau $||h||_1 = \int_{\mathbb{R}^+} h(t)dt$, qui correspond à l'espérance du nombre d'événements générés du fait de l'excitation engendrée par un seul événement. La simulation d'un processus de Hawkes peut d'ailleurs se réaliser de manière itérative, en générant un processus de Poisson homogène d'intensité $\mu > 0$ puis en simulant pour chaque événement le nombre d'événements générés par l'auto-excitation ainsi que leurs délais d'apparition. La boucle est presque sûrement finie si et seulement si $\int_{\mathbb{R}^+} h(t)dt < 1$, (cf. Møller & Rasmussen, 2005). De la même manière que la distribution d'un loi Pareto généralisée n'admet d'espérance finie

seulement si son paramètre de forme $\gamma$ est strictement inférieur à 1, les résultats ergodiques, assimilables à la loi des grands nombres, s'appliquent sur les processus de Hawkes uniquement si $\int_{\mathbb{R}^+} h(t)dt < 1$, comme le montre Bacry, Delattre, Hoffmann, et Muzy (2013).

Si cette dernière est supérieure ou égale à 1, le processus de Hawkes est dit instable. Nous considérerons le cas d'un processus de Hawkes stable, avec un noyau $h$ de norme $L^1$ strictement inférieure à 1.

### 3.2.2 Propriétés des processus de Hawkes

L'estimation d'un processus de Hawkes linéaire peut reposer sur la maximisation de la vraisemblance, dans le cas d'une famille de noyau paramétrique. De manière alternative ou dans le cadre d'une estimation non paramétrique introduite par (Bacry & Muzy, 2016), l'espérance et l'auto-corrélation d'un processus de Hawkes stable permettent d'estimer l'intensité d'un processus de Hawkes linéaire.

Nous présentons ci-dessous de manière succincte les écritures analytiques de son espérance et de sa variance.

**Proposition 3.2.1 (Extrait du théorème 2 de (Bacry et al., 2013))**
*Soit $N^{(\mu,h)}$ un processus de Hawkes linéaire, auto-excitant de noyau $h$. Notons $\Lambda_{N^{(\mu,h)}}$ son compensateur. Si le processus est stable, à savoir que $||h||_1 < 1$, alors son espérance est finie et définie par :*

$$\mathbb{E}[N^{(\mu,h)}(t)] = \mathbb{E}[\Lambda_{N^{(\mu,h)}}(t)] = \mu\left(t + \int_0^t H(t-s)sds\right),$$

*avec $H(t) = \sum_{k\geq 1} h^{*k}(t)$, et $(h^{*k})_{k\geq 1}$ définie récursivement par $h^{*1}(t) = h(t)$, et pour $k \geq 2$, $h^{*k}(t) = \int_0^t h(t-s)h^{*(k-1)}(s)ds$, correspondant au produit de convolution $h * h^{*(k-1)}$. Remarquons que par construction, $\int_{\mathbb{R}^+} H(t)dt = \frac{||h||_1}{1-||h||_1} < 1$.*

Nous présentons maintenant un résultat sur l'expression analytique de la structure de covariance des processus de Hawkes, déjà étudiée en version infinitésimale, par exemple dans la proposition 2 de (Bacry, Dayri, & Muzy, 2012).

**Théorème 3.2.2 (Extrait du Théorème 2.4 de Hillairet et Réveillac (2023))**
*Soit $N^{(\mu,h)}$ un processus de Hawkes linéaire, auto-excitant de noyau $h$, associé à la fonction $H(t) = \sum_{k\geq 1} h^{*k}(t)$ telle que définie précédemment. Pour tout couple de réels positifs*

$(s,t)$, *notons* $Cov_{N^{(\mu,h)}}(s,t)$ *la structure de covariance des variations de* $N^{(\mu,h)}$, *égale à* $\mathbb{E}[N^{(\mu,h)}(s)N^{(\mu,h)}(t)] - \mathbb{E}[N^{(\mu,h)}(s)]\mathbb{E}[N^{(\mu,h)}(t)]$. *Si* $\int_{\mathbb{R}^+} h(t)dt < 1$, *alors la structure de covariance des variations s'exprime de la manière suivante :*

$$Cov_{N^{(\mu,h)}}(s,t) = \mu \int_0^s \left(1 + \int_0^v H(w)dw\right)\left(1 + \int_v^s H(w-v)dw\right)\left(1 + \int_v^t H(w-v)dw\right)dv$$

Nous présentons également un résultat introduisant une écriture alternative de l'intensité d'un processus de Hawkes.

**Proposition 3.2.3 (Proposition 2.1 de Jaisson et Rosenbaum (2015))**
*Soit* $N^{(\mu,h)}$ *un processus de Hawkes linéaire et son compensateur* $\Lambda^{(\mu,h)}$. *Soit* $M_{N^{(\mu,h)}}$ *défini par* $M_{N^{(\mu,h)}} = N^{(\mu,h)} - \Lambda^{(\mu,h)}$, *martingale puisque* $h(0) = 0$. *Si* $\int_{\mathbb{R}^+} h < 1$, *alors son intensité* $\lambda_{N^{(\mu,h)}}$ *peut s'exprimer de la manière suivante :*

$$\lambda_{N^{(\mu,h)}}(s) = \mu\left(1 + \int_0^s H(s-u)du\right) + \int_0^s H(s-z)dM_{N^{(\mu,h)}}(z).$$

D'autre part, le comportement de processus de Hawkes à la limite du régime stationnaire, précisément sous des hypothèses de la forme $\int_{\mathbb{R}^+} h_T(t)dt \underset{T\to+\infty}{\to} 1$, est étudié par Jaisson et Rosenbaum (2016).

### 3.2.3 Exemples de noyaux de processus de Hawkes

Une première famille de noyau de processus de Hawkes intéressante est celle des noyaux exponentiels, dont les éléments dépendent d'un couple de paramètres positifs $\theta_0 = (\alpha, \beta)$ et s'écrivent de la manière suivante :

$$h_{(\alpha,\beta)}(0) = 0 \text{ et pour } t > 0,\ h_{(\alpha,\beta)}(t) = \alpha e^{-\beta t}.$$

Un processus de Hawkes de type exponentiel est ainsi stable si et seulement si $\alpha < \beta$. Le noyau exponentiel est souvent utilisé puisque le processus de Hawkes associé est Markovien. Par ailleurs, supposons observer une trajectoire entre $[0;T]$. Nous pouvons calculer de manière analytique les limites à gauche des événements de l'intensité ainsi que la valeur du compensateur en $T$, permettant ensuite de calculer la vraisemblance :

$$\text{pour } h_{(\alpha,\beta)}, \begin{cases} \lim\limits_{t \overset{<}{\to} \tau_k} \lambda^{(\mu,\theta_0)}_{N^{(\mu,\theta_0)}}(t) = \mu + \alpha \sum_{i=1}^{k-1} e^{-\beta(\tau_k - \tau_i)}, \\ \\ \Lambda^{(\mu,\theta_0)}_{N^{(\mu,\theta_0)}}(T) = \mu T + \frac{\alpha}{\beta} \sum_{i=1}^{N^{(\mu,\theta_0)}(T)} \left(1 - e^{-\beta(T-\tau_i)}\right). \end{cases}$$

Un autre type de noyau pouvant être présenté est une généralisation du noyau exponentiel.Défini par trois paramètres positifs $\theta_1 = (\alpha, \beta, \gamma)$, il s'écrit de la manière suivante :

$$h_{(\alpha,\beta,\gamma)}(0) = 0 \text{ et pour } t > 0,\ h_{(\alpha,\beta,\gamma)}(t) = \alpha\gamma t^{\gamma-1} e^{-\beta t^{\gamma}},$$

Un cas particulier intéressant est celui d'un paramètre $\gamma<1$. En effet, le noyau n'est alors pas borné, tout en conservant la stabilité du processus si et seulement si $\alpha < \beta$. De la même manière que ci-dessus, nous pouvons calculer les éléments de la vraisemblance d'une observation entre 0 et $T$ :

$$\text{pour } h_{(\alpha,\beta,\gamma)}, \begin{cases} \lim\limits_{t \overset{<}{\to} \tau_k} \lambda^{(\mu,\theta_1)}_{N^{(\mu,\theta_1)}}(t) = \mu + \alpha\gamma\tau_k^{\gamma-1} \sum_{i=1}^{k-1} e^{-\beta(\tau_k - \tau_i)}, \\ \\ \Lambda^{(\mu,\theta_0)}_{N^{(\mu,\theta_1)}}(T) = \mu T + \frac{\alpha}{\beta} \sum_{i=1}^{N^{(\mu,\theta_1)}(T)} \left(1 - e^{-\beta(T-\tau_i)^{\gamma}}\right). \end{cases}$$

Nous n'utiliserons cependant pas ce type de noyau dans notre contribution.

## 3.3 Accumulation en fréquence des risques numériques

De manière générale, la survenance concomitante d'un certain nombre de sinistres est un défi pour les organismes d'assurance. En effet, l'ampleur de l'accumulation peut remettre en cause le principe de mutualisation entre les expositions assurées et affecter l'efficience de la gestion des sinistres. Par exemple, lors d'événements de grande ampleur comme des catastrophes naturelles, une part significative des expositions géographiquement proches peuvent être impactées dans une fenêtre temporelle limitée. Cette accumulation peut par conséquent créer un pic de demande de gestion de sinistres induisant la prolongation des délais de gestion et l'augmentation des coûts. Ainsi, l'impact d'une accumulation en fréquence s'en trouve-t-il démultiplié.

Dans le cas des catastrophes naturelles, l'origine des accumulations est bien comprise. A titre d'illustration, les sinistres liés à l'impact de la sécheresse sur les bâtis sont générés par des retraits et gonflements des terres argileuses. Autre exemple, les sinistres liés aux inondations se concentrent dans les contres-bas des cours d'eaux surchargés. Certains organismes d'assurance s'appuient donc sur des modèles physiques des aléas naturels et sur la cartographique de leurs expositions pour quantifier les accumulations potentielles. De manière complémentaire, certains transfèrent en partie ce risque au marché de la réassurance, ces acteurs permettant une mutualisation géographique et temporelle. En France, le régime des catastrophes naturelles rend obligatoire cette mutualisation auprès de la Caisse Centrale de Réassurance, détenue par l'Etat et permettant une garantie de crédit étatique. Certains risques sont toutefois exclus comme les tempêtes en France métropolitaine, la grêle et la neige.

En revanche, les risques numériques ne dépendent pas d'aléas naturels mais de perturbations malveillantes de réseaux de transfert et de traitement de l'information numérique. De part son évolution dans un milieu créé par l'homme, la transitivité de ces risques se rapproche certainement davantage de celle des risques de crédit. De plus, le phénomène de concentration des acteurs financiers s'observe également dans le secteur du numérique. Nous pouvons par exemple citer la position incontournable des cinq doigts de la main numérique, les "GAFAM", mais également l'utilisation généralisée de sous-jacents communs comme le langage de programmation C ou des outils numériques répandus. Un impact significatif sur l'une de ces fondations aurait de facto des conséquences systémiques, comme celles observées lors de la faillite de Lehman Brothers. Les différences sont en même temps nombreuses. Les interconnexions numériques sont bien plus denses que les dépendances

financières, le digital étant plus perméable et davantage décentralisé. Par ailleurs, la réglementation du numérique est relativement plus souple que celle du secteur financier, créant une hétérogénéité plus forte dans les capacités de résilience.

Afin de contribuer à la quantification des accumulations potentielles des risques numériques, nous proposons d'analyser de manière statistique les accumulations observées, en ajustant des processus de comptage de type Hawkes. Cette étude pourrait compléter les analyses informatiques de propagation et de concentration des risques numériques.

### 3.3.1 Application à la fréquence des failles de données

Nous commençons par étudier des données de comptage d'événements par l'intermédiaire de variables aléatoires discrètes correspondant à l'observation des processus de comptage à la fin de l'observation $T$. Les données de sinistres d'assurance en lien avec les systèmes d'informations n'étant pas publiques, nous avons dans un premier temps analysé une base de données publique répertoriant des failles de données personnelles aux Etats-Unis, la base de données Privacy Right Clearinghouse (PRC). Cependant, cette base d'événements assimilables à des sinistres n'est pas jointe à une base de clients. Nous avons donc moins d'informations qu'un organisme d'assurance, notamment concernant l'exposition. Par conséquent, nous proposons de reconstituer des informations de portefeuille pour structurer nos données en une base de clients et une base de sinistres. Nous envisageons deux possibilités :

- nous considérons un premier scénario dans lequel les clients correspondent aux entreprises ayant fait face à au moins deux failles répertoriées dans la base PRC. Parmi eux, nous supposons que les clients sinistrés se limitent aux entreprises ayant au moins trois événements recensés.
- nous proposons enfin un second scénario dans lequel les clients correspondent aux organisations côtées à la bourse New York Stock Exchange (NYSE) et ayant fait face à au moins une faille répertoriée dans la base PRC. Parmi eux, nous supposons que les clients sinistrés se limitent auux entreprises ayant au moins fait face à deux failles de données.

Nous faisons ainsi implicitement l'hypothèse que les premières identifications publiques de failles relatives à une entreprise fiabilisent les identifications des futures failles. Pour les deux scénarios, nous obtenons des informations de sinistralité tronquées. Sur ces données de comptage agrégées, nous avons ajusté deux familles de loi de probabilité : la loi de

Poisson et la loi géométrique. Sur la base du critère d'information d'Akaike permettant de comparer ces deux familles aux complexités différentes, nous avons préféré la loi géométrique. En effet, celle-ci traduit dans les deux cas une probabilité d'occurrence annuelle de faille de données de l'ordre de 10%, variant selon le secteur d'activité. La performance de la loi géométrique suggère que le processus sous-jacent n'appartient pas à la famille des processus de Poisson homogènes.

### 3.3.2 Analyse des messages d'un réseau social évoquant des rançongiciels

Plusieurs scénarios peuvent aboutir à une accumulation en fréquences de sinistres numériques, par exemple la déstabilisation d'un jalon central comme un hébergement ou l'utilisation d'une faille d'un programme informatique répandu. La propagation de rançongiciels est également un scénario crédible et qui s'observe régulièrement à des échelles très variables. Une des plus médiatisée est celle du rançongiciel "WannaCry" détectée le 12 mai 2017. Elle impacte dans cette première journée au moins 200 000 ordinateurs dans plus de 150 pays, en exploitant une faille de sécurité pourtant corrigée depuis mars 2017 par Microsoft. Une autre propagation est celle du rançongiciel "Ryuk", détectée le 31 octobre 2021. Elle est sans doute la plus génératrice de revenus pour les hackers avec un montant total de rançons d'environ 50 millions d'euros.

Afin de mieux comprendre les phénomènes d'accumulation en fréquences, nous étudions de manière indépendante sept rançongiciels générant chacun une diffusion médiatique sur la toile du réseau social Twitter. Dans notre analyse, les attaques correspondent à des tweets. Pour chaque rançongiciel, nous observons les temps $(\tau_i)_{i\in\mathbb{N}}$ auxquels un tweet a été publié. Nous considérons chaque trajectoire comme la réalisation d'un processus de comptage, noté $\mathfrak{I}(t)$.

A partir de cette dynamique de propagation sous-jacente $\mathfrak{I}(t)$, nous définissons deux processus ponctuels. Le premier, noté $\mathfrak{A}(t)$, compte le nombre de victimes nécessitant une assistance, en supposant que le besoin d'assistance suite à un sinistre survenu en $\tau$ se concentre sur $[\tau;\tau+\delta[$, avec $\delta > 0$. Le second, noté $\mathfrak{R}(t)$, compte le nombre de victimes ne nécessitant plus d'assistance. Par construction, le nombre de victimes nécessitant une assistance $\mathfrak{A}(t)$ correspond bien au nombre de victimes impactées $\mathfrak{I}(t)$, auquel est retranché le nombre de victimes rétablies $\mathfrak{R}(t)$. Les processus s'expriment ainsi analytiquement.

$$\begin{cases} \mathfrak{I}(t) &= \sum_{\tau_i} \mathbb{1}_{\{\tau_i \leq t\}} \\ \\ \mathfrak{R}(t) &= \sum_{\tau_i} \mathbb{1}_{\{\tau_i \leq t-\delta\}} \\ \\ \mathfrak{A}(t) &= \sum_{\tau_i} \mathbb{1}_{\{t-\delta < \tau_i \leq t\}} \end{cases}$$

Notons que le processus d'intérêt $\mathfrak{A}(t)$ n'est pas un processus de comptage puisqu'il n'est pas nécessairement croissant. En revanche, ses évolutions sont bien unitaires et sa trajectoire est constante par morceaux. Ce type de processus est par exemple étudié pour les risques numériques par Hillairet et Lopez (2021) sur la base de modèles de propagations épidémiologiques. Dans le cadre d'une diffusion par processus de Hawkes, le processus $\mathfrak{A}(t)$ caractérise une dynamique de population semblable à celles étudiées par Boumezoued (2016).

### 3.3.3 Inégalités de concentration relatives aux besoins en assistance

Notre objectif est d'étudier les accumulations maximales du nombre de victimes nécessitant une assistance en même temps afin de dimensionner les ressources d'assistance d'un organisme d'assurance. Nous étudions ainsi le supremum du processus $\mathfrak{A}(t)$ sur un intervalle d'observation $[0;T]$, noté $\sup_{0\leq t\leq T}\mathfrak{A}(t)$. Plus précisément, nous cherchons à déterminer une inégalité de concentration telle que ci-dessous afin de majorer par $\alpha$ la probabilité de dépasser les capacités d'assistance, pour chaque seuil de ressources $x$.

$$\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{A}(t) > x\right) \leq \alpha.$$

Cette quantification des accumulations potentielles des victimes nécessitant une assistance est obtenue en trois étapes :

- d'abord en étudiant la différence entre le processus $\mathfrak{A}(t)$ et son espérance, la proposition 3.2 de la contribution explicitant la décomposition suivante :

$$\mathfrak{A}(t) = \mathbb{E}[\mathfrak{A}(t)] + \mathfrak{M}(t),$$

- puis, en se focalisant sur $\mathfrak{M}(t)$ et plus précisément sur sa norme infinie, nous obtenons, sous certaines conditions, une inégalité sur $\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{M}(t) > x\right)$, à l'étape

C.6 de la preuve. Ce résultat découle d'une disjonction de cas relative au compensateur du processus $\mathfrak{I}(t)$, noté $\Lambda(t)$, comme explicité ci-dessous.

$$\begin{aligned}
\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right) &= \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\cap\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\\
&+ \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\cap\left\{\Lambda(T)\geq\mathfrak{c}_u^T T\right\}\right)\\
&\leq \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\mid\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\mathbb{P}\left(\Lambda(T)<\mathfrak{c}_u^T T\right)\\
&+ \mathbb{P}\left(\Lambda(T)\geq\mathfrak{c}_u^T T\right)\\
&\leq \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\mid\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\\
&+ \mathbb{P}\left(\Lambda(T)\geq\mathfrak{c}_u^T T\right).
\end{aligned}$$

Conditionnellement à l'événement aléatoire d'un compensateur borné en $T$, l'étape C.3 de la preuve explicite la borne obtenue sur le supremum du processus $\mathfrak{M}(t)$, en appliquant un résultat de (Guével, 2021). Par ailleurs, la probabilité d'observer un compensateur dépassant la borne précédemment définie en $T$ est obtenue à l'étape C.4 de la preuve. Cela repose sur un résultat de (Reynaud-Bouret & Roy, 2007), supposant le noyau $h$ du processus de Hawkes borné et nul à partir d'un certain laps de temps $\Delta$.

- enfin, à l'étape C.7 de la preuve, nous analysons le comportement limite de $\mathbb{E}[\mathfrak{A}(t)]$, afin de borner en probabilité les événements de dépassement des capacités d'assistance $\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{A}(t)>x\right)$.

Nous proposons de rappeler ci-dessous ce résultat (corollaire 3.3 de la contribution) ainsi que les deux hypothèses suffisantes à son application.

**Hypothèse 3.3.1** *Nous supposons que le phénomène d'auto-excitation est limité dans le temps, c'est-à-dire que le noyau $h$ du processus de Hawkes est nul à partir d'un certain temps.*

$$\exists\Delta<+\infty \text{ tel que pour tout } t>\Delta,\ h(t)=0.$$

**Hypothèse 3.3.2** *Nous supposons par ailleurs que le noyau $h$ du processus de Hawkes est borné.*

$$||h||_\infty := \sup_{s\in[0;\Delta]} h(s) < +\infty.$$

**Théorème 3.3.3**

*Soit $N^{(\mu,h)}$ un processus de Hawkes linéaire auto-excitant, $\Lambda_{N^{(\mu,h)}}$ son compensateur et $u$ un réel strictement positif. Si $||h||_1 < 1$ et si les hypothèses 3.3.1 et 3.3.2 sont vérifiées, alors il existe un $t_u > 0$, dépendant de $(u, \Delta, ||h||_1)$, tel que pour tout $T \geq t_u$, le supremum du processus $\mathfrak{A}(t)$ sur $[0;T]$ est borné en probabilité, de la manière suivante :*

$$\mathbb{P}\left(\sup_{0\leq t\leq T} \mathfrak{A}(t) > x\right) \leq \exp\left(-\mathfrak{c}_u^T T \mathfrak{H} I\left(\frac{x - \mu\delta\left(1 + ||H||_1\right)}{\mathfrak{c}_u^T T \mathfrak{H}}\right)\right) + \left(3 + \frac{\mu e}{u + \log(T)}\right) e^{-u},$$

- *avec* $||H||_1 = \frac{||h||_1}{1-||h||_1}$,
- $\mathfrak{H} = 1 + ||H||_1$ *de manière générale,*
- *et* $\mathfrak{H} = 1 + \int_0^\delta H(u)du$ *lorsque* $H$ *est décroissante.*

**Remarque 3.3.4** *Dans la borne obtenue, $u \in \mathbb{R}^{+*}$ est quelconque tant que $T > t_u$. Cette contrainte n'admet en revanche pas de formule fermée puisque la dépendance de $t_u$ en fonction de $u$ n'est pas connue. En supposant que $T > t_u$ pour $u \in [u_w; u_m]$, la borne peut être appliquée avec un $u^*$ optimal qui la minimise sur $[u_w; u_m]$. Des algorithmes d'optimisation permettraient par exemple d'estimer ce compromis entre les deux parties du majorant.*

La preuve est détaillée à l'article D de la contribution, dans laquelle nous estimons le processus de Hawkes sur des tweets liés à des ransomwares puis appliquons ce résultat pour illustrer les potentielles accumulations en fréquences. Notre application illustre en particulier la force auto-excitante du rançongiciel Wannacry sur les données analysées. Pour autant, une étude complémentaire de la compréhension de $t_u$ permettrait de rationaliser l'application de ce résultat.

# Conclusion

Au travers de ce manuscrit, nous avons présenté des solutions mathématiques personnalisées dans l'objectif de contribuer à la résolution de certains enjeux théoriques soulevés par l'assurance des risques numériques. Notons que les résultats proposés sont certainement perfectibles et généralisables. L'application concrète de ces méthodes sur une base de données de sinistralité numérique nécessiterait par ailleurs une attention particulière, afin d'ajuster les paramètres annexes aux données. Par exemple la constante de pénalisation pour l'arbre de régression construit selon la vraisemblance d'une loi de Pareto généralisée, ou encore le noyau pour l'estimation d'une copule dans un contexte de censure à droite ou pour l'estimation d'un processus de Hawkes et la majoration des scénarios d'accumulation. En conclusion de ce manuscrit, nous proposons pour chacun des trois chapitres des suggestions d'applications et des perspectives théoriques.

Au premier chapitre, nous avons adapté l'algorithme d'arbre de régression au contexte de la classification de risques à partir de valeurs extrêmes. Cette approche permet en particulier d'identifier une structure claire des profils de risque en distinguant les poids des différentes queues de distribution. En pratique, cette méthode pourrait s'appliquer à l'analyse de sinistres graves et contribuer à la rédaction d'un guide de souscription pour sélectionner des risques sur la base de l'expérience acquise. Pour les risques assurés au sein d'un portefeuille, cette méthode pourrait aider aux prises de décisions d'évolution tarifaire en intégrant le poids de la sinistralité grave de manière différenciée. En ce qui concerne les perspectives, les généralisations de l'arbre de régression semblent bienvenues, telles que la forêt aléatoire, le gradient boosting et l'initialisation d'un réseau de neurones (cf. Biau, Scornet, & Welbl, 2019). Leurs analyses théoriques permettraient certainement une mise en perspective appréciable des résultats obtenus.

Au second chapitre, nous avons ajusté les estimations semi-paramétriques et non-paramétriques de copules dans un contexte de données censurées à droite. Cette méthode permet notamment d'estimer la dépendance entre les coûts de diverses garanties d'un sinistre numérique. Elle s'applique ainsi au processus opérationnel de provisionnement,

consistant à valoriser les engagements relatifs aux sinistres non encore clôturés. Pour autant, nous supposons les coûts des sinistres ouverts censurés à zéro, alors qu'en pratique, plusieurs règlements sont réalisés dans le cadre de la gestion des sinistres ouverts, par exemple dans le cadre d'une indemnisation partielle ou d'une expertise. L'application de cette méthodologie à une base de données réelle et sa mise en concurrence avec des méthodologies plus classiques permettraient de préciser les situations d'applications particulièrement pertinentes.

Au troisième chapitre, nous avons proposé une majoration de l'accumulation d'événements issus d'un processus de Hawkes. Ce résultat permet en particulier de contribuer aux prises de décisions relatives à l'appétit au risque des organismes d'assurance, mais également à la mise en place d'une veille sur le portefeuille afin de détecter les prémices d'accumulations extrêmes. Toutefois, dans le cadre de la mise en pratique des résultats, une étude approfondie des constantes impliquées dans la majoration, théorique ou empirique, serait appréciable afin de valider l'application sur une base de données réelles. L'utilisation des processus ponctuels pour enrichir la compréhension de la fréquence de sinistralité semble une piste pertinente, d'autant plus dans le contexte de l'assurance des risques numériques (cf. Hillairet, Réveillac, & Rosenbaum, 2023).

Nous avons ainsi exploré une partie des outils de mathématiques permettant d'apporter des éléments de réflexions à la quantification des risques numériques pour l'assurance. D'autres approches existent et pourraient être étudiées et appliquées aux risques numériques. Par exemple, la criminalité numérique ayant pour principales motivations la rentabilité, l'espionnage ou la déstabilisation, les interactions entre assurés et individus malveillants pourraient être modélisées en les considérant comme des agents économiques au comportement rationnel. Cette approche suggérée par (Didier Parsoire, 2019) repose sur l'application de concepts de la théorie des jeux, qui semble particulièrement pertinente pour ces risques générés par les humains. Les contributions mathématiques ne suffisent pas à rendre la gestion des risques numériques pérenne. De manière générale, les pouvoirs publics, les organismes d'assurances et les acteurs de la sécurité numérique sont complémentaires et essentiels pour aider les entreprises à mettre en œuvre le virage de la gestion des risques numériques qui est globalement nécessaire pour atteindre des capacités de résilience satisfaisantes (cf. Hassler, 2019).

# Bibliographie


Akritas, M. G. (2000). The central limit theorem under censoring. *Bernoulli*, *6*(6), 1109–1120. Consulté le 2023-04-04, sur http://www.jstor.org/stable/3318473

Autorité de régulation des communications électroniques, d. p. e. d. l. d. d. l. p. A. (2022). (http://www.data.gouv.fr/fr/datasets/barometre-du-numerique/)

Bacry, E., Dayri, K., & Muzy, J.-F. (2012). Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, *85*, 1–12.

Bacry, E., Delattre, S., Hoffmann, M., & Muzy, J.-F. (2013). Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, *123*(7), 2475–2499.

Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, *1*(01), 1550005.

Bacry, E., & Muzy, J.-F. (2016). First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, *62*(4), 2184-2202. doi: 10.1109/TIT.2016.2533397

Balkema, A. A., & de Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, *2*(5), 792 – 804. Consulté sur https://doi.org/10.1214/aop/1176996548 doi: 10.1214/aop/1176996548

Beirlant, J., & Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Computational statistics & data analysis*, *42*(4), 595–619.

Beirlant, J., & Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, *89*(1), 97–118.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2004). *Statistics of extremes : theory and applications*. John Wiley & Sons.

Biau, G., Scornet, E., & Welbl, J. (2019, December). Neural Random Forests. *Sankhya A : The Indian Journal of Statistics*, *81*(2), 347-386. Consulté sur https://ideas.repec.org/a/spr/sankha/v81y2019i2d10.1007_s13171-018-0133-y.html doi:

10.1007/s13171-018-0133-y

Biener, C., Eling, M., & Wirfs, J. H. (2015). Insurability of cyber risk : An empirical analysis. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *40*, 131–158.

Böhme, R., & Kataria, G. (2006). Models and measures for correlation in cyber-insurance. In *Weis* (Vol. 2, p. 3).

Bonnet, A., Dion, C., Gindraud, F., & Lemler, S. (2021, juillet). *Neuronal Network Inference and Membrane Potential Model using Multivariate Hawkes Processes.* Consulté sur `https://hal.archives-ouvertes.fr/hal-03309709` (working paper or preprint)

Boumezoued, A. (2016). Population viewpoint on hawkes processes. *Advances in Applied Probability*, *48*(2), 463–480. doi: 10.1017/apr.2016.10

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bremaud, P. (1981). *Point processes and queues : Martingales dynamics*. Springer Verlag, New York. doi: https://doi.org/10.1007/978-1-4684-9477-8

Chavez-Demoulin, V., Embrechts, P., & Hofert, M. (2015). An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates. *Journal of Risk and Insurance*, *83*(3), 735–776.

Chavez-Demoulin, V., & McGill, J. (2012). High-frequency financial data modeling using hawkes processes. *Journal of Banking Finance*, *36*(12), 3415-3426. Consulté sur `https://www.sciencedirect.com/science/article/pii/S0378426612002336` (Systemic risk, Basel III, global financial stability and regulation) doi: https://doi.org/10.1016/j.jbankfin.2012.08.011

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151. Consulté le 2023-07-10, sur `http://www.jstor.org/stable/2335289`

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer London. Consulté sur `https://doi.org/10.1007/978-1-4471-3675-0` doi: 10.1007/978-1-4471-3675-0

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*(2), 129–164. Consulté le 2023-07-13, sur `http://www.jstor.org/stable/2983950`

Cox, D. R., & Lewis, P. A. (1966). The statistical analysis of series of events.

Daniel Zajdenweber, P. V. J. S. J. C. V. M. E. Q. M.-C. D. C. D. N. A. J. P. P. M., Philippe Cotelle. (2020). Cyber(in)sécurité. *Les cahier de l'assurance*, *123*, 19–61.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Bulletins de l'Académie Royale de Belgique*, *65*(1), 274–292. Consulté sur `https://www.persee.fr/doc/barb_0001-4141_1979_num_65_1_58521` doi: 10.3406/barb.1979.58521

des juristes, L. C. (2018). *Assurer le risque cyber*.

des juristes, L. C. (2021). *Le droit pénal à l'épreuve des cyberattaques*.

Didier, D. (2019). Qu'est-ce que le « numérique » ? regards sur le champ lexical qui l'accompagne ». *Comprendre et maîtriser les excès de la société numérique*, 17-22. Consulté sur `https://www.cairn.info/comprendre-et-maitriser-les-exces-de-la-societe--9782810906994-page-17.htm`

Didier Parsoire, S. H. (2017). La couverture du cyber-risque. *Revue d'économie financière*, *126*, 169–182. doi: 10.3917/ecofi.126.0169

Didier Parsoire, S. H. (2019). Le prix du risque cyber. *Revue d'économie financière*, *133*, 155–170. doi: 10.3917/ecofi.133.0155

D. J. Daley, D. V.-J. (2008a). *An introduction to the theory of point processes : Volume i*. Springer. Consulté sur `https://doi.org/10.1007/978-0-387-49835-5` doi: 10.1007/978-0-387-49835-5

D. J. Daley, D. V.-J. (2008b). *An introduction to the theory of point processes : Volume ii*. Springer. Consulté sur `https://doi.org/10.1007/978-0-387-49835-5` doi: 10.1007/978-0-387-49835-5

Doob, J. L. (1953). *Stochastic processes*. John Wiley  Sons, New York.

Einmahl, U., Mason, D. M., et al. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, *33*(3), 1380–1403. doi: https://doi.org/10.1214/009053605000000129

Eling, M. (2018). *Cyber risk and cyber risk insurance : Status quo and future research* (Vol. 43). Springer.

Eling, M., & Schnell, W. (2016). What do we know about cyber risk and cyber risk insurance ? *The Journal of Risk Finance*, *17*(5), 474–491.

Embrechts, P., Lindskog, F., & Mcneil, E. (2001). Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, *8*. doi: 10.1016/B978-044450896-6.50010-8

Embrechts, P., McNeil, A. J., & Straumann, D. (2002). Correlation and dependence in risk management : Properties and pitfalls. In M. A. H. Dempster (Ed.), *Risk management : Value at risk and beyond* (p. 176–223). Cambridge University Press. doi: 10.1017/CBO9780511615337.008

Embrechts, P., Rüdiger, F., & Mcneil, A. J. (2015). *Quantitative risk management : Concepts, techniques and tools revised edition*. Economics Books, Princeton University Press.

Errais, E., Giesecke, K., & Goldberg, L. R. (2010). Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, *1*(1), 642-665. Consulté sur `https://doi.org/10.1137/090771272` doi: 10.1137/090771272

Eurobarometer. (2022). (`http://data.europa.eu/88u/dataset/S2280_FL496_ENG`)

Faure-Muntian, V. (2021). Rapport sur la cyber-assurance.

Fermanian, J.-D., Radulovic, D., & Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, *10*(5), 847 – 860. Consulté sur `https://doi.org/10.3150/bj/1099579158` doi: 10.3150/bj/1099579158

Fermanian, J.-D., & Wegkamp, M. H. (2012). Time-dependent copulas. *Journal of Multivariate Analysis*, *110*, 19-29. Consulté sur `https://www.sciencedirect.com/science/article/pii/S0047259X12000590` (Special Issue on Copula Modeling and Dependence) doi: https://doi.org/10.1016/j.jmva.2012.02.018

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis* (Vol. 169). John Wiley & Sons.

Frank, M. J. (1979). On the simultaneous associativity of f(x, y) and x+y - f(x, y). *Aequationes Math*, *19*(19), 194-226. Consulté sur `https://doi.org/10.1007/BF02189866`

Fréchet, M. (1960). Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *28*(1/2), 10–32. Consulté le 2023-07-01, sur `http://www.jstor.org/stable/1401846`

Geerdens, C., Janssen, P., & Veraverbeke, N. (2016). Large sample properties of non-parametric copula estimators under bivariate censoring. *Statistics*, *50*(5), 1036-1055. Consulté sur `https://doi.org/10.1080/02331888.2015.1119149` doi: 10.1080/02331888.2015.1119149

Genest, C., & MacKay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics /*

*La Revue Canadienne de Statistique*, *14*(2), 145–159. Consulté le 2023-07-10, sur http://www.jstor.org/stable/3314660

Genest, C., & Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association*, *88*(423), 1034–1043.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *The annals of statistics*, *11*(1), 49–58.

Gilles Bénéplanc, J. N. F.-X. A. E. F. L. Z. A. T.-M. N., Philippe Lemoine. (2018). Se protéger face aux cyberattaques. *Les cahier de l'assurance*, *113*, 15–57.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, *44*(3), 423–453. Consulté le 2023-07-23, sur http://www.jstor.org/stable/1968974

Gribkova, S., & Lopez, O. (2015). Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics*, *42*(4), 925–946. Consulté le 2023-08-06, sur http://www.jstor.org/stable/24586868

Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized pareto distribution. Consulté sur https://doi.org/10.1080/00401706.1993.10485040 doi: 10.1080/00401706.1993.10485040

Guével, R. L. (2021). Exponential inequalities for the supremum of some counting processes and their square martingales. *Comptes Rendus. Mathématique*, *359*(8), 969–982. doi: 10.5802/crmath.206

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, *55*(292), 698-707. doi: 10.1080/01621459.1960.10483368

Haan, L. D., & Ferreira, A. (2006). *Extreme value theory*. Springer New York. Consulté sur https://doi.org/10.1007%2F0-387-34471-3 doi: 10.1007/0-387-34471-3

Hassler, A. (2019). Assurer, réassurer et titriser les cyber-risques. *Les cahier de l'assurance*, *120*, 135–141.

Hawkes, A. G. (1971a, 07). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B*, *33*. doi: 10.1111/j.2517-6161.1971.tb01530.x

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*, 83-90.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance : a review. *Quantitative Finance*, *18*(2), 193-198. Consulté sur https://doi.org/10.1080/14697688.2017.1403131 doi: 10.1080/14697688.2017.1403131

HCJPFP. (2022, 01). Rapport sur l'assurabilité des risques cyber.

Heyde, C. C. (Ed.). (1997). *Quasi-likelihood and its application*. Springer New York. Consulté sur https://doi.org/10.1007%2Fb98823 doi: 10.1007/b98823

Hillairet, C., & Lopez, O. (2021). Propagation of cyber incidents in an insurance portfolio : counting processes combined with compartmental epidemiological models. *Scandinavian Actuarial Journal*, *8*(2), 671-694. doi: 10.1080/03461238.2021.1872694

Hillairet, C., & Réveillac, A. (2023). Explicit correlations for the hawkes processes. *arXiv preprint arXiv :2304.02376*.

Hillairet, C., Réveillac, A., & Rosenbaum, M. (2023). An expansion formula for hawkes processes and application to cyber-insurance derivatives. *Stochastic Processes and their Applications*, *160*, 89-119. Consulté sur https://www.sciencedirect.com/science/article/pii/S0304414923000455 doi: https://doi.org/10.1016/j.spa.2023.02.012

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73 – 101. Consulté sur https://doi.org/10.1214/aoms/1177703732 doi: 10.1214/aoms/1177703732

Insua, D. R., Couce-Vieira, A., & Musaraj, K. (2018). Some risk analysis problems in cyber insurance economics. *Studies of Applied Economics*, *36*(1), 181–194.

Jacobs, J. (2014). Analyzing ponemon cost of data breach. *Data Driven Security*, *11*.

Jaisson, T., & Rosenbaum, M. (2015). Limit theorems for nearly unstable hawkes processes. *The Annals of Applied Probability*, *25*(2), 600–631. Consulté sur http://www.jstor.org/stable/24519929

Jaisson, T., & Rosenbaum, M. (2016). Rough fractional diffusions as scaling limits of nearly unstable heavy tailed hawkes processes. *The Annals of Applied Probability*, *26*(5), 2860–2882. Consulté sur http://www.jstor.org/stable/24810116

JO. (2023, 01). LOI n° 2023-22 du 24 janvier 2023 d'orientation et de programmation du ministère de l'intérieur. *Journal officiel de la République française*(0021). Consulté sur https://www.legifrance.gouv.fr/eli/loi/2023/1/24/IOMD2223411L/jo/texte

Kanno-Youngs, Z., & Sanger, D. E. (2021). U.S. Accuses China of Hacking Microsoft. *The New York Times*.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits.

*Journal of the American Statistical Association*, *96*(454), 589-604. doi: 10.1198/016214501753168271

Lewis, P. A. W., & Shedler, G. S. (1976). Simulation of nonhomogeneous poisson processes with log linear rate function. *Biometrika*, *63*(3), 501–505. Consulté le 2023-07-18, sur http://www.jstor.org/stable/2335727

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14-23. Consulté sur https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.8 doi: https://doi.org/10.1002/widm.8

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review / Revue Internationale de Statistique*, *82*(3), 329–348. Consulté sur http://www.jstor.org/stable/43298996

Lopez, O. (2007). *Réduction de dimension en présence de données censurées* (Theses, ENSAE ParisTech). Consulté sur https://pastel.archives-ouvertes.fr/tel-00195261

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin : The Journal of the IAA*, *23*(2), 213–225. doi: 10.2143/AST.23.2.2005092

McNeil, A. J., & Nešlehová, J. (2009). Multivariate archimedean copulas, d-monotone functions and -norm symmetric distributions. *The Annals of Statistics*, *37*(5B), 3059–3097. Consulté sur http://www.jstor.org/stable/30243736

Meurant, S., & Cardon, R. (2021, 10). Rapport d'information fait au nom de la délégation aux entreprises relatif à la cybersécurité des entreprises. (678).

Meyer, P.-A. (1971). Démonstration simplifiée d'un théorème de Knight. *Séminaire de probabilités de Strasbourg*, *5*, 191–195. Consulté sur http://www.numdam.org/item/SPS_1971__5__191_0/

Møller, J., & Rasmussen, J. G. (2005). Perfect simulation of hawkes processes. *Advances in applied probability*, *37*(3), 629–646.

NAIC. (2019). Report on the cybersecurity insurance and identity theft coverage supplement.

nationale de la sécurité des systèmes d'information, A. (2018). (La méthode EBIOS Risk Manager)

nationale de la sécurité des systèmes d'information, A. (2020). Etat de la menace rançongiciel à l'encontre des entreprises et institutions.

nationale de la sécurité des systèmes d'information, A. (2022). Panorama de la cyberme-

nance 2022.

Nelsen. (2007). *An introduction to copulas*. Springer New York. Consulté sur https://books.google.pt/books?id=B3ONT5rBv0wC

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in Statistics - Theory and Methods*, *15*(11), 3277-3285. doi: 10.1080/03610928608829309

Nicholas J Tierney, M. J. H., Fiona A Harden, & Mengersen, K. L. (2015). Using decision trees to understand structure in missing data. *BMJ open*, *5*(6). doi: doi:10.1136/bmjopen-2014-007450

of Insurance Supervisors, I. A. (2023, 04). Global Insurance Market Report (GIMAR) SPECIAL TOPIC EDITION Cyber.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, *27*(1), 23-31. doi: 10.1109/TIT.1981.1056305

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, *83*(401), 9–27. Consulté le 2023-07-18, sur http://www.jstor.org/stable/2288914

Omer Yoachimik, A. F., Julien Desgats. (2021). Cloudflare atténue une attaque DDoS record de 71 millions de requêtes par seconde. Consulté sur https://blog.cloudflare.com/fr-fr/cloudflare-mitigates-record-breaking-71-million-request-per-second-ddos-attack-fr-fr/

Papangelou, F. (1974). On the palm probabilities of processes of points and processes of lines. *In Stochastic geometry*, 114–147.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, *3*(1), 119–131.

Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Bachelier (Paris). Consulté sur http://catalogue.bnf.fr/ark:/12148/cb31129269s doi: ark:/12148/bpt6k110193z

Reynaud, F. (2016). Le piratage de Yahoo ! est le plus important vol de données de l'histoire. *Le Monde*.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., & Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, *4*, 1–41.

Reynaud-Bouret, P., & Roy, E. (2007). Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, *13*(5), 883 –

896. Consulté sur https://doi.org/10.36045/bbms/1170347811 doi: 10.36045/bbms/1170347811

Romanosky, S., Ablon, L., Kuehn, A., & Jones, T. (2019, 02). Content analysis of cyber insurance policies : how do carriers price cyber risk ? *Journal of Cybersecurity*, *5*(1). Consulté sur https://doi.org/10.1093/cybsec/tyz002 (tyz002) doi: 10.1093/cybsec/tyz002

Rubin, I. (1972). Regular point processesand their detection. *IEEE Transactions on information theory*, *18*(5), 547–557. Consulté sur https://ieeexplore.ieee.org/abstract/document/1054897

Sanger, D. E. (2012). Obama Order Sped Up Wave of Cyberattacks Against Iran. *The New York Times*.

Satten, G. A., & Datta, S. (2001). The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, *55*(3), 207–210. Consulté le 2023-03-30, sur http://www.jstor.org/stable/2685801

Schweizer, B., & Sklar, A. (1983). *Probabilistic metric spaces*. North Holland. Consulté sur https://books.google.fr/books?id=b0_vAAAAMAAJ

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, *18*(3), 764 – 782. Consulté sur https://doi.org/10.3150/11-BEJ387 doi: 10.3150/11-BEJ387

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, *51*(4), 1384–1399. Consulté le 2023-08-06, sur http://www.jstor.org/stable/2533269

Sklar, M. (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.

Stute, W. (1995). The Central Limit Theorem Under Random Censorship. *The Annals of Statistics*, *23*(2), 422 – 439. Consulté sur https://doi.org/10.1214/aos/1176324528 doi: 10.1214/aos/1176324528

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 1089–1102.

Stute, W., & Wang, J.-L. (1993). The Strong Law under Random Censorship. *The Annals of Statistics*, *21*(3), 1591 – 1607. Consulté sur https://doi.org/10.1214/aos/1176349273 doi: 10.1214/aos/1176349273

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586-598. Consulté sur https://doi.org/10.1198/106186004X2165 doi: 10.1198/106186004X2165

Suru, A. (2020). *Assurance iard, les dessous d'un secteur qui vous protège*. Economica.

Sébastien Héon, F. D. (2013). L'analyse du risque cyber, emblématique d'un dialogue nécessaire. *Sécurité et stratégie*, *14*, 44–52. doi: 10.3917/sestr.014.0044

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 28–76.

Terry M. Therneau, E. J. A., & Mayo. (2022). An introduction to recursive partitioning using the rpart routines.

Therneau, T., & Mayo. (2022). User written splitting functions for rpart.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, *33*(3), 357–375. Consulté sur `http://dx.doi.org/10.1002/cjs.5540330304` doi: 10.1002/cjs.5540330304

Union, I. T. (2022). (Global Connectivity Report)

Van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Watanabe, S. (1964). On discontinuous additive functionals and lévy measures of a markov process. *Japanese journal of mathematics :transactions and abstracts*, *34*, 53-70.

Wrede, D., Stegen, T., & Graf von der Schulenburg, J.-M. (2020). Affirmative and silent cyber coverage in traditional insurance policies : Qualitative content analysis of selected insurance products from the german insurance market. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *45*, 657–689.

Yanagimoto, T., & Okamoto, M. (1969). Partial orderings of permutations and monotonicity of a rank correlation statistic. *Annals of the Institute of Statistical Mathematics*, *21*(21), 489-506. Consulté sur `https://doi.org/10.1007/BF02532273`

# Article A

# Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance

# Cyber claim analysis using Generalized Pareto Regression Trees with applications to insurance

Sébastien Farkas[1], Olivier Lopez[1], Maud Thomas[1]

**Abstract**

With the rise of the cyber insurance market, there is a need for better quantification of the economic impact of this risk and its rapid evolution. Due to the heterogeneity of cyber claims, evaluating the appropriate premium and/or the required amount of reserves is a difficult task. In this paper, we propose a method for cyber claim analysis based on regression trees to identify criteria for claim classification and evaluation. We particularly focus on severe/extreme claims, by combining a Generalized Pareto modeling—legitimate from Extreme Value Theory—and a regression tree approach. Coupled with an evaluation of the frequency, our procedure allows computations of central scenarios and of extreme loss quantiles for a cyber portfolio. Finally, the method is illustrated on a public database.

**Key words:** Cyber insurance; Extreme value analysis; Regression Trees; Generalized Pareto Distribution.

**Short title:** Cyber risk using GP regression trees

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France, E-mails: sebastien.farkas@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr, maud.thomas@sorbonne-universite.fr

# 1 Introduction

Cyber risk is a natural consequence of the digital transformation. Digital technologies induce new vulnerabilities for economic actors, with a rapid evolution of practices, threats, and behaviors. With the increase of cyber threats, insurance contracts appear as fundamental tools to improve the resilience of society.

However while the cyber insurance market is growing fast (see for example the report of European Insurance and Occupational Pensions Authority (EIOPA), 2019), risk analysis faces a lack of consistent and reliable data in a context where the amount of claims is particularly volatile (see Matthews, 2019). Therefore, quantifying this emerging and evolving risk is a difficult task. In this paper, we propose to analyze cyber claims via regression trees in order to constitute clusters of cyber incidents. These clusters achieve a compromise between homogeneity and a sufficient size to allow a reliable statistical estimation of the risk. A particular attention is devoted to large claims, for which heavy tail distributions are fitted. The study of large claims raises the question of insurability of the risk, and the clustering technique we propose may help to separate types of incidents or circumstances according to whether they can be covered without endangering risk pooling. In the present work, we develop a regression tree methodology specifically adapted to the study of heavy-tailed distributions, and discuss its relevance to embrace the challenges of cyber risk quantification. A first contribution is methodological: by adapting the methodology of regression trees to extreme value regression purpose, we provide a flexible and still intelligible modeling tool, that can be valuable for a wide range of risks (not only in the field of cyber). On the other hand, we aim to discuss their practical behavior on a real cyber related database. In order to allow reproducibility of our work, we consider a publicly available database. Through this analysis, we show how the output of our method can be used for cyber insurance risk management, and to identify stylized facts useful for practitioners. Furthermore, public databases on cyber events are an important source of information to complement the information (usually poor due to the relative novelty of the risk) available for insurance companies. We hope that the detailed description of the path of our analysis can help to improve the integration of such public information in cyber risk quantification.

Topics recently addressed in cyber insurance are reviewed in Biener et al. (2015), Eling and Schnell (2016) and Marotta et al. (2017). However most of these approaches are performed from the point of view of a cyber analyst. For instance, Fahrenwaldt

et al. (2018) study the topology of infected networks, and Insua et al. (2019) gather expert judgments using an Adversarial Risk Analysis. Eling and Loperfido (2017) and Edwards et al. (2016) developed more established insurance modeling methods illustrated on the Privacy Rights Clearinghouse (PRC) database (available for public download at `https://privacyrights.org/data-breaches`). PRC database has also been studied by Maillart and Sornette (2010). It gathers data breaches events for which a severity indication is given (through the volume of breached data), making it valuable for insurance applications. On the other hand, this database is not fed by an insurance portfolio but by various sources of information, each reporting heterogeneous types of claims. In particular, the exposure (that is the number of entities exposed to the risk in the scope of PRC organization) is blur.

In the present paper, we consider the same PRC database to illustrate our methodology, that can be easily extended to other types of data. The method we develop is adapted to detect such instabilities in this context of a database fed by sources of information which variety may disturb the evaluation of the risk. We especially focus on "extreme" events, that is events for which the severity of the claim is larger than a fixed (high) threshold, seeking to gain further insight on the impact of the characteristics of companies and of the circumstances on a cyber event. Therefore, relying on regression trees inference and extreme value theory, we introduce a statistical methodology that takes into account both the heterogeneity and the extreme features. In addition, we propose an insurance pricing and reserving framework based on assumptions on the exposure and on the costs of data breaches in order to take advantage of the PRC database within the realms of possibility.

Regression trees are good candidates to understand the origin of the heterogeneity, since they allow to perform regression and classification simultaneously. Since the pioneer works of Breiman et al. (1984) who introduced CART algorithm (Clustering And Regression Tree), regression trees have been used in many fields, including industry (see e.g. González et al., 2015), geology (see e.g. Rodriguez-Galiano et al., 2015), ecology (see e.g. De'ath and Fabricius, 2000), claim reserving (see e.g. Lopez et al., 2016). A nice feature of this approach is to introduce nonlinearities in the way the distribution is modeled, while furnishing an intelligible interpretation of the final classification of response variables. A further argument inf the favor of the use of regression trees is the simplicity of the algorithm: such models are fitted to the data via an iterative decomposition. The splitting criterion depends on the type of problems one wishes to investigate: the standard CART algorithm uses a quadratic loss since it aims at performing mean-regression. Alternative

loss functions may be considered as in (Chaudhuri and Loh, 2002) in order to perform quantile regression or in (Su et al., 2004) for log-likelihood loss for example. Loh (2011, 2014) provide detailed descriptions of regression trees procedures and a review of their variants. In the present paper, we use different types of splitting criteria, with a particular attention devoted to the tail of the distribution of the claim size, which describes the behavior of extreme events. We therefore use a Generalized Pareto distribution to approximate the tail of the distribution—which is at the core of the "Peaks Over Threshold" procedure in extreme value theory (see e.g. Beirlant et al., 2004; Pickands, 1975)—with parameters depending on the classes defined by the regression tree.

The rest of the paper is organized as follows. In Section 2, we give a short presentation of the PRC database, its advantages and its inconsistencies. The general description of regression trees and their adaptation to extreme value analysis is done in Section 3. These methodologies are applied to the PRC database in Section 4, leading to a model for the severity of claims. This model is combined with a frequency model in Section 5.2, in order to quantify the impact of this analysis on (virtual) insurance portfolios.

# 2 A public data breaches database

The Privacy Rights Clearinghouse (PRC) database is one of the few publicly available databases on cyber events which associates a quantification of the severity with a claim. This piece of information is crucial from an insurance perspective: evaluation of the risk associated with a policyholder requires to estimate the probability of being a victim of a cyber event (or the frequency of occurrence of such events), and to quantify the potential random loss. Regarding the severity, PRC database does not directly provide the loss associated with an event, but reports the number of records (that is the number of user accounts) affected by the breach. This number is correlated to the financial impact of the claim, which can be approximatively retrieved through a formula given in Jacobs (2014) which will be described later on in Section 5.1. We describe the database in Section 2.1. A focus on the sources feeding the database is done in Section 2.2. This short overview helps us to identify some characteristics and inconsistencies of cyber data summarized in Section 2.3, and will motivate the use of the methodology developed in the rest of the paper.

## 2.1 Description of the database

Privacy Rights Clearinghouse is a nonprofit organization founded in 1992 which aims at protecting US citizens privacy. Especially, PRC has maintained a chronology since 2005, listing companies that have been involved in data breaches affecting US citizens. This article is based on a download of this database made on January 23 2019, corresponding to 8860 cyber events on companies, mainly American companies. Among them, only 8298 events were kept for our analysis, since we eliminated duplicated and/or inconsistent events (e.g. information on the targeted company are sometimes not consistent).

The PRC database gathers information regarding each cyber event (its type, the number of records affected by the breach, a description of the event) and its victim (the targeted company name, its activities, its localization). These variables and their modalities are summarized in Tables 1 to 3. Additional statistics are shown in the supplementary material (Section 1).

Table 1: List of the available variables in the PRC database.

| PRC database | Variable |
|---|---|
| Victim data | Name of organization |
| | Sector of organization |
| | Geographic position of organization |
| Event data | Source of release |
| | Date of release |
| | Type of breach |
| | Number of affected records |
| | Description of the event |

Table 2: Labels for activity sectors of victims in the PRC database.

| | |
|---|---|
| BSF | Businesses - Financial and Insurance Services |
| BSO | Businesses - Other |
| BSR | Businesses - Retail/Merchant - Including Online Retail |
| EDU | Educational Institutions |
| GOV | Government & Military |
| MED | Healthcare, Medical Providers & Medical Insurance Services |
| NGO | Nonprofits |

Table 3: List of the types of data breaches as labelled in the PRC database.

| | |
|---|---|
| CARD | Fraud involving debit and credit cards that is not accomplished via hacking |
| HACK | Hacked by outside party or infected by malware |
| INSD | Insider (someone with legitimate access intentionally breaches information) |
| PHYS | Includes paper documents that are lost, discarded or stolen (non electronic) |
| PORT | Lost, discarded or stolen laptop, PDA, smartphone, memory stick, CDs, hard drive, data tape, etc. |
| STAT | Stationary computer loss (lost, inappropriately accessed, discarded or stolen computer or server not designed for mobility) |
| DISC | Unintended disclosure (not involving hacking, intentional breach or physical loss) |
| UNKN | Unknown |

## 2.2 Multiple sources feeding the database

In this section, we focus on the variable "Source of release". The PRC organization gathers cyber events from different sources, which can be clustered into four groups:

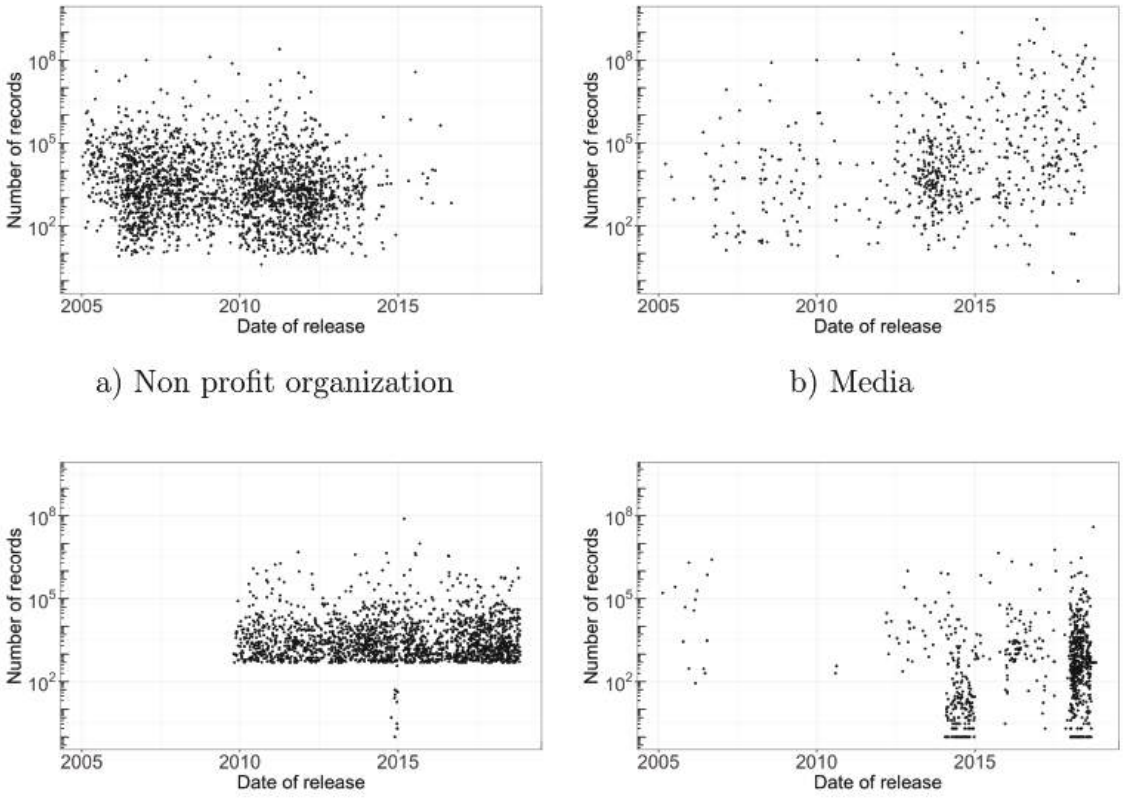
- US Government Agencies on the federal level: in the healthcare domain, the Health Insurance Portability and Accountability Act (HIPAA) imposes a notification to the Secretary of the U.S. Department of Health and Human Services for each breach that affects 500 or more individuals, (U.S. HHS department, n.d.-b). Those notifications are reported online with free access on the breach portal (U.S. HHS department, n.d.-a).

- US Government Agencies on the state level: since 2018, every state has a specific legislation related to data breaches. Differences have been studied by Privacy Rights Clearinghouse (2019). Particularly, there is no uniformity on the threshold (in terms of number of victims) above which a notification becomes mandatory. Some states publicly release notifications, which is the case of California through the online portal (State of California, n.d.), but this is not systematic.

- Media: PRC organization monitors media to list data breaches that have received extensive media coverage.

- Non profit organizations: the PRC database includes the data breaches reported by other non profit organizations than PRC, for instance (Databreaches.net, n.d.).

While merging different sources of notifications increases the scope of the PRC chronology, it also introduces some heterogeneity among the reported events, since each source reports a particular kind of claims. Additionally, the proportion of reported events from a given source fluctuates through time, as shown in Figures 1 and 2.




a) Non profit organization

b) Media

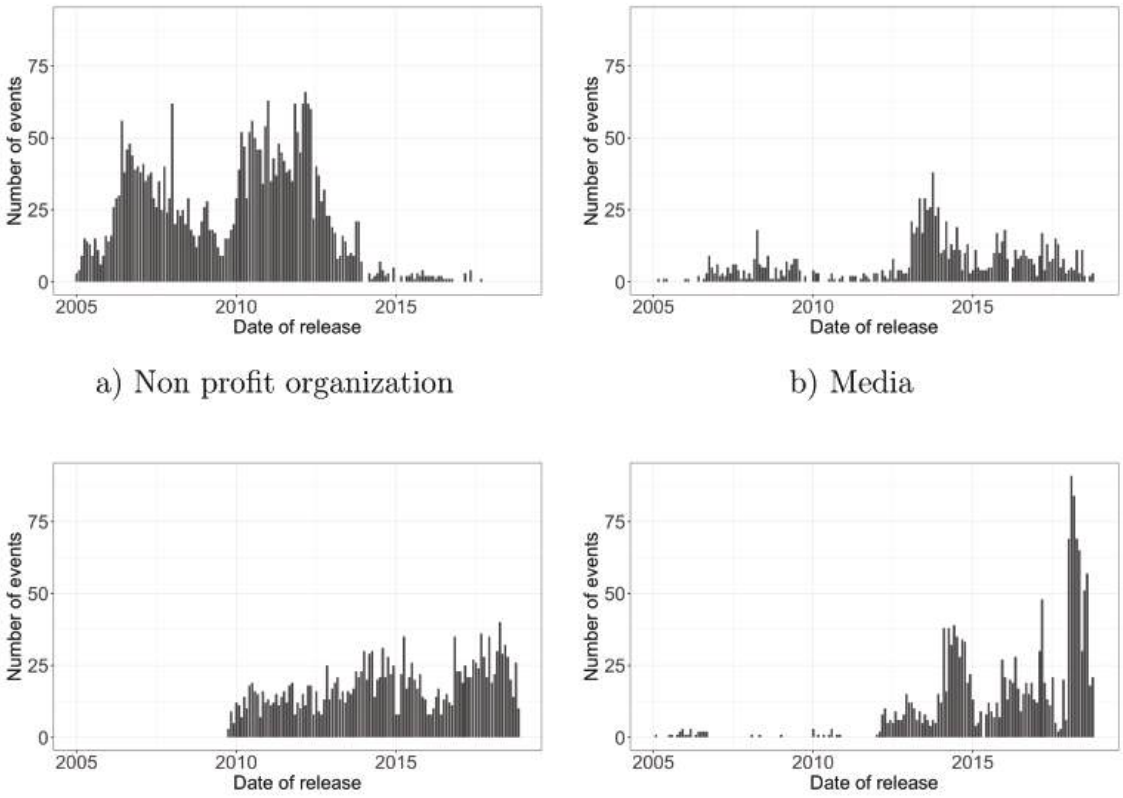
c) US GA: Federal - HIPAA

d) US GA: State

Figure 1: Scatter plots of data breaches listed in the PRC database (the x-axis is the date of the release and the y-axis is the number of records) depending on the source of information.

## 2.3 Heterogeneity and inconsistencies in PRC database

The way the database has been fed has evolved over time. These changes have had an impact on our main objective, which is to analyze the severity of these events. Indeed one may for example guess that cyber claims that were exposed by media are more likely to




a) Non profit organization

b) Media

c) US GA: Federal - HIPAA

d) US GA: State

Figure 2: Barplots of data breaches listed in the PRC database (the x-axis is the date of the release and the y-axis is the number of records) depending on the source of information.

be more "spectacular" (and hence more severe). This intuition will be confirmed by the quantitative results of Section 3.

Moreover, a short descriptive analysis of the severity variable ("number of records", see Table 4) shows that it is highly volatile. One can note an important difference between the median of the number of records (2 000) and the empirical mean (1.821 millions) because the latter is mainly driven by extreme events (the largest having 3 billions of records). This important dispersion is expected, due to the extreme variety of situations considered in the database. This pleads for reducing this heterogeneity by introducing appropriate risk classes, and in which we could separate the sources of information if they appear to be correlated with the severity of the claim. To determine such classes, our procedure relies on regression trees which are described in Section 3 below. They offer the advantage to perform an automatic clustering, without any a priori on the covariates.

Table 4: Descriptive statistics for the variable "Number of records" depending on the source of information (first column). $q_\alpha$ denotes the empirical $\alpha-$quantile, that is such that $\alpha\%$ of observations are smaller than $q_\alpha$.

| | Number | Mean | $q_{0.25}$ | Median | $q_{0.75}$ | $q_{0.9}$ | $q_{0.95}$ | Max |
|---|---|---|---|---|---|---|---|---|
| Total | 6 160 | 1 821 682 | 597 | 2 000 | 10 891 | 70 000 | 300 000 | 3 000 000 000 |
| US GA: Federal - HIPAA | 1 949 | 84 358 | 981 | 2 300 | 8 009 | 28 440 | 75 016 | 78 800 000 |
| US GA: State | 888 | 89 377 | 20 | 4 010 | 2 403 | 18 000 | 63 825 | 40 000 000 |
| Media | 595 | 16 208 786 | 1 400 | 11 266 | 137 193 | 4 420 000 | 41 029 090 | 3 000 000 000 |
| Nonprofit organization | 2 309 | 422 623 | 380 | 2 000 | 14 000 | 86 333 | 247 200 | 250 000 000 |
| Unknown | 419 | 853 736 | 958 | 2 300 | 9 154 | 30 194 | 61 863 | 191 000 000 |

# 3 Regression Trees and extreme value analysis

Regression trees are a convenient tool when one wants to simultaneously predict a response and filter heterogeneity by determining clusters among the data. In the sequel, $Y$ denotes a response variable (a "cost" variable representing the severity of the claim), and $\mathbf{X} \in \mathbb{R}^d$ some covariates (the circumstances of the claim, the victim(s), the source which detected the event...). Our observation set is composed of i.i.d. replications $(Y_i, \mathbf{X}_i)_{1\leq i\leq n}$ of $(Y, \mathbf{X})$. Regression trees aim at determining "rules" to gather observations in risk classes depending on the values of their characteristics $\mathbf{X}_i$. Therefore they are particularly adapted to the situations where the variety of profiles of $\mathbf{X}_i$ induces some heterogeneity. The CART algorithm, used to compute the trees, is presented in Section 3.1. Depending on the purpose of regression trees (typically, in our situation, depending on whether we wish to investigate the center or the tail of the distribution), an appropriate loss function has to be defined in order to evaluate the quality of the tree and define splitting rules for the clustering part of the algorithm. Generalized Pareto regression trees, introduced in Section 3.2, rely on a splitting rule which is designed to focus on the tail of the distribution, due to key results in extreme value theory.

## 3.1 Regression Trees

Regression Trees are modeling tools that allow one to introduce modeling of (nonlinear) heterogeneity between the observations, by splitting them into classes on which different regression models are fitted. The aim is to retrieve a regression function $m^* = \arg\min_{m\in\mathcal{M}} E[\phi(Y, m(\mathbf{X}))]$, where, again, $Y$ is our response variable (the severity of a cyber claim in our case), $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ is a set of covariates, $\mathcal{M}$ is a class of target

functions on $\mathbb{R}^d$ and $\phi$ is a loss function that depends on the quantity we wish to estimate (see Section 3.1.2).

In the following, we will consider three different types of functions $\phi$ :

- the quadratic loss $\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$ corresponds to the situation where the objective is the conditional mean $m^*(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ and $\mathcal{M}$ is the set of functions of $\mathbf{x}$ with finite second order moment;
- the absolute loss $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$, where $m^*$ is the conditional median;
- a log-likelihood loss $\phi(y, m(\mathbf{x})) = -\log f_{m(\mathbf{x})}(y)$, where $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ is a parametric family of densities. This corresponds to the case where one assumes that the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ belongs to the parametric family $\mathcal{F}$ for all $\mathbf{x}$, with parameter $m(\mathbf{x})$ depending on $\mathbf{x}$.

This split of the data is performed in an iterative way, by finding at each step an appropriate simple rule (that is a condition on the value of some covariate) to separate the data into two more homogeneous classes. The procedure includes two phases: a "growing" phase which corresponds to the CART algorithm, and a "pruning" step which consists in the extraction of a subtree from the decomposition obtained in the initial phase. Pruning can therefore be understood as a model selection procedure. In Section 3.1.1, we describe a general version of the CART algorithm, and explain in Section 3.1.2 how an estimation of a regression model can be deduced from a tree obtained in this first phase. The pruning step is then described in Section 3.1.3.

### 3.1.1 Growing step: construction of the maximal tree

The CART algorithm consists in determining iteratively a set of "rules" $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \to R_j(\mathbf{x})$ to split the data, aiming at optimizing some objective function (also referred to as splitting criterion). More precisely, for each possible value of the covariates $\mathbf{x}$, $R_j(\mathbf{x}) = 1$ or 0 depending on whether some conditions are satisfied by $\mathbf{x}$, with $R_j(\mathbf{x})R_{j'}(\mathbf{x}) = 0$ for $j \neq j'$ and $\sum_j R_j(\mathbf{x}) = 1$. In case of regression trees, these partitioning rules have a particular structure, since they can be written as $R_j(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_1 \leq \mathbf{x} < \mathbf{x}_2}$ for some $\mathbf{x}_1 \in \mathbb{R}^d$ and $\mathbf{x}_2 \in \mathbb{R}^d$, and the comparison symbols have to be understood as componentwise comparisons. In other terms, if $d = 1$, rules can be identified as partitioning segments, if $d = 2$ they are rectangles (hyper-rectangles in the general case). The determination of these rules from one step to another can be represented as a binary tree, since each rule $R_j$ at

step $k$ generates two rules $R_{j1}$ and $R_{j2}$ (with $R_{j1}(\mathbf{x}) + R_{j2}(\mathbf{x}) = 0$ if $R_j(\mathbf{x}) = 0$) at step $k+1$. The algorithm can be summarized as follows:
**Step 1:** $R_1(\mathbf{x}) = 1$ for all $\mathbf{x}$, and $n_1 = 1$ (corresponds to the root of the tree).
**Step k+1:** Let $(R_1, ...R_{n_k})$ denote the rules obtained at step $k$. For $j = 1, \ldots, n_k$,

- if all observations such that $R_j(\mathbf{X}_i) = 1$ have the same characteristics, then keep rule $j$ as it is no longer possible to segment the population;
- else, rule $R_j$ is replaced by two new rules $R_{j1}$ and $R_{j2}$ determined in the following way: for each component $X^{(l)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, define the best threshold $x_{j\star}^{(l)}$ to split the data, such that $x_{j\star}^{(l)} = \arg\min_{x^{(l)}} \Phi(R_j, x^{(l)})$, with

$$\begin{aligned}\Phi(R_j, x^{(l)}) &= \sum_{i=1}^{n} \phi(Y_i, \widehat{m}(R_j)) R_j(\mathbf{x}) \\ &- \sum_{i=1}^{n} \phi(Y_i, m_{l-}(\mathbf{X}_i, R_j)) \mathbf{1}_{X_i^{(l)} \leq x^{(l)}} R_j(\mathbf{x}) \\ &- \sum_{i=1}^{n} \phi(Y_i, m_{l+}(\mathbf{X}_i, R_j)) \mathbf{1}_{X_i^{(l)} > x^{(l)}} R_j(\mathbf{x}),\end{aligned}$$

  where

$$\begin{aligned}\widehat{m}(R_j) &= \arg\min_{m \in \mathcal{M}} \sum_{i=1}^{n} \phi(Y_i, m(\mathbf{X}_i)) R_j(\mathbf{X}_i), \\ m_{l-}(x, R_j) &= \arg\min_{m \in \mathcal{M}} \sum_{i=1}^{n} \phi(Y_i, m(\mathbf{X}_i)) \mathbf{1}_{X_i^{(l)} \leq x} R_j(\mathbf{X}_i), \\ m_{l+}(x, R_j) &= \arg\min_{m \in \mathcal{M}} \sum_{i=1}^{n} \phi(Y_i, m(\mathbf{X}_i)) \mathbf{1}_{X_i^{(l)} > x} R_j(\mathbf{X}_i).\end{aligned}$$

  Then, select the best component index to consider: $\widehat{l} = \arg\min_l \Phi(R_j, x_{j\star}^{(l)})$.
  Define the two new rules $R_{j1}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(\widehat{l})} \leq x_{j\star}^{(\widehat{l})}}$, and $R_{j2}(\mathbf{x}) = R_j(\mathbf{x}) \mathbf{1}_{x^{(\widehat{l})} > x_{j\star}^{(\widehat{l})}}$.
- Let $n_{k+1}$ denote the new number of rules.

**Stopping rule:** stop if $n_{k+1} = n_k$.

As it has already been mentioned, this algorithm has a binary tree structure. The list of rules $(R_j)_{1 \leq j \leq n_k}$ are identified with the leaves of the tree at step $k$, and the number of leaves of the tree is increasing from step $k$ to step $k+1$.

In this version of the CART algorithm, all covariates are continuous or $\{0,1\}-$valued. For qualitative variables with more than two modalities, they must be transformed into binary variables, or the algorithm must be slightly modified so that the splitting step of each $R_j$ should be done by finding the best partition into two groups on the values of the modalities that minimizes the loss function. This can be done by ordering the modalities with respect to the average value—or the median value—of the response for observations associated with this modality.

The stopping rule can also be slightly modified to ensure that there is a minimal number of points of the original data in each leaf of the tree at each step.

### 3.1.2 From the tree to the regression function

From a set of rules $\mathcal{R} = (R_j)_{j=1,\dots,s}$, an estimator $\widehat{m}^{\mathcal{R}}$ of the function $m$ is given by

$$\widehat{m}^{\mathcal{R}}(\mathbf{x}) = \sum_{j=1}^{s} \widehat{m}(R_j) R_j(\mathbf{x}).$$

The final set of rules $\mathcal{R}^M$ obtained from the CART algorithm is called the maximal tree. This leads to a trivial estimator of $m$, since either the number of observations in a leaf is one, or all observations in this leaf have the same characteristics $\mathbf{x}$. The pruning step consists in extracting a subtree from the maximal tree, achieving a compromise between simplicity and good fit.

### 3.1.3 Selection of a subtree: pruning algorithm

For the pruning step, a standard way to proceed is to use a penalized approach to select the appropriate subtree (see Breiman et al., 1984; Gey and Nédélec, 2005). A subtree $\mathcal{S}$ of the maximal tree is associated with a set of rules $\mathcal{R}^{\mathcal{S}} = (R_1^{\mathcal{S}}, \dots, R_{n_{\mathcal{S}}}^{\mathcal{S}})$ of cardinality $n_{\mathcal{S}}$. One then selects the subtree $\widehat{\mathcal{S}}(\alpha)$ that minimizes the criterion

$$C_\alpha(\mathcal{S}) = \sum_{i=1}^{n} \phi(Y_i, m^{\mathcal{R}^{\mathcal{S}}}(\mathbf{X}_i)) + \alpha n_{\mathcal{S}}, \tag{3.1}$$

among all subtrees of the maximal tree, where $\alpha$ is a positive constant. Hence, the trees with large numbers of leaves (i.e. of rules) are penalized compared to smaller ones. To determine this tree $\widehat{\mathcal{S}}(\alpha)$, it is not necessary to compute all the subtrees from the maximal tree. It suffices to determine, for all $K \geq 0$, the subtree $\mathcal{S}_K$ which minimizes the criterion (3.1) among all subtrees $\mathcal{S}$ with $n_{\mathcal{S}} = K$, and then to choose the tree $\mathcal{S}_K$ which minimizes

the criterion with respect to $K$. From (Breiman et al., 1984, p.284–290), these $\mathcal{S}_K$ are easy to determine, since $\mathcal{S}_K$ is obtained by removing one leaf to $\mathcal{S}_{K+1}$.

The penalization constant $\alpha$ is chosen using a test sample or $k-$fold cross-validation. In the first case, data are split into two parts before growing the tree (a training dataset of size $n$ and a test sample which is not used in computing the tree). In the second case, the dataset is randomly split into $k$ parts which successively act as training or test sample.

Let $\widehat{\alpha}$ denote the penalization constant calibrated using the test sample or the $k-$fold cross-validation approach, our final estimator is then $\widehat{m}(\mathbf{x}) = m^{\widehat{\mathcal{S}}(\widehat{\alpha})}(\mathbf{x})$.

## 3.2 Generalized Pareto Regression trees for analyzing the tail of the distribution

Since the severity of cyber events is highly volatile, it seems necessary to develop a specific approach for the tail of distribution. In Section 3.2.1, we recall why Generalized Pareto (GP) distributions naturally appear in the analysis of heavy-tailed variables. This motivates our GP trees described in Section 3.2.2.

### 3.2.1 Peaks over threshold method for extreme value analysis

Extreme value analysis is the branch of statistics which has been developed and broadly used to handle extreme events, such as extreme floods, heat waves episodes or extreme financial losses (Embrechts et al., 1997; Katz et al., 2002). Given a series of independent and identically distributed observations $Y_1, Y_2, \ldots$ with an unknown survival function $\bar{F}$ (that is $\bar{F}(y) = P(Y_1 > y)$). A natural way to define extreme events is to consider the values of $Y_i$ which exceed some high threshold $u$. The excesses above $u$ are then defined as the variables $Y_i - u$ given that $Y_i > u$. The asymptotic behavior of extreme events is characterized by the distribution of the excesses which is given by

$$\overline{F}_u(y) = P[Y_1 - u > y \mid Y_1 > u] = \frac{\overline{F}(u+y)}{\overline{F}(u)},\ y > 0\,.$$

If $\overline{F}$ satisfies the following property

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma},\ \forall y > 0, \tag{3.2}$$

with $\gamma > 0$, then

$$\lim_{u\to\infty} \sup_{y>0} |\overline{F}_u(y) - \overline{H}_{\sigma_u,\gamma}(y)| = 0 \tag{3.3}$$

for some $\sigma_u > 0$ and $\overline{H}_{\sigma_u,\gamma}$ necessarily of the form

$$\overline{H}_{\sigma_u,\gamma}(y) = \left(1 + \gamma \frac{y}{\sigma_u}\right)^{-1/\gamma}, \ y > 0. \tag{3.4}$$

Here, $\sigma_u > 0$ is a scale parameter and $\gamma > 0$ is a shape parameter, which reflects the heaviness of the tail distribution. Especially, if $\gamma \in ]0; 1[$, the expectation of $Y$ is finite whereas if $\gamma \geq 1$ the expectation of $Y$ is infinite. In our situation of highly volatile severity variables, the assumption $\gamma > 0$ is reasonable and supported by the empirical results of (Maillart and Sornette, 2010) (who even estimated $\gamma > 1$). The result from (Balkema and De Haan, 1974) states that, if the survival function of the normalized excesses above a high threshold $u$ weakly converges toward a non-degenerate distribution, then the limit is a Generalized Pareto distribution (see also Pickands, 1975).

In practice, the so-called Peaks over Threshold (PoT) method has been widely used since 1990 (see Coles, 2001; Davison and Smith, 1990). It consists in choosing a high threshold $u$ and fitting a GP distribution on the excesses above that threshold $u$. The estimation of the parameters $\sigma$ and $\gamma$ may be done by maximizing the GP likelihood. The choice of the threshold $u$ implies a balance between bias and variance. Too low a threshold is likely to violate the asymptotic basis of the model, leading to bias; too high a threshold will generate few excesses with which the model can be estimated, leading to high variance. The standard practice is to choose as low a threshold as possible, subject to the limit model providing a reasonable approximation.

**Remark 3.1** *Property* (3.2) *is called regular variation. When $\gamma > 0$, we say that $\overline{F}$ is heavy-tailed, meaning that its tail decreases polynomially. Usual distributions as Pareto, Cauchy and Student distributions satisfy this property. For more details, see (De Haan and Ferreira, 2007, Appendix B).*

### 3.2.2 Generalized Pareto Regression Trees

When it comes to studying the severity of cyber claims, we expect to see a potential heterogeneity in the tail of the distribution. In order to improve the precision of our analysis, a natural idea is to study the impact of the circumstances of the claim and of the characteristics of the victim on the response variable. In our regression framework, for each value of the covariate $\mathbf{x}$, we assume the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ to be heavy-tailed, but the parameters $\gamma$, $\sigma$ (and the threshold $u$ above which the GP distribution approximation seems satisfactory) depend on $\mathbf{x}$. More precisely, this means

that (3.2) becomes

$$\lim_{t\to\infty} \frac{\overline{F}(ty|\mathbf{x})}{\overline{F}(y|\mathbf{x})} = y^{-1/\gamma(\mathbf{x})}, \forall y > 0, \tag{3.5}$$

where $\overline{F}(y|\mathbf{x}) = \mathbb{P}(Y \geq y|\mathbf{X} = \mathbf{x})$ with $\gamma(\mathbf{x}) > 0$ for all $\mathbf{x}$, and (3.3) becomes

$$\lim_{u(\mathbf{x})\to\infty} \sup_{y>0} |\overline{F}_{u(\mathbf{x})}(y \mid \mathbf{x}) - \overline{H}_{\sigma_{u(\mathbf{x})}(\mathbf{x}),\gamma(\mathbf{x})}(y)| = 0. \tag{3.6}$$

where $\overline{F}_{u(\mathbf{x})}(y \mid \mathbf{x}) = P[Y - u(\mathbf{x}) > y \mid Y > u(\mathbf{x}), \mathbf{X} = \mathbf{x}]$.

The idea is then to apply the procedure of Section 3 to the observations $(Y_i - u(\mathbf{X}_i), \mathbf{X}_i)$ for which $Y_i \geq u(\mathbf{X}_i)$, using the Generalized Pareto log-likelihood as split function, that is

$$\phi(y, m(\mathbf{x})) = -\log(\sigma(\mathbf{x})) - \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{y\gamma(\mathbf{x})}{\sigma(\mathbf{x})}\right),$$

where $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ (we use the notation $\sigma(\mathbf{x}) = \sigma_{u(\mathbf{x})}(\mathbf{x})$ to simplify). The function $u(\mathbf{x})$ is an input of the procedure, and has to be taken so that the GP distribution fit seems appropriate for all considered values of covariates. The practical choice of this function is a delicate problem (see Section 4 in Beirlant and Goegebeur, 2004). To simplify, we consider in the following a fixed threshold $u(\mathbf{x}) = u$ for all values of covariates $\mathbf{x}$. The threshold $u$ is chosen large enough so that the GP approximation is correctly fitted to the data (practical choice of this parameter will be discussed in Section 4.2, see also Remark 3.2 below). In the end, the leaves of the tree identify classes, each corresponding to different tail behaviors (that is with different values of $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$, the function $m$ being constant on each leaf.

Compared to competing approaches in extreme value regression, the advantage of the procedure is to introduce discontinuities in the regression function while parametric approaches, like in (Beirlant and Goegebeur, 2003), suppose a form of linearity. More flexible nonparametric approaches, as in (Beirlant and Goegebeur, 2004), rely on smoothing techniques that require covariates to be continuous. Chavez-Demoulin et al. (2015) propose a semiparametric framework to separate the continuous covariates from the discrete ones. Smoothing splines are used to estimate nonparametrically the continuous part, while the influence of discrete covariates is captured by a parametric function. Due to the nice properties of this technique applied on operational risk data in (Chavez-Demoulin et al., 2015), we compare the results of our GP regression tree approach to their procedure in Section 4.3.

**Remark 3.2** *As already stated, the conditional version of* (3.4) *used in extreme value regression leads to the introduction of a threshold function* $u(\mathbf{x})$ *that potentially depends*

*on* $\mathbf{x}$. *A possibility would be to adapt the CART algorithm to select, at each step, a choice of threshold that could be different in each leaf. However, this complexifies considerably the technique, and we did not consider it.*

# 4 PRC database analysis with regression trees

In this section, we apply the different variations of the regression tree approach of Section 3 to the response variable $Y =$ "Number of Records" in the PRC database. Let us note that, despite its name, this variable can be considered as continuous, since this number of records takes a wide range of values (see Table 4) with few ties (caused by rounded numbers). Section 4.1 describes regression tree analysis of the central part of the distribution, while the tail part is considered in Section 4.2, applying GP trees. Comparison with the fit of a GAM model as in Chavez-Demoulin et al. (2015) is shown in Section 4.3. Section 4.4 shows how our two regression tree approaches (one for the central part of the distribution, one for the tail) can be combined to provide a global analysis of the distribution. A discussion on the insurability of cyber risk —which, from a probabilist point of view, is closely related to the value of the tail parameter $\gamma$—is done in Section 4.5.

## 4.1 Central part of the severity distribution

In order to estimate the conditional mean $E[Y|\mathbf{X} = \mathbf{x}]$, with a regression tree, the loss function $\phi$ has to be chosen as the quadratic loss $\phi(y, m(\mathbf{x})) = (y-m(\mathbf{x}))^2$. The conditional mean is particularly important in view of computing a pure premium in insurance (pure premium corresponds to estimating the expectation of the cost, which requires to estimate the frequency of occurrence and the mean value of a claim), but this indicator is not robust, due to its sensitivity to large observations. Let us also observe that this conditional expectation may even not be defined for some values of $\mathbf{X}$ since $Y$ is heavy-tailed. Since the variable $Y$ we study is highly volatile, investigating the conditional median of the distribution of $Y|\mathbf{X} = \mathbf{x}$ (that is $med(Y|\mathbf{X} = \mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq 1/2\}$, where $F(y|\mathbf{x}) = P(Y \leq y|\mathbf{X} = \mathbf{x})$) may be more stable. Estimating the conditional median corresponds to the choice of the absolute loss as the loss function, that is $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$.

We fit regression trees using these two loss functions. These trees are computed using the `R` package `rpart` (see Therneau and Clinic, 2018), by using a user defined split function. The pruning step has been done thanks to a 10–fold cross validation used for

error measurement and the selection of a proper subtree. The obtained trees are shown in Figure 3.

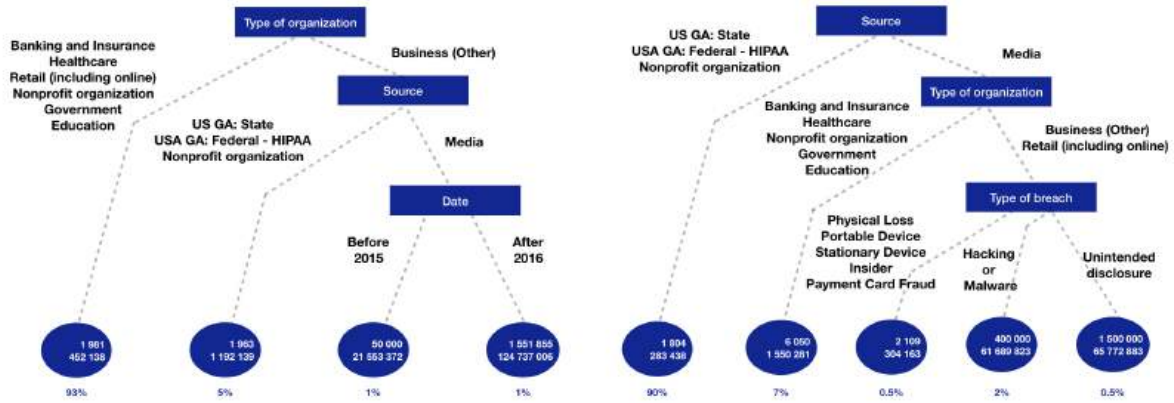



Figure 3: Trees obtained from the CART algorithm based on the quadratic (left-hand side) and the absolute (right-hand side) losses. For each leaf, the value of the empirical median (first line) and mean (second line) are given. Percentage of observations affected to each leaf is mentioned.

The structure of the trees is different for the conditional median compared to the conditional expectation, although some similarities exist. For example, the category of victims "Business (Other)" seems generally associated with higher severity: for the mean tree, all events are gathered in the same leaf, except for those affecting this category of targets, which are associated with the largest predicted values. The picture is slightly different for the median tree: the highest predicted values are still linked with the "Business (Other)" category, but only under particular circumstances. In both cases, the Media source is generally associated with larger events.

The leaves of the trees determine clusters. If one wants to get a distribution for the claim severity, a distribution can be fitted on each leaf, see the supplementary material (Section 1.3) for more details.

**Remark 4.1** *Our procedure consists in first determining clusters (using regression trees with $L^2$ or $L^1$ loss), and then fitting log-normal distributions to each leaf. This last step is only required if one wishes to have a global model for the distribution of $Y$. One could directly use a log-normal log-likelihood as split criterion to obtain different clusters that should improve the log-normal fit. The reason for not choosing this path is because our purpose is essentially to understand which covariates drive the central part of the distri-*

*bution (in order to compare it to the study of the tail, which is our main objective), but comparisons with a direct log-normal fit can be found in Section 2.1 of the supplementary material. The $L^2$ and $L^1$ trees are supposed to provide clusters that are based on the expectation or the median, with no particular assumption on the conditional distribution of $Y$ in each leaf. In fact, fitting these trees can be done even in the case where all the leaves do not correspond to the same family of distribution (one may fit a gamma distribution in one leaf, a log-normal in another).*

## 4.2 Tail part of the severity distribution

In view of applying the GP regression tree approach of Section 3.2.2, our first task is to determine the threshold $u$ above which the GP distribution approximation seems reasonable. This choice is made from the Hill plot (shown in the appendix, Figure 6) (see Resnick, 2007, pp 85–89 for more details on Hill plots). From the shape of the curve, we chose $u = 27\ 999$ (which corresponds to a stabilization of the Hill plot) which leads to keep the 1 000 highest observations (around 16% of the total number of breaches). Let us note that Hill plots are not designed for regression methods. In our context, as already pointed in Remark 3.2, one could look at thresholds depending on the covariates. See also Section 4 in Beirlant and Goegebeur (2004) who discussed this choice of thresholds in extreme value regression, and Section 4 of the supplementary material.

Figure 4 shows the obtained GP tree (fitted using the library `rpart` in `R`, with the appropriate user defined loss function), and variable importance is evaluated in Table 8 (more details on the computation of variable importance can be found in the supplementary material, Section 4). The confidence intervals for the parameters estimates in each leaf are reported in Table 5. Goodness-of-fit for the different leaves is shown through quantile-quantile plots in Section 7.2, see Figure 7. Let us first note that the structure of the GP tree is quite different from the ones obtained from the central part of the distribution. The estimated values of the shape and scale parameters on each leaf have first to be compared to the values obtained if we fit a GP distribution to the whole set of observations greater than $u$. In this case, maximum likelihood estimation leads to $\widehat{\sigma} = 48\ 243$ (the 95% confidence interval is $[40\ 685; 55\ 802]$) and $\widehat{\gamma} = 2.16$ (the 95% confidence interval is $[1.96; 2.36]$). The worst case scenario, corresponding to the leaf with shape estimate 3.26, is even worse than this benchmark. Yet, the two other leaves, representing 82% of the extreme events, are "lighter" (although still associated with a shape parameter greater than 1, that is such that the expectation is not finite).

Table 5: Generalized Pareto parameters estimated by the Generalized Pareto Regression Tree based on excesses and the 95% confidence intervals (given under brackets).

| | Leaf 1 | Leaf 2 | Leaf 3 |
|---|---|---|---|
| $\gamma$ | 1.43 [1.21;1.64] | 1.72 [1.41;2.04] | 3.26 [2.62;3.91] |
| $\sigma.10^{-5}$ | 0.36 [0.29;0.43] | 0.76 [0.55;0.97] | 1.82 [0.98;2.67] |

Moreover, let us observe that the major part of these events corresponds to a shape parameter equal to 1.43, which is close to the estimate of the tail distribution index provided by Maillart and Sornette (2010).

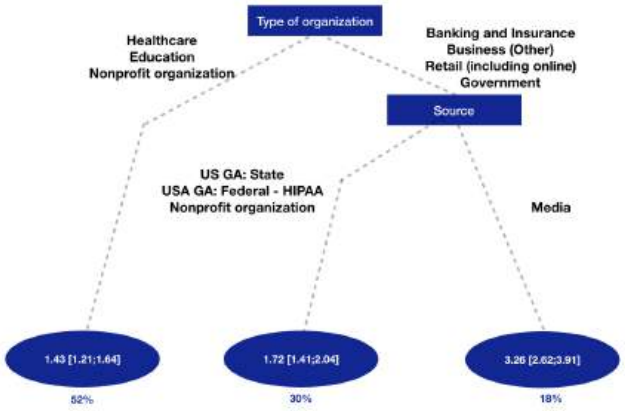



Figure 4: Tree obtained from the CART algorithm based on the Generalized Pareto log-likelihood splitting rule (fitted on the observations exceeding the threshold $u$). For each leaf, the estimates of $\gamma$ and their 95% confidence intervals are given.

**Remark 4.2** *The value $\widehat{\gamma} = 2.16$ obtained from the whole sample implies that $E[Y] = \infty$. This indicates that the quadratic based regression method may not only lack robustness, but leads to ill-defined estimates (since the conditional expectation is not defined, at least for some leaves in the tree).*

## 4.3 Comparison with Generalized Additive Models

To compare the GP regression tree with competing extreme value regression approaches, we implemented the methodology developed by Chavez-Demoulin et al. (2015), that is using a Generalized Additive Model based on GP distributions for studying the tail (that is for $Y \geq u$). We will use the notation GAM GPD to refer to this technique. A short

description of this technique is provided in the supplementary material (Section 3.1), along with estimates for the values of the model parameters.

Table 6 compares the fits of the GP tree with GAM GPD. Classical GP distribution fit (that is, using the PoT approach and without taking attention to the impact of the covariates) is also considered as a benchmark. We see that, in terms of log-likelihood and Akaike criterion (AIC), both regression techniques significantly improve this benchmark model, with a slightly better fit for the GP tree.

Table 6: Comparison of extreme value theory methodologies

| | Covariates used for $\sigma$ | Covariates used for $\gamma$ | LL | AIC |
|---|---|---|---|---|
| GP distribution | - | - | -2122 | 4249 |
| GPD GAM | Organization and Source | Date and Organization | -2031 | 4098 |
| GP tree | Type of organization and Source | Type of organization and Source | -2024 | 4072 |

## 4.4 Global distribution analysis

The GP tree of Figure 4 only provides an analysis of the distribution above a threshold $u$. If one wishes a global distribution, one must combine this approach with an analysis of the central part of the distribution. On the other hand, the analysis of Section 4.1 provides such a global analysis, but without taking the tail into account. Moreover, going back to the trees of Figure 3, one can notice that, in each leaf, there is a significant difference between the value of the mean and the value of the median, as it is the case in the global set of observations (see Section 2.3). This invites us to look at a regression tree computed using the same method as in Section 4.1 (using absolute loss since less sensitive to large observations, the tree obtained via quadratic loss is shown in the supplementary material) but only on observations smaller than the threshold $u$. To summarize, observations are cut in two parts: observations with $Y \leq u$ are fitted using a regression tree based on absolute loss, while observations larger than $u$ are fitted using the GP tree of Section 4.2.

This leads to the regression tree of Figure 5. We see that the gap between the empirical median and the empirical mean in each leaf has been drastically reduced. On the other hand, the tree has a different structure than the one obtained from the global set of observations in Figure 3, which shows that the presence of extreme values influences the obtained clusters.

A log-normal distribution (truncated by $u$) is fitted on the leaves of the absolute tree. The corresponding parameters are listed in Table 7.

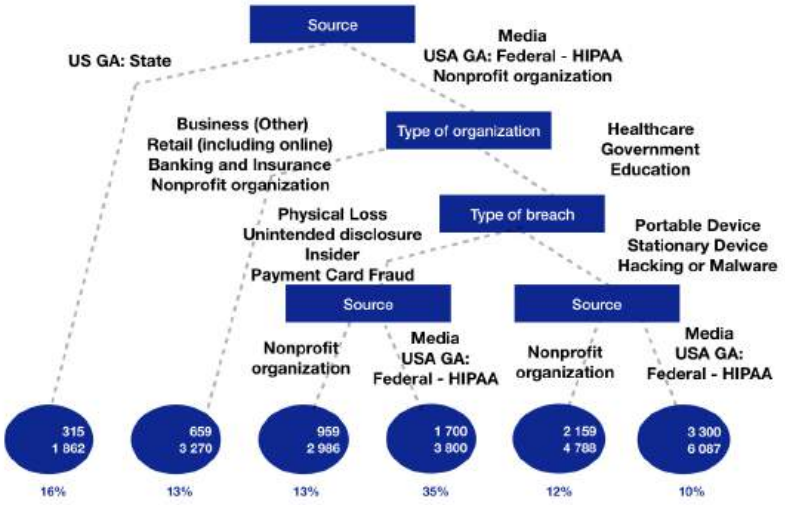



Figure 5: Tree obtained from the CART algorithm based on the absolute loss fitted on the observations such that the variable "Number of Records" is less than $u$. For each leaf, the median (above) and the mean (below) are given.

Table 7: Truncated log normal parameters estimated by the absolute loss tree based on data below $u$ . The parameter $\mu$ is the location parameter (expectation of the logarithm of the variable) and $\sigma$ the scale parameter (standard deviation of the logarithm of the variable). Leaves are numerated from left to right according to the representation of the tree from Figure 5. 95% intervals are given in brackets

| | Leaf 1 | Leaf 2 | Leaf 3 |
|---|---|---|---|
| $\mu$ | 5.62 [5.30;5.94] | 6.79 [6.54;7.04] | 6.95 [6.73;7.16] |
| $\sigma$ | 3.37 [3.13;3.61] | 2.46 [2.26;2.65] | 2.19 [2.02;2.35] |
| | Leaf 4 | Leaf 5 | Leaf 6 |
| $\mu$ | 7.64 [7.57;7.70] | 8.20 [7.85;8.54] | 8.72 [8.36;9.07] |
| $\sigma$ | 1.31 [1.26;1.36] | 2.27 [2.04;2.49] | 1.91 [1.69;2.13] |

To obtain the global distribution of the variable $Y =$ "Number of records", the combination of the results of the trees from Figures 5 and 4 and Table 7 is done in the following way. We consider that the conditional distribution of $Y$ is a mixture variable with same distribution as $\delta Z_1 + (1-\delta)Z_2$, where :

- $\delta$ is a Bernoulli random variable independent from $\mathbf{X}$, and $p = P(\delta = 1)$ is the probability for an observation $Y_i$ to be smaller than the threshold $u$;
- $Z_1|\mathbf{X} = \mathbf{x}$ has a distribution given by the absolute tree of Figure 5 (where each leaf is

associated with a truncated log-normal distribution determined by the parameters of Table 7);

- $Z_2|\mathbf{X} = \mathbf{x}$ has a distribution given by the GP tree of Figure 4;
- $\delta$ is independent from $(Z_1, Z_2)$ and $Z_1$ and $Z_2$ are independent conditionally to $\mathbf{X}$.

Let us recall that our estimate for $p$, in the PRC case, is the proportion of observations whose number of records is smaller than $u$, that is 0.84.

To complete this section, Table 8 reports the variable importance for both trees involved in this scheme. This confirms the relevance of separating the tail from the center of the distribution, since the variables driving the tail are different (at least in term of hierarchy) from the ones driving the center.

Table 8: Variable importance for the absolute tree of Figure 5 and for the Generalized Pareto tree of Figure 4 (in %).

| | Source | Type of breach | Type of organization | Year |
|---|---|---|---|---|
| Central part tree | 47 | 17 | 18 | 17 |
| Tail part tree | 35 | - | 48 | 17 |

## 4.5 Insurability of cyber risk

If we focus only on the tail of the distribution, the model fitted by the GP regression tree induces a mixture of three GP distributions for the unconditional distribution of $Y$. The advantage, compared to fitting a single GP distribution to all data larger than $u$, is that the tail index that the resulting shape index tends to be too pessimistic. Theoretically speaking, the tail index estimation of the global distribution should converge towards the worst tail index of the elements of the GP mixture. The GP tree technique presents the advantage to allow identification of some groups of claims that are still associated with a heavy tail behavior, but with more moderate consequences (in our example, all three leaves of the tree in Figure 4 corresponds to an infinite expectation, but let us recall that we are working with a proxy variable for the real amount of a claim). Hence we argue that using such techniques on more elaborate insurance databases can be a valuable tool to identify which types of cyber risks should be excluded from the policies (if the insurance company is unable to manage it), and potentially be used to reduce the premium if the insured population is associated with a lower risk.

# 5 An illustration on virtual cyber portfolios

The statistical approach performed in Section 4 is done using all the covariates present in the public PRC database. The aim is to achieve the best possible understanding of what drives the severity of cyber events. On the other hand, if one wishes to combine this analysis with an insurance perspective, an adaptation has to be made. It is the purpose of the present section to explain how this can be done. The question of coupling public database with intern information (from the history of the portfolio) is indeed fundamental in the context of cyber insurance, due to the lack of experience on the risk for many companies.

In this paper, we only address how to use the GP regression trees to project the result of a cyber insurance portfolio. In a real-life situation, this has to be combined with more reliable (but poorer) intern data. We perform simulations on four portfolios of 1 000 policies, where each portfolio is composed of policyholders coming from only one of the following sectors of activities: BSF, BSO, BSR, or MED. The simulations use the different models we fitted on data. Nevertheless, the severity analysis we performed in Section 4 must be completed by three additional assumptions to produce an evaluation of the cost:

1. a transformation $f$ that maps a number of records $Y$ to a financial loss $f(Y)$;
2. a frequency analysis to model the occurrence of cyber claims, that is a distribution for $N_i$= number of incidents for the $i$th policyholder within 1 year;
3. once a claim has occurred, a probability distribution to determine the type of incident: indeed, since the type of breach has been seen to have a significant impact on the distribution of the claim size, we need to distinguish between these different categories of claims.

The total loss of the portfolio is then

$$S = \sum_{j=1}^{1000} \sum_{j=1}^{N_i} f(Y_{i,j}),$$

where $(Y_{i,j})_{1 \leq i \leq n, 1 \leq j \leq N_i}$ are the number of records for the claims of policyholder $i$ (the number of records are supposed independent from $N_i$ in this simple model). The distribution of $S$ is then deduced from the points 1 to 3 above. In Sections 5.1 to 5.3, we address successively each of these points. We then explain the simulation procedures we use to evaluate the total loss of each portfolio in Section 5.4.

## 5.1 Loss quantification of a data breach

Jacobs (2014) provided a model to transform a volume of data breach $Y$ into a financial loss $L = f(Y)$. This model, which has also been used in Eling and Loperfido (2017), is based on data from Ponemon used Cost of Data Breach (CODB) reports of 2013 and 2014. The formula is

$$\log(L) = 7.68 + 0.76 \log(Y). \tag{5.1}$$

A limit for this formula and analysis is that, in 2014, data gathered by the Ponemon Institute LLC was restricted. Indeed, the highest observed data breach had a size of 100 000 records, far from the highest one of the actual PRC database (which is 3 billions). Hence we propose to use a modified version of (5.1), using additional information contained in the 2018 CODB report, in which, "for the first time, [one] attempt[s] to measure the cost of a data breach involving more than one million compromised records, or what [one] refer[s] to as a mega breach".

Since only two costs of mega breaches are publicly available in the 2018 CODB report, we performed a rough fit of a linear relationship between $\log L$ and $\log Y$, based on four points detailed in Table 9. These four points are the two mega breaches, and two artificial points obtained, for moderate breaches, by the application of Formula (5.1). This presents the advantage to take Formula (5.1) into account and benefit from the fact that it has been calibrated on a large (non public) database, while using the additional information on mega breaches.

This leads to the following formula that will be used in our loss quantification,

$$\log(L) = 9.59 + 0.57 \log(Y). \tag{5.2}$$

Table 9: Data breaches used to calibrate Formula (5.2): the costs of moderate breaches have been computed using Formula (5.1); the mega breaches are the only two communicated in CODB 2018.

| | Moderate breaches | | Mega breaches | |
|---|---|---|---|---|
| Number of records | 10 000 | 100 000 | 1 000 000 | 50 000 000 |
| Costs (in $) | 2 373 458 | 13 657 827 | 39 490 000 | 350 000 000 |
| Costs per record (in $) | 237 | 137 | 39 | 7 |

The difference between the results of Formulas (5.1) and (5.2) is shown in the supplementary material (Section 3.2). The results are relatively close for the most part of the events contained in the PRC database, but less pessimistic for the largest ones.

Clearly, we do not claim that Formula (5.2) is accurate for the association of a financial loss to the number of records. Our purpose is only to have a rough approximation of it. From the (public) data we have at our disposal, there is no way to pretend one is able to perform this evaluation with a good statistical precision. In practice, based on real loss data, the analysis that we provide can be seen as a rough benchmark that clearly needs to be improved by the use of more precise information.

Let us also note that Romanosky (2016) also studied the cost of data breaches using a private database gathering cyber events and associated losses. However, the obtained calibration requires information which is unavailable in the database used in this paper (but should be known from an insurance company when dealing with a real portfolio).

**Remark 5.1** *The GP regression tree of Figure 4 has been done on the variable $Y$ and not on the loss variable $f(Y)$. This choice has been done because we wanted to focus on the most reliable data, while Formula* (5.2) *is an approximation. However, the shape parameter of the GP distribution of $f(Y)$ can be easily deduced. Let us recall that this parameter is of most importance, since it gives us the decay of the survival function of $f(Y)$ (if this parameter is larger or equal to 1, $f(Y)$ has no expectation, and hence can be considered as "non-insurable" in a simplified vision of the problem). If $P(Y \geq y) \sim Cy^{-1/\gamma}$, where $\gamma > 0$ is the shape parameter of $Y$ and $C$ is a constant, considering $f(y) = \exp(\alpha + \beta \log y)$ leads to*

$$P(f(Y) \geq z) = P\left(Y \geq \exp\left(\frac{\log z - \alpha}{\beta}\right)\right) \sim C \exp\left(-\frac{\alpha}{\beta\gamma}\right) z^{-\frac{1}{\beta\gamma}}.$$

*Hence, the shape parameter of $f(Y)$ is $\beta\gamma$. In* (5.2)*, $\beta = 0.57$. Hence, the three leaves of the tree of Figure 4 have respective shape parameters* 0.82, 0.98, 1.86. *If we do not separate our claims into these three classes of risk, the shape parameters would have been* $0.57 \times 2.16 = 1.23$. *All of these numerical results should be taken carefully: the question of insurability is not so simple as determining if a GP shape parameter is smaller than one or not (and let us observe that, with Formula* (5.1)*, all shapes parameters would have been greater than 1), but it still shows the importance to distinguish tail behaviors depending on the covariates in order to identify more clearly which type of risks can be managed and which cannot.*

## 5.2 Frequency analysis

To provide an insurance pricing methodology, estimation of the annual frequency of claims is mandatory. The PRC database is not adequate to estimate this quantity rigorously. Nevertheless, we present here a possible way to roughly evaluate this frequency. This seems important for, at least, two reasons: 1) we want to provide an order of magnitude for the cost of cyber contracts ; 2) even for an insurance company with a cyber portfolio, it is likely that frequency would be poorly estimated only based on internal historical data: since the risk is new, the number of reported claims would be too small to perform an accurate estimation. Hence, we believe that the combination of these information with external information—including public databases like PRC—is essential to improve the evaluation of the risk.

An important issue with the PRC database is the lack of knowledge of the exposure to the risk. Typically, it is impossible to know from such data which part of the increase of reported claims along time is caused by an evolution of the risk, and which is caused by an instability in the way the database is fed. This can be seen, for example, from Figure 2. For example, the choice of PRC to stop gathering data breaches revealed by nonprofit organizations as from 2013 and a peak of data released by the media between 2015 and 2016 may be observed. Moreover, Bisogni et al. (2017) claim that the majority of data breaches proves to be unreported.

Hence, we propose two heuristics to derive a frequency analysis from the PRC database:

(H1) we restrain ourselves to companies listed in the PRC database that have been breached at least twice according to the PRC database. Since almost 90% of companies listed in PRC are reported only once, one may fear that the information about them is not completely reliable. On the other hand, a repeatedly reported company has more chances to have its major breaches exhaustively reported in the database. The frequency is estimated from companies that have been breached multiple times, considering that we are dealing with 1–truncated data.

(H2) we restrain ourselves to companies quoted on the New York Stock Exchange (NYSE) that have been breached at least once according to the PRC database. This idea has first been suggested by Wheatley et al. (2016). Here, 94% of companies of NYSE are absent from the PRC database. Assuming that no breach occurred for all of them seems unrealistic and would considerably lower the frequency: their absence is more likely due to the fact that these breaches have not been reported by the processes

of PRC. If a company is associated with 0 claim, it is therefore not certain that this absence from PRC is really caused by the absence of a breach, or by the fact that this entity was not in the scope of PRC. Hence, we consider that data from these companies is 0–truncated.

In the following, we consider two portfolios corresponding either to case (H1) (PRC portfolio) or case (H2) (NYSE portfolio). A summary of descriptive count statistics for both portfolios is given in the supplementary material (Section 1.2).

To model the number of claims striking a portfolio, we fit a Generalized Linear Model (GLM), considering the sector of activity as a covariate. For the PRC portfolio, we consider the sectors BSF, BSO, BSR, EDU, GOV and MED only, deliberately excluding the NGO sector because of lack of data on this category. The NYSE portfolio does not contain companies from sectors EDU, GOV and NGO. We consider two cases: a GLM based on a Poisson distribution, and one on a geometric distribution (for all $k \geq 0$, the probability that a geometric distribution is $k$ is $p(1-p)^k$, where $p$ is a parameter taking values in $(0,1)$). More precisely, these two models can be written as

$$g(E[N|\mathbf{X}]) = \mathbf{X}\beta \text{ , with } \begin{cases} N \sim \mathcal{P}(\lambda) \quad \text{and} \quad g(x) = \log(x). \\ \quad \text{or} \\ N \sim \mathcal{G}(p) \quad \text{and} \quad g(x) = \log\left(\frac{x}{1-x}\right). \end{cases} \tag{5.3}$$

On the PRC database, fitting indicators can be found in the supplementary material, Section 1.2, showing that the geometric GLM seems more adequate than the Poisson one.

## 5.3 Type of incident

The frequency of claims determined in Section 5.2 does not include the variety of cyber incidents: it is a global frequency, regardless the type of claims. If we want to simulate the impact on our insurance portfolio, we must simulate also a type of event once an event occurred. In our simulation scheme, the idea is to use a multinomial random variable to draw the type of event. We assume that the parameters only depend on the type of activity of the victim (which is the only variable available for the insurance company, among those present in the regression trees).

Let $S$ denote an indicator of the sector of activity, and $M$ denote the type of breach. We can write

$$P(M = m|S = s) = \frac{\mathrm{e}^{\beta_{s,0}+\beta_{s,m}}}{\sum_{m'} \mathrm{e}^{\beta_{s,0}+\beta_{s,m'}}},$$

where $\beta_{s,0}$ corresponds to a reference category (here we took as reference category the incidents for which the type of organization is unknown).

In full generality, this would lead to the estimation of a large number of coefficients, with to few data to calibrate them. To reduce the number of parameters, we used a LASSO dimension reduction technique (the log-likelihood is penalized using a $L^1-$penalty on the fitted coefficients $\beta_{s,m}$, with a parameter tuned through 10-fold cross validation, see e.g. Tibshirani (1996)). The matrix of fitted coefficients can be found in Section 1.2 of the supplementary material.

## 5.4 Results

We now show the impact of these models on our virtual portfolios. We recall that we consider four portfolios with 1 000 policyholders, each composed of entities of a single category among BSF, BSR, EDU and MED. The losses of each portfolio are simulated according to the following procedure:

1. For each policyholder, we simulate a number of claims under the geometric model of Section 5.2.

2. For each claim, we determine which type of incident has caused the claim from the multinomial distribution of Section 5.3.

3. We simulate the number of records accordingly to four methodologies, assuming, in each case, that the distribution is the same as the one given by one single source of information (US GA State or Media):

    - Clustering: we use the tree obtained with the absolute loss from Figure 3 to determine risk classes. The distribution of the claims in each leaf of the tree is considered as log-normal using the following set of parameters: $(7.56, 2.66)$ for leaf 1, $(8.88, 3.11)$ for leaf 2, $(8.19, 3.25)$ for leaf 3, $(12.67, 4.19)$ for leaf 4, $(13.47, 4.42)$ for leaf 5 (leaves numerated from left to right, first parameter (resp. second) is the expectation (standard deviation) of the logarithm of the log-normal variable).
    - GP regression tree: we use the combination of the trees of Figure 4 and 5, as described in Section 4.4. For the central part, log-normal distributions are used, with the fitted parameters of Table 7.

- GAM GPD: for comparison, we considered the approach developed by Chavez-Demoulin et al. (2015), which is exposed in detailed in the supplementary material.

4. We use (5.2) to convert this number of records into a financial loss (a comparison with the use of (5.1) can be found in the supplementary material, Section 3.2).

Results of these simulation procedures are summarized in Table 10. Let us first remark that, regarding the clustering approach based on a single tree (built using absolute loss), the difference between the median quantile $q_{0.5}$ and $q_{0.9}$ is much smaller than for the two other approaches. This was expected, due to the use of a GP distribution to model the tail for the last two models. On the other hand, the order of magnitude of all tree-based methods is much smaller than for the GAM GPD approach, although all sectors generally keep the same ranking in terms of severity from one model to another.

It is also interesting to notice that, in our tree-based methods, separating the tail from the central part of the distribution pushes up the value of the median quantile of the loss (of course the push on the $q_{0.9}$ quantile was expected, because a specific model has been done on the tail of the distribution). Through this phenomenon, one can observe once again the benefit of separating "extreme" observations from the others: their presence in the sample distorts the fitting of the tree and of the log-normal distributions in the leaves, even though we chose a relatively stable procedure through the use of the absolute loss.

Table 10: Comparison of median and 0.9–quantile depending on the methodology used (through columns) and the additional hypothesis regarding the source of information and the frequency portfolio (through lines). Quantities are given in million of dollars and have been obtained after 10 000 simulations.

| Modeling methodology | | | Clustering | | GAM GPD | | GP tree | |
|---|---|---|---|---|---|---|---|---|
| Source | Frequency | Organization | $q_{0.5}$ | $q_{0.9}$ | $q_{0.5}$ | $q_{0.9}$ | $q_{0.5}$ | $q_{0.9}$ |
| US GA State | (H1) | BSF | 286 | 424 | 2 561 | 58 764 | 433 | 1 320 |
| | | BSO | 363 | 522 | 4 806 | 156 055 | 572 | 1 755 |
| | | BSR | 235 | 358 | 1 851 | 41 563 | 351 | 1 074 |
| | | MED | 305 | 447 | 497 | 2 446 | 284 | 574 |
| | (H2) | BSF | 342 | 498 | 3 426 | 75 515 | 532 | 1 558 |
| | | BSO | 202 | 317 | 1 445 | 41 952 | 292 | 916 |
| | | BSR | 244 | 374 | 2 077 | 46 042 | 365 | 1 140 |
| | | MED | 223 | 345 | 332 | 1 695 | 203 | 434 |
| Media | (H1) | BSF | 884 | 1 491 | 3 651 | 82 832 | 3 455 | 78 978 |
| | | BSO | 23 686 | 62 795 | 6 857 | 223 796 | 5 602 | 123 629 |
| | | BSR | 12 236 | 38 421 | 2 695 | 58 175 | 2 511 | 51 070 |
| | | MED | 942 | 1 556 | 744 | 3 744 | 348 | 648 |
| | (H2) | BSF | 1 056 | 1 747 | 5 110 | 105 458 | 4 604 | 106 816 |
| | | BSO | 11 860 | 37 837 | 2 086 | 60 200 | 1 900 | 40 166 |
| | | BSR | 13 069 | 40 128 | 3 005 | 66 493 | 2 698 | 55 113 |
| | | MED | 683 | 1 204 | 508 | 2 478 | 252 | 475 |

# 6 Conclusion

In this paper, we applied regression trees as a valuable tool for analyzing cyber claims. For reproducibility purpose, all models have been fitted on a public database, the PRC database. Although this database, widely used in the literature, presents serious drawbacks and inconsistencies as we discussed it intensively throughout the paper, the methodology can be easily extended to other private databases, and several conclusions we draw can be generalized. The first observation is the heterogeneity of cyber events in terms of severity. This is, of course, a well known fact. However the regression tree approaches allow a clarification and a quantification of some characteristics that create this heterogeneity. For example, some sectors of activity (Healthcare, Education, Nonprofit organization) seem to have significantly lighter tail than the others (see Figure 4). Moreover, it appears that the central part of the distribution does not behave like the tail—in the sense that the impact of the covariates on this right tail does not seem to be identical

to what we can observe on the core of the distribution. Among the categories of targeted organizations which are associated with the lightest tail, one can observe that Healthcare and Education are mainly affected to the right-hand side of the tree describing the central part of the distribution (see Figure 5), meaning that the severity of claims striking them is, in average, higher. This shows the importance of a separate analysis of "typical" claims, and "extreme" ones. This dissemblance between what drives the center and what drives the tail of the distribution is not specific to cyber, but is probably reinforced by the various profiles of cyber criminals (home-made attacks versus larger scale criminal organizations). Finally, the results on our analysis based on GP trees reveal that there may be a significant operational impact if we pay attention to clustering types of "extreme" claims.

We want to emphasize this last point: our analysis tends to acknowledge that a classical peaks over threshold approach (that is ignoring the influence of covariates on the shape parameter) leads to considering the whole tail of the distribution as too heavy. On the other hand, identifying some clusters for extreme events could at least be interesting for designing appropriate risk management strategies for some type of claims. Our purpose is not to draw a clear line between which criterion should be used to exclude or not some type of claims from the perimeter of insurance contracts, our data are not accurate enough to elaborate precise recommendations. Nevertheless we strongly advocate for developing such regression approaches to better understand and manage extreme claims.

Regarding estimation of the frequency, the approach we took is very approximative due to the lack of consistency of data. Nevertheless, this analysis seemed to us essential in order to show how a whole insurance pricing and reserving methodology can be developed. Moreover, due to the relative novelty of the risk, the information gathered by insurance companies are sufficiently recent to take advantage on additional sources of (public) data. Hence we believe that a promising field of research is to find a proper way for companies to combine internal data and these external sources, provided that a rigorous statistical analysis has first identified and corrected their biases.

# 7 Appendix

## 7.1 Hill plot

Figure 6 shows the Hill plot for the number of records (see Resnick, 2007, pp 85–89 for more details on Hill plots). From the shape of the curve, we chose $u = 27\ 999$ (which

corresponds to a stabilization of the Hill plot) which leads to keep the 1 000 highest observations (around 16% of the total number of breaches).

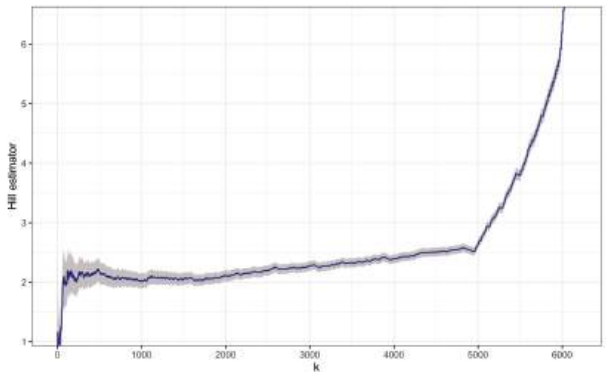



Figure 6: Hill plot for the number of records.

## 7.2 Goodness of fit for GP tree and comparison tests

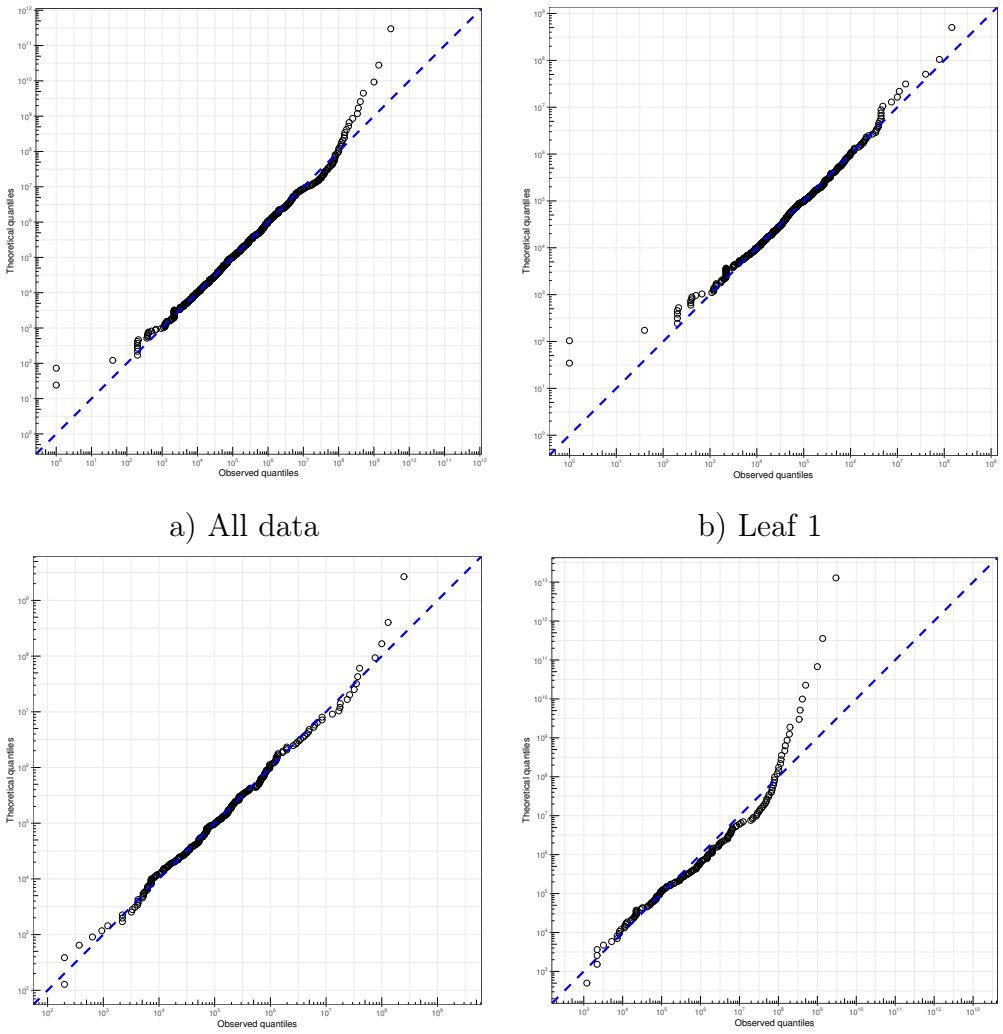
Figure 7 gathers quantile-quantile plots corresponding to each leaf of our final GP tree of Figure 4. After fitting the GP tree of Figure 4, we check that the three clusters can be considered dissimilar enough so that they can not be grouped into a single one (which would considerably simplify the study by performing standard extreme value analysis methodologies, i.e. without taking covariates into account). A Kolmogorov-Smirnov test (see Section 6.9 in Lehmann and Romano, 2006) has been used to compare the empirical distribution of each couple of leaves. The $p$-values are given in Table 11. They suggest a rejection of the null hypothesis. We also considered a likelihood ratio test (see Section 12.4.4 in Lehmann and Romano, 2006) which uses the particular structure of GP distribution. This consists in computing the difference between the log-likelihood obtained from the tree to the log-likelihood obtained when a single GP distribution is fitted to the whole set of observations. The value of this test statistic is 169.8, leading to a $p-$value lower than $2.2 \times 10^{-16}$, which once again suggests a significant improvement of the fit.

**Acknowledgement:** The authors acknowledge funding from the project *Cyber Risk Insurance: actuarial modeling*, Joint Research Initiative under the aegis of Risk Fundation, with partnership of AXA, AXA GRM, ENSAE and Sorbonne Université.

**R codes:** The code is made publicly available at `https://bitbucket.org/sebastien_farkas/cyber_claim_analysis_gpd_regression_trees/`

Table 11: Statistics and p-values of the two sample Kolmogorov Smirnov tests computing on samples of the leaves of the Generalized Pareto tree, two by two.

| Leaf of the first sample | Leaf of the second sample | KS statistic | KS p-value |
|---|---|---|---|
| 1 | 2 | 0.22 | $1.94 \times 10^{-8}$ |
| 1 | 3 | 0.44 | $< 2.2 \times 10^{-16}$ |
| 2 | 3 | 0.32 | $8.07 \times 10^{-11}$ |




a) All data

b) Leaf 1

c) Leaf 2

d) Leaf 3

Figure 7: Quantile quantile plots of Generalized Pareto distribution fits on all observations exceeding the threshold $u$ (Figure a)) and on samples of each of the 3 leaves of the Generalized Pareto tree of Figure 4 (Figures b), c) and d)).

# References


Balkema, A. A., and De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 792–804.

Beirlant, J., and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Computational statistics & data analysis*, *42*(4), 595–619.

Beirlant, J., and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, *89*(1), 97–118.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of Extremes: theory and applications*. John Wiley & Sons Ltd., Chischester.

Biener, C., Eling, M., and Wirfs, J. H. (2015). Insurability of Cyber Risk: An empirical analysis. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *40*(1), 131–158.

Bisogni, F., Asghari, H., and Van Eeten, M. J. (2017). Estimating the size of the iceberg from its tip: An investigation into unreported data breach notifications. In *Proceedings of 16th annual workshop on the economics of information security 2017*.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Advanced Books and Software, Belmont, CA.

Chaudhuri, P., and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, *8*(5), 561–576.

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2015). An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates. *Journal of Risk and Insurance*, *83*(3), 735–776.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag, London.

Databreaches.net. (n.d.). *Databreaches reporting.* (`https://www.databreaches.net/about/`)

Davison, A. C., and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(3), 393–425.

De'ath, G., and Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192.

De Haan, L., and Ferreira, A. (2007). *Extreme Value Theory: an introduction*. Springer Science & Business Media.

Edwards, B., Hofmeyr, S., and Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, *2*(1), 3–14.

Eling, M., and Loperfido, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, *75*, 126–136.

Eling, M., and Schnell, W. (2016). What do we know about cyber risk and cyber risk insurance? *Journal of Risk Finance*, *17*(5), 474–491.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events, volume 33 of applications of mathematics*. Springer-Verlag, Berlin.

European Insurance and Occupational Pensions Authority (EIOPA). (2019). Cyber risk for insurers - challenges and opportunities. Retrieved from `https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa_cyber_risk_for_insurers_sept2019.pdf`

Fahrenwaldt, M. A., Weber, S., and Weske, K. (2018). Pricing of cyber insurance contracts in a network model. *ASTIN Bulletin*, *48*(3), 1175–1218.

Gey, S., and Nédélec, E. (2005). Model Selection for CART Regression Trees. *IEEE Transactions on Information Theory*, *51*(2), 658–670.

González, C., Mira-McWilliams, J., and Juárez, I. (2015). Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission & Distribution*, *9*(11), 1120–1128.

Insua, D. R., Vieira, A. C., Rubio, J. A., Pieters, W., Labunets, K., and Rasines, D. G. (2019). An Adversarial Risk Analysis Framework for Cybersecurity. *Risk Analysis*.

Jacobs, J. (2014). *Analyzing Ponemon Cost of Data Breach.* Retrieved from `http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/`

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, *25*(8-12), 1287–1304.

Lehmann, E. L., and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, *82*(3), 329–348.

Lopez, O., Milhaud, X., and Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, *10*(2), 2685–2716.

Maillart, T., and Sornette, D. (2010). Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, *75*(3), 357–364.

Marotta, A., Martinelli, F., Nanni, S., Orlando, A., and Yautsiukhin, A. (2017). Cyber-insurance survey. *Computer Science Review*, *24*, 35–61.

Matthews, D. (2019). Report on the cybersecurity insurance and identity theft coverage supplement.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, *3*(1), 119–131.

Ponemon, L. (2018). Cost of a data breach study: global overview. *Benchmark research sponsored by IBM Security independently conducted by Ponemon Institute LLC*.

Privacy Rights Clearinghouse. (2019). Retrieved from `https://privacyrights.org/data-breaches`

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818.

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, *2*(2), 121–135.

State of California. (n.d.). *California list of Data Security Breaches.* (`https://oag.ca.gov/privacy/databreach/list`)

Su, X., Wang, M., and Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586–598.

Therneau, T., and Clinic, M. (2018). User written splitting functions for rpart.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

U.S. HHS department. (n.d.-a). Retrieved from `https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf`

U.S. HHS department. (n.d.-b). Retrieved from `https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html`

Wheatley, S., Maillart, T., and Sornette, D. (2016). The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, *89*(7).

# Article B

# Generalized Pareto Regression Trees for extreme events analysis

# Generalized Pareto Regression Trees for extreme event analysis

Sébastien Farkas[1], Antoine Heranval[1,2], Olivier Lopez[1], Maud Thomas[1]

**Abstract**

This papers derives finite sample results to assess the consistency of Generalized Pareto regression trees introduced by Farkas et al. [2021] as tools to perform extreme value regression for heavy-tailed distributions. This procedure allows the constitution of classes of observations with similar tail behaviors depending on the value of the covariates, based on a recursive partition of the sample and simple model selection rules. The results we provide are obtained from concentration inequalities, and are valid for a finite sample size. A misspecification bias that arises from the use of a "Peaks over Threshold" approach is also taken into account. Moreover, the derived properties legitimate the pruning strategies, that is the model selection rules, used to select a proper tree that achieves a compromise between simplicity and goodness-of-fit. The methodology is illustrated through a simulation study, and a real data application in insurance for natural disasters.

**Key words:** Extreme value theory; Regression trees; Concentration Inequalities; Generalized Pareto Distribution.

**Short title:** GP regression trees

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France.

E-mails: sebastien.farkas@sorbonne-universite.fr, antoine.heranval@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr, maud.thomas@sorbonne-universite.fr

[2] Mission Risques Naturels, 1 rue Jules Lefebvre, 75009 Paris, France

# 1 Introduction

Extreme value theory (EVT) is the branch of statistics which has been developed and broadly used to handle extreme events, such as extreme floods, heat wave episodes or extreme financial losses [Katz et al., 2002, Embrechts et al., 2013]. One of the key results behind the success of this approach was proved by Balkema and de Haan [1974], who established the ability of the Generalized Pareto (GP) family to approximate the tail of a distribution. This property allows the statistician to find information from the largest observations of a random sample to extrapolate the tail. This yields the so-called Peaks over Threshold (PoT) method introduced by Smith [1984] which consists in fitting a GP distribution to the excesses above some (high) suitably chosen threshold. In a regression framework, the parameters of this GP distribution depend on covariates reflecting the fact that different values of these covariates may result in a different tail behavior of the response variable [see e.g. Davison and Smith, 1990, Smith, 1989]. In this paper, we study the use of regression trees to perform GP regression on the excesses for heavy-tailed distributions. This ensemble method, introduced by Breiman et al. [1984], determines clusters of similar tail behaviors depending on the value of the covariates, based on a recursive partition of the sample and simple model selection rules. In the present work, we provide theoretical results and empirical evidence on the consistency of such a procedure and of these selection rules. The result we provide are based on concentration inequalities, in order to hold for finite sample sizes. The main difficulty stands in the misspecification of the model and on handling the fact that the distributions are heavy tailed.

Tail regression is a challenging task. Several papers have been interested in extreme quantile regression, Chernozhukov [2005] and, Wang et al. [2012] derive extreme quantile estimators assuming a linear form for the conditional quantile. Gardes and Stupfler [2019] and Velthoen et al. [2019] use conditional intermediate-level quantiles to extrapolate above the threshold and deduce estimators for extreme conditional quantiles. Another approach is to model the parameters of the GP distribution as functions of the covariates e.g. as local polynomials [Beirlant and Goegebeur, 2004] or as generalized additive models [Chavez-Demoulin et al., 2015, Youngman, 2019]. More and more approaches in extreme value regression use machine learning methods. Carreau and Vrac [2011] present a new class of stochastic downscaling models, the conditional mixture models (CMM) which builds on a neural network. CMM are mixture models whose parameters are functions of predictor variables. Rietsch et al. [2013] address the issue of the optimization of the spatial design of a network of existing weather stations by combining EVT with neural networks.

Very recently, Velthoen et al. [2021] proposed a gradient boosting procedure to estimate conditional GP distribution. Several works [Richards and Huser, 2022, Pasche and Engelke, 2022, Allouche et al., 2022] have proposed methodologies based on neural networks for extreme quantile regression. Finally, Gnecco et al. [2022] have developed a method for extreme quantile regression using random forests. Their extremal random forest estimates the parameter of a GP distribution conditionally on the predictor vector using local likelihood maximization. Finally, two works consider piece-wise stationary marginal and dependence model to estimate the meteorological and oceanographic variables [Ross et al., 2018, Barlow et al., 2023].

Regression trees, introduced by Breiman et al. [1984] along with the CART algorithm (for Classification And Regression Trees), are flexible tools to perform a regression and clustering task simultaneously, with the ability to deal with discrete and smooth covariates simultaneously. They have been used in various fields, including industry [González et al., 2015], geology [see e.g. Rodriguez-Galiano et al., 2015], ecology [see e.g. De'ath and Fabricius, 2000], claim reserving in insurance [Lopez et al., 2016]. Through the iterative splitting algorithm used in CART, nonlinearities are introduced in the way the distribution is modeled, while furnishing an intelligible interpretation of the final classification of response variables. The splitting criterion—used to iteratively separate observations into clusters with similar behaviors—depends on the type of problems one is considering. While the standard CART algorithm relies on mean-squared criterion to perform mean-regression, alternative loss functions have been considered as in [Chaudhuri and Loh, 2002] for quantile regression, or in [Su et al., 2004] who used a log-likelihood based loss. Loh [2011, 2014] provide detailed descriptions of regression trees procedures and a review of their variants. In this paper, building on the result of Balkema and de Haan [1974] , we use a GP log-likelihood loss, as in [Farkas et al., 2021], to perform extreme value regression.

The rest of the paper is organized as follows. In Section 2, we introduce notations and describe the GP regression tree algorithm. Section 3 lists the main results of this paper, that is deviation bounds for the regression tree estimator for finite sample size, and consistency of the "pruning" (that is model selection) strategy. Empirical results are gathered in Section 4, which provides a simulation study, and a real data analysis in natural disaster insurance. Detailed proofs of the technical results are shown in the Appendix.

# 2 Regression trees for extreme value analysis

This section describes the estimation method (GP regression trees) that we consider in this paper, and which has already been introduced by Farkas et al. [2021]. Some classical results in EVT are given in Section 2.1 to motivate the GP approximation. Regression trees adapted to this context are described in Section 2.3.

## 2.1 Extreme value theory and regression

Let us consider independent and identically distributed observations $Y_1, Y_2, \ldots$ with an unknown survival function $\overline{F}$ (that is $\overline{F}(y) = P(Y_1 > y)$). A natural way to define extreme events is to consider the values of $Y_i$ which have exceeded some high threshold $u$. The excesses above $u$ are then defined as the variables $Y_i - u$ given that $Y_i > u$. The asymptotic behavior of extreme events is characterized by the distribution of the excesses which is given by

$$\overline{F}_u(z) = P[Y_1 - u > z \mid Y_1 > u] = \frac{\overline{F}(u+z)}{\overline{F}(u)},\ z > 0\,.$$

Pickands [1975] showed that, if $\overline{F}$ satisfies the following property

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0},\ \forall y > 0, \tag{2.1}$$

with $\gamma_0 > 0$, then

$$\lim_{u\to\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}_{\sigma_0,\gamma_0}(z)| = 0 \tag{2.2}$$

for some $\sigma_0 > 0$ and $\overline{H}_{\sigma_0,\gamma_0}$ necessarily belongs to the Generalized Pareto (GP) distributions family which distribution function is of the form

$$\overline{H}_{\sigma_0,\gamma_0}(z) = \left(1 + \gamma_0 \frac{z}{\sigma_0}\right)^{-1/\gamma_0},\ z > 0,$$

where $\sigma_0 > 0$ is a scale parameter and $\gamma_0 > 0$ is a shape parameter, which reflects the heaviness of the tail distribution. Especially, if $\gamma_0 \in (0, 1)$, the expectation of $Y_1$ is finite whereas if $\gamma_0 \geq 1$ the expectation of $Y_1$ is infinite. More details on these results can be found in e.g. [Coles, 2001, Beirlant et al., 2004].

Note that in full generality, the shape parameter $\gamma_0 \in \mathbb{R}$. However, the applications we have in mind, such as in Section 4, concern natural catastrophes which fall into the domain of heavy-tailed distributions, that is distributions for which $\gamma_0 > 0$. We therefore

choose here to focus on the case $\gamma_0 > 0$. Besides, in this paper, we derive non-asymptotic results on the consistency of a procedure on the GP log-likelihood (see Section 3). The derivation of such results requires some smoothness on the GP log-likelihood, which is satisfied for $\gamma_0 > 0$, but not for all $\gamma_0 \in \mathbb{R}$.

The so-called Peaks over Threshold (PoT) method is widely used [see Davison and Smith, 1990, Coles, 2001]. It consists in choosing a high threshold $u$ and fitting a GP distribution on the excesses above that threshold $u$. The estimation of the parameters $\sigma_0$ and $\gamma_0$ may be done by maximizing the GP likelihood. The choice of the threshold $u$ can be understood as a compromise between bias and variance: the smaller the threshold, the less valid the asymptotic approximation, leading to bias; on the other hand, a too high threshold will generate few excesses to fit the model, leading to high variance. In practice, threshold selection is a challenging task. The existing methods for the choice of the threshold $u$ relies on graphical diagnostics or on computational approaches based on supplementary conditions (that depend on unknown parameters) on the underlying distribution function $F$ [see Scarrott and MacDonald, 2012] . However, it should be mention that some recent works model GP distribution upper tail (with $\gamma_0 > 0$) and the remaining of the full distribution in one step, which allows one to overcome the challenging issue of threshold selection [Tencaliec et al., 2020, Huang et al., 2019].

In the present paper, we consider a regression framework, that is, our goal is to estimate the impact of some random covariates $\mathbf{X}$ on the tail of the distribution of a response variable $Y$. The previous convergence result (2.2) holds, but for quantities $\sigma_0$, $\gamma_0$ and $u$ that may depend on $\mathbf{X}$. More precisely, this means that, if we assume that $\gamma_0(\mathbf{x}) > 0$ for all $\mathbf{x}$ (which is the assumption that we will make throughout this paper), then (2.1) becomes

$$\lim_{t\to\infty} \frac{\overline{F}(ty \mid \mathbf{x})}{\overline{F}(y \mid \mathbf{x})} = y^{-1/\gamma_0(\mathbf{x})}, \forall y > 0, \tag{2.3}$$

where $\overline{F}(y \mid \mathbf{x}) = \mathbb{P}(Y \geq y \mid \mathbf{X} = \mathbf{x})$ [see Beirlant et al., 2004, and references therein], and (2.2) becomes

$$\lim_{u(\mathbf{x})\to\infty} \sup_{z>0} |\overline{F}_{u(\mathbf{x})}(z \mid \mathbf{x}) - \overline{H}_{\sigma_{0u(\mathbf{x})}(\mathbf{x}),\gamma_0(\mathbf{x})}(z)| = 0. \tag{2.4}$$

where $\overline{F}_{u(\mathbf{x})}(z \mid \mathbf{x}) = P[Y - u(\mathbf{x}) > z \mid Y > u(\mathbf{x}), \mathbf{X} = \mathbf{x}]$.

Therefore, in this regression framework, the PoT approach consists now in the estimation of the function $\boldsymbol{\theta}_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{x}))^t$ (where $a^t$ denotes the transpose of a vector $a$).

## 2.2 Framework

We now suppose that we have observed $(Y_i, \mathbf{X}_i)_{1\leq i\leq n}$ a $n$-sample of $(Y, \mathbf{X})$, where $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)$ belongs to a compact set $\mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathbb{R}$. In the approach described thereafter, each covariate can be either discrete or smooth, and it is not necessary that they are all of the same nature. Recall that the PoT approach consists in considering observations such that $Y_i \geq u(\mathbf{X}_i)$.

In this paper, we will restrain ourselves to the case where the function $u(\mathbf{x}) = u$. To allow an adaptive choice of this parameter, our results hold uniformly for $u \in [u_{\min}, u_{\max}]$ (see Section 3), with $u_{\min}$ and $u_{\max}$ such that

1. $u_{\min}$ is defined as the $1 - k_n/n$ quantile of $F$, that is

$$\mathbb{P}(Y \geq u_{\min}) = \frac{k_n}{n},$$

   where $k_n$ be an intermediate sequence, that is $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$,

2. $u_{\max}$ is defined such that

$$\mathbb{P}(Y \geq u_{\max}) = \frac{u_0 k_n}{n},$$

   for some constant $u_0 \leq 1$.

   Note that $u_{\min}$ and $u_{\max}$ are functions of $n$.

Here, $k_n$ denote the average number (up to some constant) of observations on which the model is fitted. It is hence related to the rate of convergence of the procedure.

**Remark 1.** *Our results easily extend to the case where* $u(\mathbf{x}) = \sum_{j=1}^{m} u_j \mathbf{1}_{\mathbf{x}\in\mathcal{X}_j}$, *where* $(\mathcal{X}_j)_{1\leq j\leq m}$ *are subsets of the space of covariates. Another possible extension would be to assume that* $u(\mathbf{x}) = f(\beta, \mathbf{x})$ *for some parameter* $\beta$ *and* $f$ *a known function. Nevertheless, a choice of such a particular threshold function seems hard to justify. Hence, we restrain ourselves to the simplest case.*

In the next section, we introduce a regression tree approach adapted to both smooth and discrete covariates, and relying on few assumptions (since the estimated regression function $\boldsymbol{\theta}_0$ does not need to be smooth).

## 2.3 GPD regression trees

Regression trees are a convenient tool to capture heterogeneous behaviors in the data [see Breiman et al., 1984]. These models aim at constituting classes of observations which have a relatively similar behavior in terms of the response variable $Y$. These classes are defined by "rules", which affect an observation to one of these classes according to the values of its covariates $\mathbf{X}$. These rules are obtained from the data through the CART (Classification And Regression Tree) algorithm, and the non-linearity of the procedure allows for an adaptation to the estimation of large classes of regression functions.

Fitting regression trees relies on a so-called "growing phase", described in our context in Section 2.3.1, which corresponds to the determination of these splitting rules. Section 2.3.2 shows how an estimator of the regression function $\boldsymbol{\theta}_0$ can be deduced from such a tree. The "pruning step", which can be understood as a model selection procedure, is described in Section 2.3.3.

### 2.3.1 Growing step: construction of the maximal tree

The CART algorithm consists in determining iteratively a set of "rules" $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \to R_\ell(\mathbf{x})$ to split the data, aiming at optimizing some objective function $\boldsymbol{\theta}^*(\mathbf{x})$ (also referred to as splitting criterion). This function $\boldsymbol{\theta}^*(\mathbf{x})$ can be seen as the minimizer of a certain risk function over a class of target functions, that is

$$\boldsymbol{\theta}^*(\mathbf{x}) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}[\phi(Y, \boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}],$$

where $\boldsymbol{\Theta} \subset \mathbb{R}^d$ represents the space parameter and $\phi$ a loss function whose choice depends on the quantity to be estimated. For instance, if $\phi$ is the quadratic (absolute) loss, then $\boldsymbol{\theta}^*$ corresponds to the conditional mean (median) of $Y$ given $\mathbf{X}$. In our case, $\phi$ will be chosen as the negative GP likelihood, that is

$$\phi(z, \boldsymbol{\theta}) = \log(\sigma) + \left(\frac{1}{\gamma} + 1\right) \log\left(1 + \frac{\gamma z}{\sigma}\right), \quad z > 0$$

where $\boldsymbol{\theta} = (\sigma, \gamma)^t \in \boldsymbol{\Theta}$. Therefore, in our case,

$$\boldsymbol{\theta}^*(\mathbf{x}) \;=\; \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}[\phi(Y - u, \boldsymbol{\theta}) \mathbf{1}_{Y>u} \mid \mathbf{X} = \mathbf{x}].$$

Note that, here, the CART algorithm is applying only to the observations $Y_i$ such that $Y_i > u$.

A set of rules $(R_\ell)_\ell$ is a set of maps such that, for all $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, $R_\ell(\mathbf{x}) = 1$ or 0 depending on whether some conditions are satisfied by $\mathbf{x}$, with $R_\ell(\mathbf{x}) R_{\ell'}(\mathbf{x}) = 0$ for

$\ell \neq \ell'$ and $\sum_\ell R_\ell(\mathbf{x}) = 1$. In case of regression trees, these partitioning rules have a particular structure, since they can be written, for quantitative covariates (the case of $\mathbf{x}$ containing qualitative variables is described in Remark 2 below), as $R_\ell(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_1 \leq \mathbf{x} < \mathbf{x}_2}$ for some $\mathbf{x}_1 \in \mathcal{X}$ and $\mathbf{x}_2 \in \mathcal{X}$, with comparison symbols to be understood as component-wise comparisons. In other terms, if $d = 1$, rules can be identified as partitioning segments, if $d = 2$ they are rectangles (hyper-rectangles in the general case). The determination of these rules from one step to another can be represented as a binary tree, since each rule $R_\ell$ at step $k$ generates two rules $R_{\ell 1}$ and $R_{\ell 2}$ (with $R_{\ell 1}(\mathbf{x}) + R_{\ell 2}(\mathbf{x}) = 0$ if $R_\ell(\mathbf{x}) = 0$) at step $k + 1$. The algorithm can be described as follows:

**Step 1:** $R_1(\mathbf{X}_i) = 1$ for all $i = 1, \ldots, n$ and $n_1 = 1$ (corresponds to the root of the tree).

**Step k+1:** Let $(R_1, ...R_{n_k})$ denote the rules obtained at step $k$. For $\ell = 1, \ldots, n_k$,

- if all observations $i$ such that $R_\ell(\mathbf{X}_i) = 1$ have the same characteristics, then keep rule $\ell$ as it is no longer possible to split the data;
- else, rule $R_\ell$ is replaced by two new rules $R_{\ell 1}$ and $R_{\ell 2}$ determined in the following way: for each component $X^{(j)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, define the best threshold $x_{\ell\star}^{(j)}$ to split the data, such that

$$\begin{aligned} x_{\ell\star}^{(j)} &= \arg\min_{x^{(j)}} \left\{ \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j-}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i) \right. \\ &\qquad \left. + \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j+}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i) \right\}, \end{aligned}$$

where

$$\begin{cases} \widehat{\boldsymbol{\theta}}_{j-}(x^{(j)}, R_\ell) & = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i), \\ \widehat{\boldsymbol{\theta}}_{j+}(x^{(j)}, R_\ell) & = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i). \end{cases}$$

Then, select the best splitting component index :

$$\begin{aligned} j_\star &= \arg\min_j \left\{ \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j-}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \right. \\ &\qquad \left. + \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j+}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \right\} \end{aligned}$$

Define the two new rules: $R_{\ell 1}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j\star)} \leq x_{\ell\star}^{(j\star)}}$, and $R_{\ell 2}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j\star)} > x_{\ell\star}^{(j\star)}}$.

- Let $n_{k+1} = n_k + 2$ denote the new number of rules.

**Stopping rule:** Stop if $n_{k+1} = n_k$.

This algorithm has a binary tree structure. The list of rules $(R_\ell)$ are identified with the leaves of the tree at step $k$, and the number of leaves of the tree is increasing from step $k$ to step $k+1$. The algorithm stops when each leaf contains only one observation or when the observations in the same leaf have the same characteristics. The stopping rule can also be slightly modified to ensure that there is a minimal number of points of the original data in each leaf of the tree at each step.

**Remark 2.** *In this version of the CART algorithm, all covariates are smooth or boolean. For qualitative variables with more than two modalities, they must be transformed into binary variables, or the algorithm must be slightly modified so that the splitting step of each $R_\ell$ should be done by finding the best partition into two groups on the values of the modalities that minimizes the loss function. This can be done by ordering the modalities with respect to the average value—or the median value—of the response for observations associated with this modality.*

### 2.3.2 From the tree to the parameter estimation

From a given set of $K$ rules $\mathcal{R} = (R_\ell)_{\ell=1,\dots,K}$, let $\mathcal{T}_\ell = \{\mathbf{x} : R_\ell(\mathbf{x}) = 1\}$, the $\ell$–th leaf of the corresponding tree. The estimator $\widehat{\boldsymbol{\theta}}^K(\mathbf{x})$ associated with the set of leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ is obtained as

$$\widehat{\boldsymbol{\theta}}^K(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\boldsymbol{\theta}}^K(R_\ell) R_\ell(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\boldsymbol{\theta}}_\ell^K \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} = \sum_{\ell=1}^{K} \begin{pmatrix} \widehat{\sigma}_\ell^K \\ \widehat{\gamma}_\ell^K \end{pmatrix} \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

The tree is obtained when the previous algorithm stops is referred to as the maximal tree and denoted $\widehat{T}_{\max}(u)$ with the set of leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K_{\max}}$, where $K_{\max}$ denotes its number of leaves. It corresponds to a trivial estimator of the objective function $\boldsymbol{\theta}^*(\mathbf{x})$ since either the number of observations in a leaf is equal to one, or all observations in this leaf share the same characteristics $\mathbf{x}$. The procedure of the growing phase is summarized in Algorithm 1.

The pruning step, presented in the next section, consists in extracting from the maximal tree $\widehat{T}_{\max}(u)$ a subtree that achieves a compromise between simplicity and goodness-of-fit.

**Algorithm 1** Growing phase

**Input:** Observations $(Y_i, \mathbf{X}_i)_{i=1,\dots,n}$ such that $Y_i > u$

$n_1 \leftarrow 1$, $R_1(\mathbf{X}_i) \leftarrow 1 \ \forall i = 1, \dots, n$ ▷ Root of the tree

**for** $\ell = 1, \dots, n_k$ **do**

**if** All observations $i$ such that $R_\ell(\mathbf{X}_i) = 1$ have the same characteristics **then**

$R_\ell \leftarrow R_\ell$ ▷ Do not change $R_\ell$

**else**

**for** $j = 1, \dots, d$ **do**

**for** $x^{(j)} \in \mathbb{R}$ **do** ▷ via a grid search

$\boldsymbol{\theta}_{j-}(x^{(j)}, R_\ell) \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i)$

$\boldsymbol{\theta}_{j+}(x^{(j)}, R_\ell) \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i)$

$x_{\ell\star}^{(j)} \leftarrow \arg\min_{x^{(j)}} \{ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j-}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x^{(j)}} R_\ell(\mathbf{X}_i)$
$+ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j+}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i) \}$

**end for**

$j_\star \leftarrow \arg\min_j \{ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j-}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} \leq x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i)$
$+ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j+}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i^{(j)} > x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \}$

**end for**

$R_{\ell 1}(\mathbf{x}) \leftarrow R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} \leq x_{\ell\star}^{(j_\star)}}$

$R_{\ell 2}(\mathbf{x}) \leftarrow R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} > x_{\ell\star}^{(j_\star)}}$

$n_{k+1} \leftarrow n_k + 2$

**end if**

**end for**

**Output:** $K_{\max}$, $(R_\ell)_{\ell=1,\dots,K_{\max}}$, $(\widehat{\boldsymbol{\theta}}_\ell)_{\ell=1,\dots,K_{\max}}$

### 2.3.3 Selection of a subtree: pruning step

For the pruning step, a standard way to proceed is to use a penalized criterion to select the appropriate subtree of $\widehat{T}_{\max}(u)$ [see Breiman et al., 1984, Gey and Nedelec, 2005]. To determine this subtree, it is not necessary to compute all the subtrees of $\widehat{T}_{\max}(u)$. It is sufficient to determine, among all the subtrees with $K$ leaves for $K \leq K_{\max}$, the subtree $\widehat{T}_K(u)$ that minimizes the following criterion

$$\frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} + \lambda K, \tag{2.5}$$

where $\lambda > 0$ denotes a penalisation constant, that can be chosen using cross-validation [see e.g. Allen, 1974, Stone, 1974]. Recall that $k_n$ is the average number of observations such that $Y_i > u$, that is the number of observations on which the CART procedure is performed. Then, it only remains to determine the final tree among the obtained list of $K_{\max}$ admissible subtrees. The trees $\widehat{T}_K(u)$, $K = 1, \ldots, K_{\max}$, are easy to determine, since $\widehat{T}_K(u)$ is obtained by removing one leaf from the tree $\widehat{T}_{K+1}(u)$ [see Breiman et al., 1984, p.284–290].

The number of leaves of the selected tree is thus obtained as the minimizer of the penalised criterion (2.5), that is

$$\widehat{K}(u) = \arg \min_{K=1,\ldots,K_{\max}} \left\{ \frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} + \lambda K \right\},$$

and the selected tree is denoted by $\widehat{T}_K(u) = \widehat{T}_{\widehat{K}(u)}(u)$.

# 3 Main results

In this section, we show that the GP regression tree procedure defined in Section 2.3 is consistent. Notations and assumptions used throughout this section are listed in Section 3.1. We then state our first main results on the consistency of a fixed tree with $K$ leaves, by separating the stochastic part of the error (Section 3.2) from the misspecification part (Section 3.3) caused by the GP approximation. The consistency of the pruning methodology is studied in Section 3.4.

## 3.1 Notations and assumptions

In order to derive our consistency results, we need the following assumptions.

**Assumption 1.** *1.* $k_n = O(n^{a_1})$, *with* $a_1 > 0$

*2. The number of leaves* $K_{\max}$ *of the maximal tree* $\widehat{T}_{\max}(u)$ *is such that* $K_{\max} \leq \kappa k_n$ *with* $0 < \kappa \leq 1$

*3. The parameter space* $\boldsymbol{\Theta}$ *is compact, that is*

$$\boldsymbol{\Theta} = [\sigma_{\min}, \sigma_n] \times [\gamma_{\min}, \gamma_{\max}],$$

*where* $\gamma_{\min}, \gamma_{\max}, \sigma_{\min} > 0$ *and* $\sigma_n = O(n^{a_2})$ *with* $a_2 > 0$.

Consider a threshold $u \in [u_{\min}, u_{\max}]$ (defined in Section 2.2) and a tree $\widehat{T}_K(u)$. We denote $\widehat{\boldsymbol{\theta}}_\ell^K(u) = \left(\widehat{\sigma}_\ell^K(u), \widehat{\gamma}_\ell^K(u)\right)^t$ the estimated parameter in each leaf $\mathcal{T}_\ell$, that is, for $\ell = 1, \ldots, K$

$$\widehat{\boldsymbol{\theta}}_\ell^K(u) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{k_n} \sum_{i=1}^n \phi(Y_i - u, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} \right\}.$$

For each $\ell = 1, \ldots, K$, this estimator is expected to be close to $\boldsymbol{\theta}_\ell^{*K}(u) = \left(\sigma_\ell^{*K}(u), \gamma_\ell^{*K}(u)\right)^t$ defined by

$$\boldsymbol{\theta}_\ell^{*K}(u) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}\left[\phi(Y - u, \boldsymbol{\theta}) \mathbf{1}_{Y > u} \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}\right]. \tag{3.1}$$

However, this quantity is not exactly our target: ideally, we wish to estimate, for $\ell = 1, \ldots, K$,

$$\boldsymbol{\theta}_{0,\ell}^K(u) = (\sigma_{0,\ell}^K(u), \gamma_{0,\ell}^K),$$

such that

$$\lim_{t \to \infty} \sup_{z > 0} |\overline{F}_t(z \mid \mathcal{T}_\ell) - \overline{H}_{\sigma_{0,\ell}^K(t), \gamma_{0,\ell}^K}(z)| = 0,$$

where $\overline{F}_t(z \mid \mathcal{T}_\ell) = \mathbb{P}(Y - t \geq z \mid \mathbf{X} \in \mathcal{T}_\ell, Y \geq t)$.

Hence, $\widehat{T}_K(u)$ denotes the tree with leaves $(\mathcal{T}_\ell)_{\ell=1,\ldots,K}$ and with parameters $\widehat{\boldsymbol{\theta}}^K(u) = (\widehat{\boldsymbol{\theta}}_\ell^K(u))_{\ell=1,\ldots,K}$. Similarly, we denote by $T_K^*(u)$ (resp. $T_{0,K}(u)$) the tree with the same leaves as $\widehat{T}_K(u)$ but with parameters $\boldsymbol{\theta}^{*K}(u) = (\boldsymbol{\theta}_\ell^{*K}(u))_{\ell=1,\ldots,K}$ (resp. $\boldsymbol{\theta}_0^K(u) = (\boldsymbol{\theta}_{0,\ell}^K(u))_{\ell=1,\ldots,K}$).

For any sequence of parameters $\boldsymbol{\theta}^K = (\boldsymbol{\theta}_\ell^K)_{\ell=1,\ldots,K}$ of a tree with $K$ leaves $(\mathcal{T}_\ell)_{\ell=1,\ldots,K}$, we denote $\boldsymbol{\theta}^K(\mathbf{x})$ the regression function defined as the following step-wise function

$$\boldsymbol{\theta}^K(\mathbf{x}) = \sum_{\ell=1}^K \boldsymbol{\theta}_\ell^K \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}.$$

In the next section, we will also need some regularity assumptions on the negative log-likelihood $y \to \phi(y - u, \boldsymbol{\theta}) \mathbf{1}_{y > u}$.

**Assumption 2.** *Let $H^{\ell}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}$ the hessian of $y \to \phi(y-u,\boldsymbol{\theta})\mathbf{1}_{y>u}$, that is*

$$H^{\ell}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}(y-u) = \begin{pmatrix} \partial^2_\sigma \phi(y-u,\boldsymbol{\theta}_1) & \partial_\sigma\partial_\gamma \phi(y-u,\boldsymbol{\theta}_2) \\ \partial_\sigma\partial_\gamma \phi(y-u,\boldsymbol{\theta}_3) & \partial^2_\gamma \phi(y-u,\boldsymbol{\theta}_4) \end{pmatrix} \mathbf{1}_{y\geq u}.$$

*Assume that there exists a constant $\mathfrak{C}_1 > 0$ such that*

$$\inf_{\substack{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4\in\boldsymbol{\Theta} \\ a,b\in\mathbb{R}}} \inf_{\substack{\ell=1,\dots,K \\ u_{\min}\leq u\leq u_{\max}}} \left| \mathbb{E}\left[ H^{\ell}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}(Y-u) \begin{pmatrix} a \\ b \end{pmatrix} \mid \mathbf{X}\in\mathcal{T}_\ell \right] \right| \geq \mathfrak{C}_1 \|(a,b)\|_\infty,$$

*where $\|(a,b)^t\|_\infty = \max(|a|,|b|)$.*

**Remark 3.** *The condition on the infimum can be relaxed: Assumption 2 comes naturally in using a Taylor expansion. Hence, the infimum with respect of $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_4$ can be restricted to $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_3$ belonging to a small neighborhood of $\boldsymbol{\theta}_1$ (and not to the whole set $\boldsymbol{\theta}$).*

We will first focus on the difference $\widehat{T}_K(u)$ and $T^*_K(u)$ in Section 3.2, which is the stochastic part of the error. Section 3.3 concerns the difference between $T^*_K(u)$ and $T_{0,K}(u)$ (and ultimately the difference between the regression functions $\widehat{\boldsymbol{\theta}}^*(\mathbf{x})$ and $\boldsymbol{\theta}_0(\mathbf{x})$) that can be understood as a misspecification term, caused by the fact that the excesses above the threshold are not exactly GP distributed. Finally, the consistency of the pruning step is shown in Section 3.4.

## 3.2 Deviation bounds for our estimator

In this section, we study the consistency of a fitted tree $\widehat{T}_K(u)$ with $K$ leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$, a subtree of the maximal tree $\widehat{T}_{\max}(u)$. For this first result, $K$ is fixed. Selection results for $K$ are provided in Theorem 3 in Section 3.4. The leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ of $\widehat{T}_K(u)$ are supposed to be fixed sets, as it is classically assumed to derive consistency of regression trees, [see e.g. Chaudhuri, 2000, Chaudhuri and Loh, 2002]. The tree $\widehat{T}_K(u)$ is identified by its leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ and the list of parameter values $\widehat{\boldsymbol{\theta}}^K_\ell(u)$ associated with each leaf $\mathcal{T}_\ell$. Considering a leaf $\mathcal{T}_\ell$, $\widehat{\boldsymbol{\theta}}^K_\ell(u)$ should ideally be close to its limit value $\boldsymbol{\theta}^{*K}_\ell(u)$, as $n$ tends to $\infty$. Hence, we introduce the "oracle" tree $\widehat{T}^*_K(u)$ which is defined by the same subdivision $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ as $\widehat{T}_K(u)$ but differs via the value of the parameters in each leaf (which is taken as $\boldsymbol{\theta}^{*K}_\ell(u)$ for leaf $\ell$). We denote $\boldsymbol{\theta}^{*K}(\mathbf{x})$ the regression function associated with $\widehat{T}^*_K(u)$.

To compare $\widehat{T}_K(u)$ and $T^*_K(u)$, the first step is to define a distance between trees. Let us define for two trees $T$ and $T'$ associated with the regression functions $\boldsymbol{\theta}(\mathbf{x}) = (\sigma(\mathbf{x}),\gamma(\mathbf{x}))^t$

and $\boldsymbol{\theta}'(\mathbf{x}) = (\sigma'(\mathbf{x}), \gamma'(\mathbf{x}))^t$ respectively,

$$\|T - T'\|_2 = \left( \int \|\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}'(\mathbf{x})\|_\infty^2 \mathrm{d}P_{\mathbf{X}}(\mathbf{x}) \right)^{1/2},$$

where $P_{\mathbf{X}}$ denotes the distribution of the covariates $\mathbf{X}$,

$$\text{and } \|\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}'(\mathbf{x})\|_\infty = \max\left(|\sigma(\mathbf{x}) - \sigma'(\mathbf{x})|, |\gamma(\mathbf{x}) - \gamma'(\mathbf{x})|\right)$$

.

The main result of this section is a deviation bound for $\|\widehat{T}_K(u) - T_K^*(u)\|_2$, which is Theorem 1 below.

**Theorem 1.** *Under Assumptions 1 and 2, there exists $\rho_0 > 0$ such that for $\beta \geq 10/(\rho_0 a_1)$ and $t \geq c_1 K(\log k_n) k_n^{-1}$, with $c_1 > 0$,*

$$\begin{aligned} &\mathbb{P}\left( \sup_{u_{\min} \leq u \leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \geq t \right) \\ &\quad \leq \; 2\left( \exp\left( -\frac{\mathcal{C}_1 k_n t}{K\beta^2 (\log k_n)^2} \right) + \exp\left( -\frac{\mathcal{C}_2 k_n t^{1/2}}{K^{1/2}\beta \log k_n} \right) \right) + \frac{\mathcal{C}_3 K}{k_n^{5/2} t^{3/2}}, \end{aligned} \tag{3.2}$$

*where $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are positive constants.*

*Moreover,*

$$\mathbb{E}\left[ \sup_{u_{\min} \leq u \leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \right] \leq \mathcal{C}_4 \frac{K\beta^2 (\log k_n)^2}{k_n}. \tag{3.3}$$

The proof of Theorem 1 is postponed to the appendix section (Section A.3). The exponential terms on the right-hand side of (3.2) come from concentration inequalities proved by Einmahl et al. [2005], while the polynomially decreasing term is related to the fact that the log-likelihood is an unbounded quantity, but that can still controlled when considering its expectation.

As a by-product, we obtain (3.3) (by integration of the bound of (3.2)). From (3.3), one can see that the $L^2-$norm of the stochastic part of the error, that is

$$\mathbb{E}\left[ \sup_{u_{\min} \leq u \leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \right]^{1/2},$$

is proportional to $K^{1/2}$, and, as expected, increases with the complexity of the tree. On the other hand, the error decreases almost at rate $k_n^{1/2}$ (up to some logarithmic factor), which is the convergence rate of standard estimators used to estimate the parameters of a GP distribution in absence of covariates.

Let us note that we do not explicitly take into account the dimension $d$ of the covariate $\mathbf{X}$ in the result of Theorem 1, in order to simplify the notations, since we implicitly consider that it is relatively small compared to $K$ and the sample size $k_n$. However, it is possible to retrieve the contribution of the dimension through the results contained in the Appendix: it appears inside the covering numbers obtained in Lemma 10 and then can be tracked through all the proofs below.

## 3.3 Misspecification bias

For $\mathbf{X} = \mathbf{x}$, the ultimate goal is to estimate the parameter set $\boldsymbol{\theta}_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{x}))^t$, introduced in (2.2), by maximization of the GP likelihood, and from the fact that the true function $\boldsymbol{\theta}_0(\mathbf{x})$ is not necessarily piecewise constant as $\boldsymbol{\theta}^*(\mathbf{x})$. The difference between $\boldsymbol{\theta}_0(\mathbf{x})$ and $\boldsymbol{\theta}^*(\mathbf{x})$ can be understood as a misspecification term due to the fact that the observations above the threshold are not exactly distributed according to a GP distribution. This bias term can be controlled under second order conditions which are standard in Extreme Value Analysis [see e.g. Beirlant et al., 2004].

Indeed, recall that assuming that the underlying distribution $\overline{F}(\cdot \mid \mathbf{x})$ satisfies Condition (2.3) guarantees that asymptotically the associate excesses above the threshold $u$ are GP distributed. For finite samples, the excesses are thus not exactly GP distributed which introduces some bias term. In order to control this bias term, a second-order condition is needed, that is a condition to control the rate of convergence in Condition (2.3). There exist numerous ways to express this second-order condition. Here, we consider the same condition as Condition C.6 in [Beirlant and Goegebeur, 2004]. First, Condition (2.3) can be translated into

$$\overline{F}(y \mid \mathbf{x}) = y^{-1/\gamma_0(\mathbf{x})}\eta(y \mid \mathbf{x}) \,, \forall y > 0, \tag{3.4}$$

where $\eta$ is a slow-varying function, that is $\eta(ty \mid \mathbf{x})/\eta(t \mid \mathbf{x}) \to 1$ as $t \to \infty$, for all $y > 0$.

**Assumption 3.** *Assume that for all* $\mathbf{x}$*, there exist a constant* $c$ *and a function* $\psi$ *such that*

$$\eta(ty \mid \mathbf{x})/\eta(t \mid \mathbf{x}) = 1 + c\psi(t) \int_1^t v^{\rho-1} \mathrm{d}v + o(\psi(t))$$

*as* $t \to \infty$ *for each* $y > 0$ *with* $\psi(t) > 0$ *and* $\psi(t) \to 0$ *as* $t \to \infty$ *and* $\rho \leq 0$.

Let us note that we could also consider the case of $c$, $\psi$ and $\rho$ depending on $\mathbf{x}$, and then assume some uniform bound over $x$ of these quantities. We chose this more restrictive formulation to simplify the notations.

The next result guarantees that the bias term tends to 0 as $u \to \infty$.

**Proposition 2.** *Under Assumptions 2 and 3, there exist a constant $c$ and a function $\psi$ such that $\psi(u) > 0$, $\psi(u) \to 0$ as $u \to \infty$, and such that, for $\mathbf{X} = \mathbf{x}$,*

$$\|\boldsymbol{\theta}_0(\mathbf{x}) - \boldsymbol{\theta}^*(\mathbf{x})\|_\infty \;\leq\; \mathfrak{C}_2(u)\frac{k_n}{n}\left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right) ,$$

*where $\mathfrak{C}_2(u)$ is a constant depending on $u$, $\gamma_{\min}$ and $\gamma_{\max}$.*

## 3.4 Consistency of the pruning step

The previous results cover the case of a tree with a fixed number of leaves $K$. In practice, the question is to select the proper subtree of $\widehat{T}_{\max}(u)$, the maximal tree obtained once the previous step of the CART procedure has stopped, with some "optimal" number of leaves, which is the objective of the pruning step described in Section 2.3.3.

As seen in Theorem 1 Equation (3.3), the stochastic part of the error put to the square increases proportionally to $K$. This is closely related to the natural inflation of the log-likelihood (which is locally quadratic) when the number of leaves increases, justifying a penalty proportional to $K$, as in [Breiman et al., 1984, Gey and Nedelec, 2005]. The aim of Theorem 3 is to corroborate this choice.

Let $K^*(u)$ denote the optimal number of leaves, that is

$$K^*(u) = \arg\min_{K=1,\dots,K_{\max}} \mathbb{E}\left[\phi(Y - u, \boldsymbol{\theta}^{*K}(\mathbf{X}))\mathbf{1}_{Y>u}\right] .$$

In words, $T^*(u) = T^*_{K^*(u)}(u)$ is the subtree of $T^*_{\max}(u)$ that achieves the closest proximity to the objective function $\mathbf{x} \to \boldsymbol{\theta}^*(\mathbf{x})$ in the sense that it maximizes the expectation of the (pseudo)-log-likelihood.

Second of all, as explained in Section 2.3.3, the selected number of leaves is defined by

$$\widehat{K}(u) = \arg\min_{K=1,\dots,K_{\max}} \left\{ \frac{1}{k_n}\sum_{\ell=1}^{K}\sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i))\mathbf{1}_{Y_i>u}\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell} + \lambda K \right\} ,$$

and $\widehat{T}(u) = T_{\widehat{K}(u)}(u)$ the corresponding selected tree.

The following Theorem 3 shows that the pruning methodology selects a tree $\widehat{T}(u)$ which approximately achieves the same rate of convergence as $\widehat{T}_{K^*(u)}(u)$, even if $K^*(u)$ is unknown, provided that the penalty constant $\lambda$ belongs to some reasonable interval.

In Theorem 3, $\Delta L(T^*(u), T^*_K(u))$ denotes the expectation of the difference of the likelihoods associated with the trees $T^*(u)$ and $T^*_K(u)$ (for a formal definition see Section A.5).

**Theorem 3.** *Let* $\mathfrak{D} = \inf_u \inf_{K<K^*(u)} \Delta L(T^*(u), T_K^*(u))$ *and suppose that there exists a constant* $c_2 > 0$ *such that the penalization constant* $\lambda$ *satisfies*

$$c_2\{\log k_n\}^{1/2} k_n^{-1/2} \leq \lambda \leq \mathfrak{D} - 2c_2\{\log(k_n)\}^{1/2} k_n^{-1/2},$$

*assuming that the right-hand side is positive. Then, under Assumptions 1 and 2, for all* $u \in [u_{\min}, u_{\max}]$,

$$\mathbb{E}\left[\|\widehat{T}(u) - T^*(u)\|_2^2\right] \leq \frac{\mathcal{C}_5 K^*(u)(\log k_n)^2}{k_n},$$

*where* $\mathcal{C}_5$ *is a constant depending on* $T^*(u)$.

The proof is given in Section A.5.

# 4 Simulation study and real data analysis

This section is devoted to the illustration of the GP regression procedure on simulated data (Section 4.1) and on a real dataset (Section 4.2). We focus on the estimation of the tail index function, since it is the most informative parameter to describe the tail of the distribution.

For both the simulations and the real data application, we used the `R` package `rpart` package for the GP CART procedure. The function `rpart` allows to fix the tuning parameter `minbucket`, which represents the minimal number of observations allowed in each leaf, that is the stopping rule. This tuning parameter was set to 50 for the simulations and 20 for the real data applications.

## 4.1 Simulations

In this section, we assess the performance of the GP regression procedure on simulated data and compare it with the competing approach proposed by Chavez-Demoulin et al. [2015]. They propose a semi-parametric framework to separate the smooth covariates from the discrete ones. Smoothing splines are used to estimate non-parametrically the smooth part, while the influence of discrete covariates is captured by a parametric function. This framework relies on a stronger assumption on the shape of the function $\boldsymbol{\theta}_0$.

We now describe the two cases considered in the simulation framework and then discuss the experiments results.

**Step-wise case**

In this first case, we consider the following regression framework: $X$ is a one dimensional variable uniformly distributed on $[0,1]$, and the response variable $Y$, conditionally on $X = x$, is distributed according to a Burr distribution of parameters $(\sigma_0(x), \gamma_0(x))$ which survival function is given by

$$\overline{F}(y \mid x) = \frac{1}{1 + (y/\sigma_0(x))^{1/\gamma_0(x)}},$$

with $\sigma_0(x) > 0$ and $\gamma_0(x)$ for all $x$. Among the two parameters, our attention, in designing the settings, is essentially motivated by the analysis of the conditional shape $\gamma_0(x)$. The reason for this particular attention is the key role it plays in the distribution of the maximum of a sample (distributed according to the conditional distribution of $Y \mid X = x$). Note that $\overline{F}(\cdot \mid x)$ satisfies Property (2.3).

The function $\gamma_0$ is taken as

$$\gamma_0(x) = \begin{cases} 0.8 & \text{if } 0 \leq x < 0.3 \\ 0.4 & \text{if } 0.3 \leq x < 0.7 \\ 0.2 & \text{if } 0.7 \leq x \leq 1. \end{cases}$$

For some $x$, the value exceeds 0.5, which corresponds to the case where the conditional variance is not defined. This case is important for risk management: if the variable $Y$ corresponds to the loss associated to a given risk, the mean-variance paradigm traditionally used by risk managers does not hold.

We then consider two settings:

1. $\sigma_0(x) = 1 - \gamma_0(x)$. This guarantees that the mean of the GP distribution is constant.
2. $\sigma_0(x) = (2^{\gamma_0(x)} - 1)/\gamma_0(x)$, here the median of the GP distribution is constant.

**Smooth case**

In this second case, we consider that $\sigma_0$ is constant equal to 1 and that $\mathbf{X}$ is no longer an one-dimensional variable uniformly distributed on $[0,1]$ but we consider a two-dimensional variable $\mathbf{X} = (X^{(1)}, X^{(2)})$. The function $\gamma_0$ is then taken as, with $t \in [0,1]$ such as $\mathbf{x} = t\mathbf{x}^{(1)} + (1-t)\mathbf{x}^{(2)}$ :

$$\gamma_0(\mathbf{x}) = 1 + \frac{\tanh(10(\mathbf{x} - 1/4))}{4} + \frac{\tanh(10(\mathbf{x} - 3/4))}{4}.$$

We simulate 1 000 replications for different sizes of the observation sample ($n$ =1000, 2500, 5000, 10 000 and 25 000) according to the described framework for all the cases. For

each sample, we consider the excesses above the 0.90-empirical quantile, which corresponds to $k_n$ =100, 250, 500, 1 000 and 2 500. For each simulated sample, we compute the regression tree procedure (GP CART), and the method based on generalized additive model (GAM) proposed by Chavez-Demoulin et al. [2015]. Next, we compute $\int(\hat{\gamma}(\mathbf{x}) - \gamma_0(\mathbf{x}))^2 \mathrm{d}\mathbf{x}$ for each estimator and also $\int \hat{\sigma}(\mathbf{x}) - \sigma_0(\mathbf{x}))^2 \mathrm{d}\mathbf{x}$ for the step-wise case. The empirical mean squared error is then obtained by averaging these errors over the 1 000 replications. Results for $\gamma_0(\mathbf{x})$ are shown in Table 1 and for $\sigma_0(\mathbf{x})$ in Table 2 (note that in the smooth case, the function $\sigma_0(\mathbf{x})$ is constant).

Table 1: Empirical mean squared errors of $\gamma_0(\mathbf{x})$ for the GP regression tree procedure (GP CART), and the GAM model for different sample sizes for a) the step-wise case with the constant mean (setting 1), b) the step-wise case with the constant median (setting 2), and c) the smooth case.

| $k_n$ | 100 | 250 | 500 | 1 000 | 2 500 |
|---|---|---|---|---|---|
| GP CART | 0.210 | 0.210 | 0.190 | 0.102 | 0.037 |
| GAM | 0.222 | 0.232 | 0.220 | 0.216 | 0.176 |

a)

| $k_n$ | 100 | 250 | 500 | 1 000 | 2 500 |
|---|---|---|---|---|---|
| GP CART | 0.212 | 0.211 | 0.197 | 0.193 | 0.187 |
| GAM | 0.222 | 0.232 | 0.220 | 0.216 | 0.176 |

b)

| $k_n$ | 100 | 250 | 500 | 1 000 | 2 500 |
|---|---|---|---|---|---|
| GP CART | 0.182 | 0.151 | 0.120 | 0.092 | 0.075 |
| GAM | 0.312 | 0.247 | 0.176 | 0.132 | 0.084 |

c)

The boxplots of the quadratic errors $\int(\widehat{\gamma}(\mathbf{x}) - \gamma_0(\mathbf{x}))^2 \mathrm{d}\mathbf{x}$ are shown for both cases (i) step-wise function for the first setting in Figure 1, (ii) step-wise function for the second setting in Figure 2 and (iii) smooth function in Figure 3. The boxplots of the quadratic

Table 2: Empirical mean squared errors for $\sigma_0(\mathbf{x})$ for the GP regression tree procedure (GP CART), and the GAM model for different sample sizes for a) the step-wise case with the constant mean (setting 1) and b) the step-wise case with the constant median (setting 2).

| $k_n$ | 100 | 250 | 500 | 1 000 | 2 500 |
|---|---|---|---|---|---|
| GP CART | 0.154 | 0.158 | 0.226 | 0.357 | 0.580 |
| GAM | 0.114 | 0.134 | 41.904 | 12.810 | 1.139 |

a)

| $k_n$ | 100 | 250 | 500 | 1 000 | 2 500 |
|---|---|---|---|---|---|
| GP CART | 1.135 | 1.148 | 1.027 | 1.036 | 1.041 |
| GAM | 0.114 | 0.134 | 41.910 | 12.810 | 1.139 |

b)

errors $\int(\widehat{\sigma}(\mathbf{x}) - \sigma_0(\mathbf{x}))^2 \mathrm{d}\mathbf{x}$ are shown for both cases (i) stepwise function for the first setting in Figure 4 and (ii) stepwise function for the second setting in Figure 5.

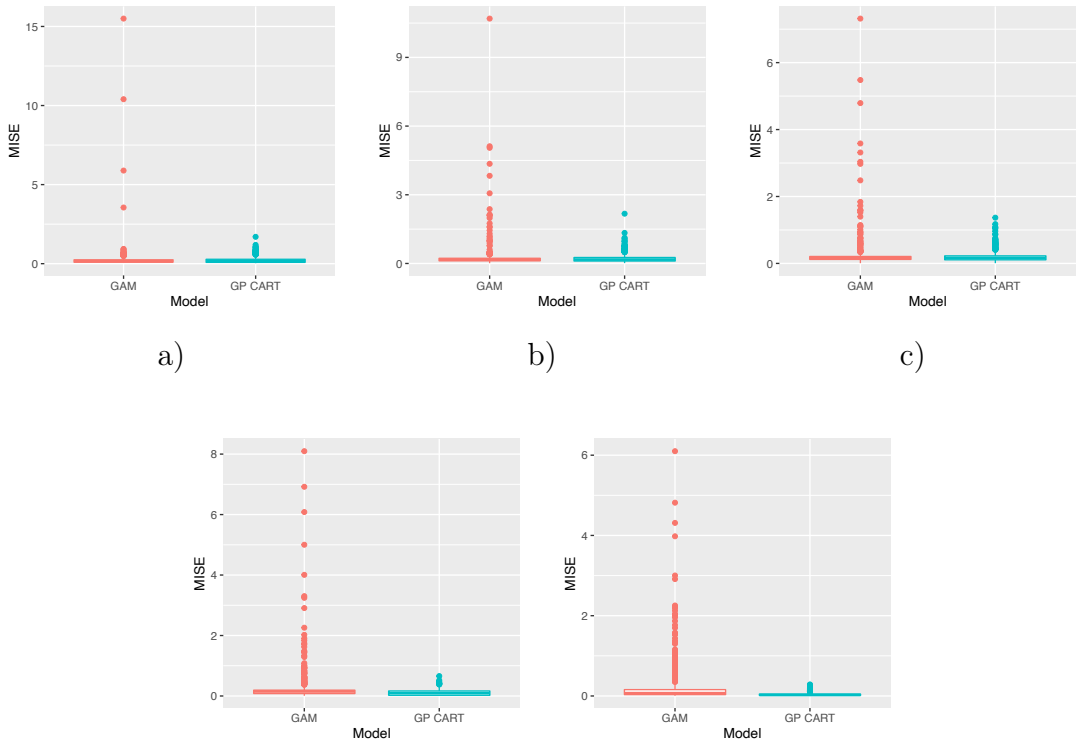



Figure 1: Boxplots of the quadratic errors for $\gamma_0(\mathbf{x})$ for each model in the step-wise case (setting 1) for a) 100 b) 250 c) 500 d) 1 000 and e) 2 500 excesses.

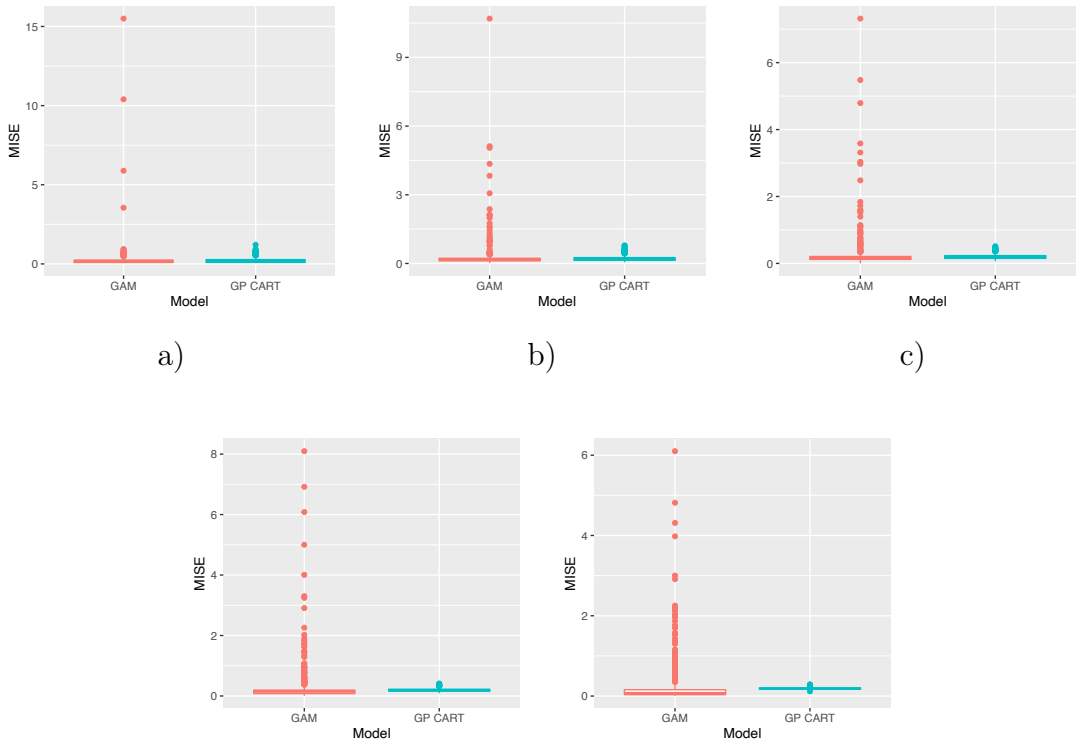



Figure 2: Boxplots of the quadratic errors for $\gamma_0(\mathbf{x})$ for each model in the step-wise case (setting 2) for a) 100 b) 250 c) 500 d) 1 000 and e) 2 500 excesses.

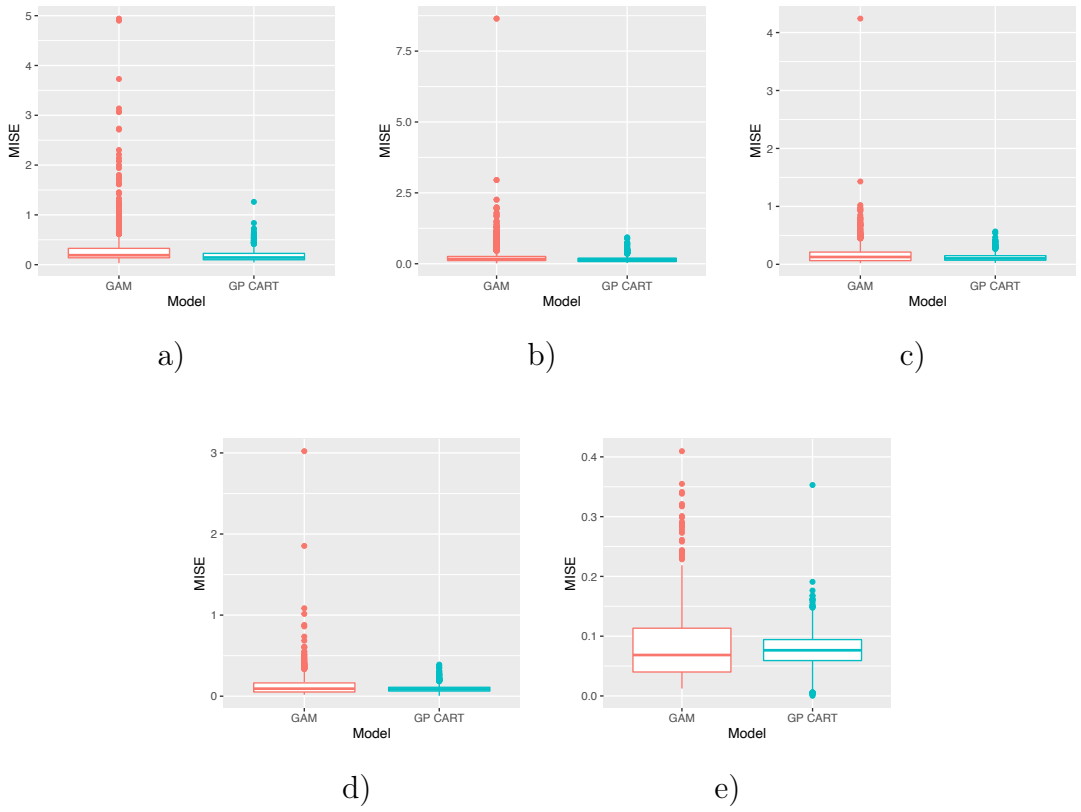



Figure 3: Boxplots of the quadratic errors for $\gamma_0(\mathbf{x})$ for each model in the smooth case for a) 100 b) 250 c) 500 d) 1 000 and e) 2 500 excesses.

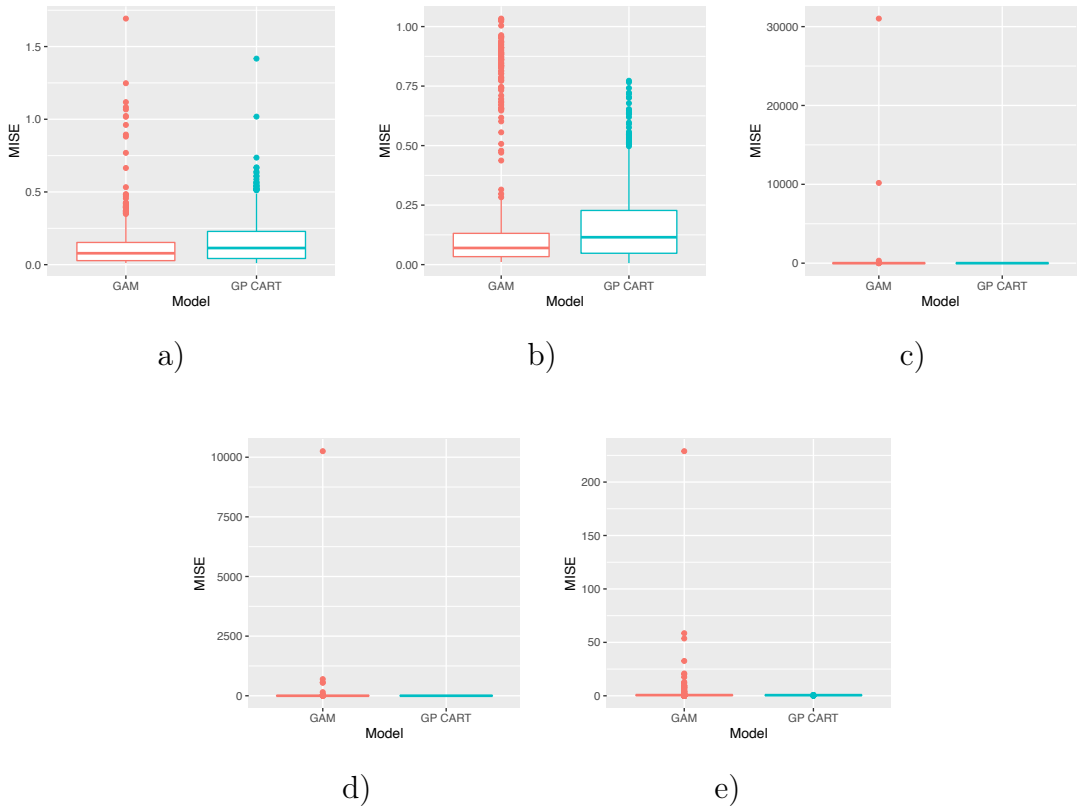



Figure 4: Boxplots of the quadratic errors for $\sigma_0(\mathbf{x})$ for each model in the step-wise case (setting 1) for a) 100 b) 250 c) 500 d) 1 000 and e) 2 500 excesses.

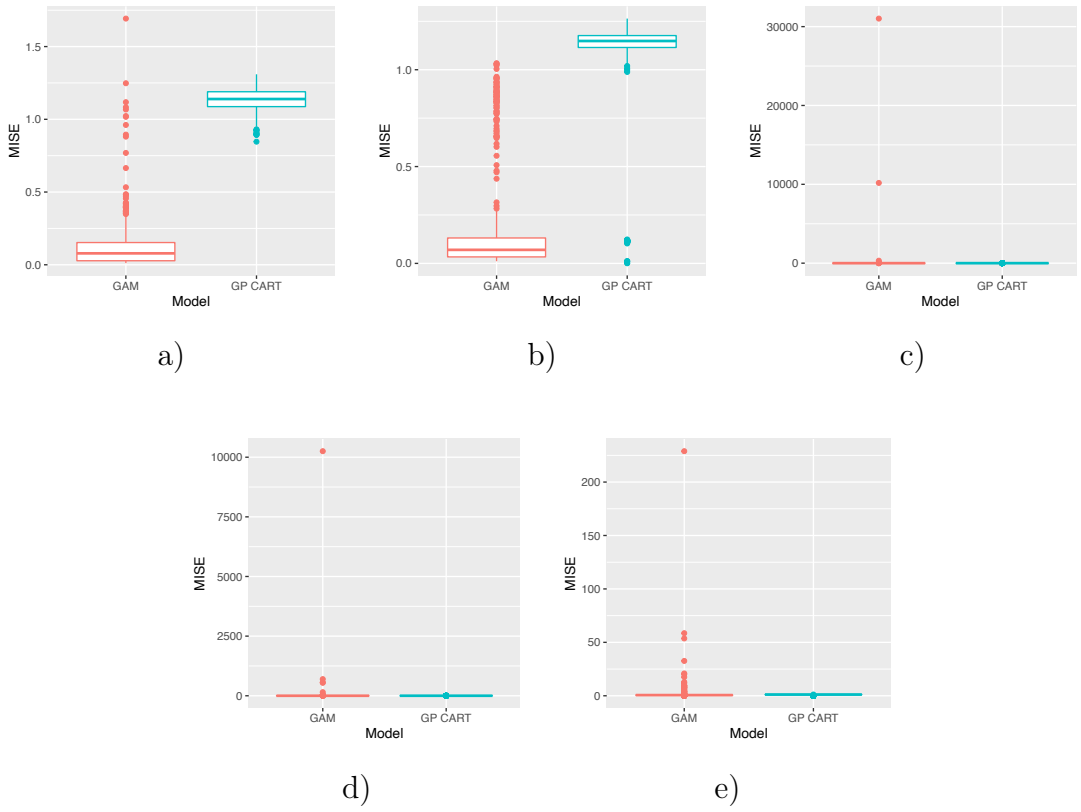



Figure 5: Boxplots of the quadratic errors for $\sigma_0(\mathbf{x})$ for each model in the step-wise case (setting 2)for a) 100 b) 250 c) 500 d) 1 000 and e) 2 500 excesses.

Concerning the performance of the estimation of the regression function $\gamma_0(\mathbf{x})$ in the step-wise cases, the GAM procedure and the GP CART procedure show similar results for $k_n$ small, and as $k_n$ gets larger the GP CART performs better than the GAM approach. Indeed, let us note that the GAM method is not designed to capture non-smooth functions like in the step-wise case. In Figures 1 and 2, the boxplots show that the quadratic errors of the GP CART procedure are more concentrated than the GAM procedure errors, that is GP CART is more stable than the GAM for the step-wise cases. Similar results can be observed for the estimation of $\sigma_0(\mathbf{x})$.

Regarding the performance of the regression function $\gamma_0(\mathbf{x})$ in the smooth case, the GAM and the GP CART procedures present similar results for small and larger sample sizes. It should be noted that again the GAM errors are generally less concentrated than the GP CART errors.

## 4.2 Prediction of the cost of flooding events in France

In order to improve the knowledge and the management of natural catastrophes, the French Federation of Insurance (FFA) is interested in the prediction of the cost of such events, especially of the most severe ones, shortly after their occurrence. These catastrophic events present some heterogeneity in their intensity depending on their characteristics, such as the affected meteorological region or the number of individual houses in flood risk area. The prediction of their cost thus becomes a challenging task. In this section, we illustrate how the GP regression tree procedure can be used to gain further insight in this heterogeneity. The ability of the procedure to design classes of events that are more homogeneous (in view of analyzing the tail of their distribution) is an appealing property in view of operation applications in insurance.

The database we consider was obtained through a partnership with the FFA, in particular with one of its dedicated technical body, the association of French insurance undertaking for natural risk knowledge and reduction (Mission Risques Naturels, MRN). It consists of all 3 100 flooding events that have been granted the status of natural catastrophe in France from 1999 to 2019 (let us note that the status "natural catastrophe" is a French specificity, with some legal consequences when an event receives this label [see Charpentier et al., 2021, MRN, 2016]).This database is fed by 13 contributors including the major French insurance companies, allowing this database to cover 70% of French non-life insurance market. The database gathers information regarding each flooding event (its cost, the meteorological region, the season, the number of affected hydrological regions, the number of individual houses and the number of professional business premises in flood-risk area). Note that, since the purpose of this database is the fast prediction of the cost of a flooding event (as soon as possible after its occurrence), the variables that are registered correspond to quantities that are available before the event, or soon after it.

The variable of interest, the total cost of a flooding event, is highly volatile. Indeed, it ranges between 0 and 394 376 000 euros with an empirical variance equal to $1.77e+14$. Figure 6 shows the average of the costs of the 10% most onerous flooding events within each meteorological region. This highlights the heterogeneity of the severity of the most severe events. Furthermore, the top ten most onerous events represent 43% of the total cost of this database and the top hundred 80%.

Now, let us recall that our goal is to understand the heterogeneity of the total cost of the most severe flooding events, that is of extreme flooding events. As explained in

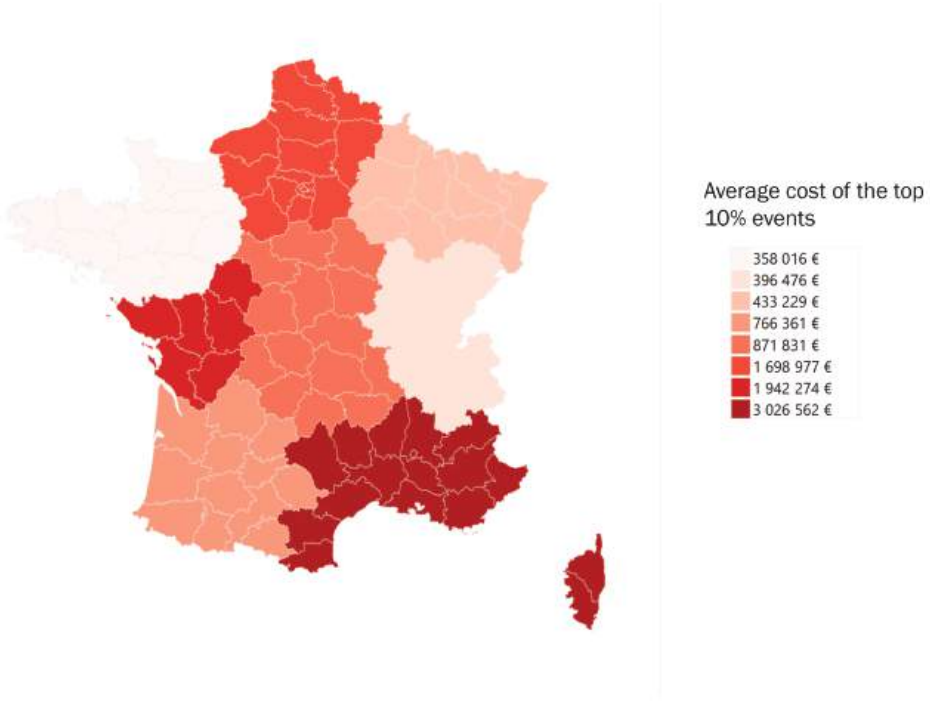



Figure 6: Cartography of the cost of flooding events in France from 1999 to 2019. For each meteorological region, the average of the costs of the 10% more onerous events is shown. The lighter red color suggesting a small cost while a darker color suggests a large cost.

Section 2.1, the definition of extreme events consists in choosing a threshold $u$, which should be chosen as a bias-variance trade-off. We chose a value of $u = 100\ 000$ based practical considerations and validated by sensitivity analyses (shown in the supplementary material, Section B). This yields 1 100 extreme events, that is for which the cost is larger than $u$.

The GP regression tree was performed on the database corresponding to the flooding events extracted from the original database for which the total cost is larger than $u$ (=100 000 euros). The variables of this database and their characteristics are summarized in Table 3. Again, it can be noticed that the cost, the variable of interest, is highly volatile.

The tree obtained from GP regression procedure is shown in Figure 7 (the quantile-quantile plots of the GP fit in each leaf are shown in the supplementary material, Section C). The tree is composed of 6 leaves, with three splits according to only 3 covariates: the number of individual houses, the number of professional business premises in flood-risk area and the number of affected meteorological regions. This seems reasonable since the

Table 3: List of quantitative and categorical variables in the database and their characteristics. For the quantitative variables, Table a) shows the minimum, the first quartile, the median, the mean, the third quartile and the maximum, and for the categorical variables, Table b) the number of observations per category.

| Variable | Min | 1st Q | Median | Mean | 3rd Q | Max |
|---|---|---|---|---|---|---|
| Cost (in euros) | 100 005 | 183 901 | 390 761 | 4 949 576 | 1 339 936 | 394 376 166 |
| Number of affected hydrological regions | 1 | 3 | 5 | 6.53 | 8 | 35 |
| Number of individual houses in flood risk area | 0 | 48 504 | 141 512 | 345 826 | 415 488 | 5 705 590 |
| Number of professional business premises in flood risk area | 0 | 17 525 | 54 921 | 168 950 | 185 772 | 2 431 039 |

a)

| Variable | Category | Number of observations |
|---|---|---|
| Meteorological regions | Center | 89 |
| | North West | 111 |
| | North | 166 |
| | North-East | 99 |
| | East | 135 |
| | South | 281 |
| | West | 49 |
| | South West | 158 |
| Seasons | Spring | 358 |
| | Summer | 336 |
| | Autumn | 251 |
| | Winter | 143 |

b)

first two covariates represent the exposure to floods, but also the population density of the affected area and the third one the extent of the flood. In each leaf, are given the

shape and scale parameters. The worst case scenario corresponds to the leaf on the far right, with a shape parameter equal to 1 and containing 9% of all flooding events. This leaf corresponds to events for which more than 9 meteorological regions are affected and more than 597 518 professional business premises are in flood-risk area. The least severe case corresponds to the third leaf from the left, with a shape parameter equal to 0.24 and containing only 3% of the events. Table 4 presents for each leaf the empirical median and mean of the costs and the theoretical median and mean of the corresponding GP distribution. Let us recall that for a GP distribution with a scale parameter $\sigma$ and a shape parameter $\gamma$, the theoretical median is given by $\sigma(2^{\gamma} - 1)/\gamma$ and the empirical mean by $\sigma/(1 - \gamma)$ for $\gamma < 1$ and $\infty$ for $\gamma \geq 1$. First of all, for every leaf, the median is much smaller than the mean suggesting that we are indeed dealing with extreme events. Then, the empirical and theoretical medians are of the same order for each leaf while the empirical and theoretical (when it exits) means are only comparable for the leaves 3 and 5 for which the shape parameter is significantly different from 1.

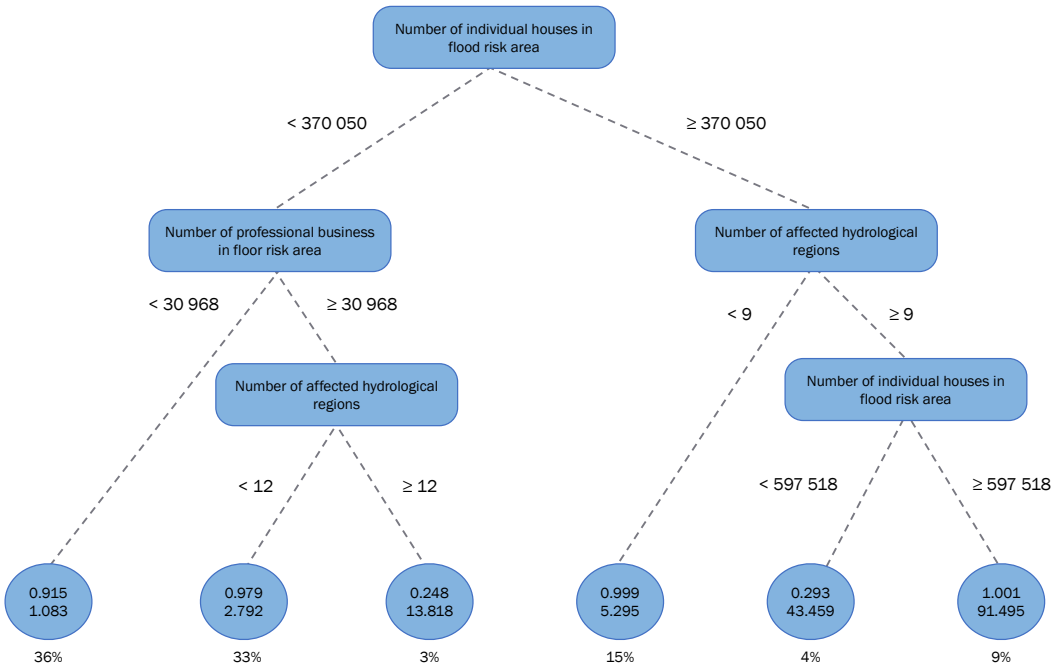



Figure 7: GP regression tree obtained for flooding events. For each leaf, the value of the shape parameter $\gamma$ (first line) and the scale parameter $\sigma$ at $10^{-5}$ (second line) are given. Percentage of observations affected to each leaf is mentioned.

| Leaf | Shape parameter | Empirical Median | Theoretical Median | Empirical Mean | Theoretical Mean |
|---|---|---|---|---|---|
| 1 | 0.915 | 207 044 | 104 793 | 711 740 | 1 366 968 |
| 2 | 0.979 | 364 513 | 276 879 | 1 325 493 | 13 168 585 |
| 3 | 0.248 | 900 945 | 1 045 203 | 1 929 512 | 1 938 357 |
| 4 | 0.999 | 578 437 | 529 377 | 3 868 125 | 807 158 756 |
| 5 | 0.293 | 2 974 918 | 3 339 911 | 6 086 955 | 6 245 812 |
| 6 | 1.001 | 9 980 686 | 9 152 030 | 37 335 807 | $\infty$ |

Table 4: Empirical median and mean, and theoretical median and mean for each leaf (in euros).

# 5 Conclusion

In this paper, we investigated the consistency of Generalized Pareto regression trees, applied to extreme value regression. The results that we derive are non-asymptotic, and allow to justify the consistency of the pruning methodology used to select a proper subtree. Let us note that the conditions under which our results hold are relatively weak, in the sense that they hold even if the tail index $\gamma$ is arbitrary close to zero (the special case $\gamma = 0$ is excluded) or large. Moreover, no regularity assumptions on the target parameters is required, due to the flexibility of the regression tree procedure.

Through the simulation study and the real data analysis, we investigated the practical performances of the methodology. The regression tree approach can be applied in various situations, and still provides interpretability of the results. On the other hand, regression trees may be unstable, since quite sensitive to some changes on the data that have been used to fit them. Hence, this work is a first step into the direction of studying other relied methodologies, like random forests [see for example Breiman et al., 1984] in this field of extreme value regression.

# A Proofs

In this Section, we present in details the proof of the results presented throughout the paper. Concentration inequalities required to obtain the results are presented in Section A.1. These inequalities are used to obtain deviation bounds in Section A.2, which are the key ingredients of the proof of Theorem 1 (Section A.3), Corollary 3.3 (Section A.2), and Theorem 3 (Section A.5). Section B shows some results on covering numbers that are required to control the complexity of some classes of functions considered in the proofs. Some technical lemmas are gathered in Section C.

## A.1 Concentration inequalities

The proofs of the main results are mostly based on concentration inequalities. The following inequality was proved initially Talagrand [1994], [see also Einmahl et al., 2005].

**Proposition 4.** *Let $(\mathbf{V}_i)_{1\leq i\leq n}$ denote i.i.d. replications of a random vector $\mathbf{V}$, and let $(\varepsilon_i)_{1\leq i\leq n}$ denote a vector of i.i.d. Rademacher variables (that is, $\mathbb{P}(\varepsilon_i=-1)=\mathbb{P}(\varepsilon_i=1)=1/2$) independent from $(\mathbf{V}_i)_{1\leq i\leq n}$. Let $\mathfrak{F}$ be a pointwise measurable class of functions bounded by a finite constant $M_0$. Then, for all $t$,*

$$\mathbb{P}\left(\sup_{\varphi\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\{\varphi(\mathbf{V}_i)-\mathbb{E}[\varphi(\mathbf{V})]\}\right\|_{\infty}>A_1\left\{E\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\right\|_{\infty}\right]+t\right\}\right)$$
$$\leq 2\left\{\exp\left(-\frac{A_2t^2}{nv_{\mathfrak{F}}}\right)+\exp\left(-\frac{A_2t}{M_0}\right)\right\},$$

*with $v_{\mathfrak{F}}=\sup_{\varphi\in\mathfrak{F}}\mathrm{Var}(\|\varphi(\mathbf{V})\|_{\infty})$, and where $A_1$ and $A_2$ are universal constants.*

The difficulty in using Proposition 4 comes from the need to control the symmetrized quantity $\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\|\right]$. Proposition 5 is due to Einmahl et al. [2005] and allows this control via some assumptions on the considered class of functions $\mathfrak{F}$.

We first need to introduce some notations regarding covering numbers of a class of functions. More details can be found for example in [van der Vaart, 1998, Chapter 2.6]. Let us consider a class of functions $\mathfrak{F}$ with envelope $\Phi$ (which means that for (almost) all $v$, $\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}$, $|f(v)|\leq\Phi(v)$). Then, for any probability measure $\mathbb{Q}$, introduce $N(\varepsilon,\mathfrak{F},\mathbb{Q})$ the minimum number of $L^2(\mathbb{Q})$ balls of radius $\varepsilon$ to cover the class $\mathfrak{F}$. Then, define

$$\mathcal{N}_{\Phi}(\varepsilon,\mathfrak{F})=\sup_{\mathbb{Q}:\mathbb{Q}(\Phi^2)<\infty}N(\varepsilon(\mathbb{Q}(\Phi^2)^{1/2}),\mathfrak{F},\mathbb{Q}).$$

**Proposition 5.** *Let $\mathfrak{F}$ be a point-wise measurable class of functions bounded by $M_0$ with envelope $\Phi$ such that, for some constants $A_3,\alpha\geq 1$, and $0\leq\sqrt{v}\leq M_0$, we have*

*(i)* $\mathcal{N}_{\Phi}(\varepsilon,\mathfrak{F})\leq A_3\varepsilon^{-\alpha}$, *for* $0<\varepsilon<1$,

*(ii)* $\sup_{\varphi\in\mathfrak{F}}\mathbb{E}\left[\varphi(\mathbf{V})^2\right]\leq v$,

*(iii)* $M_0\leq\frac{1}{4\alpha^{1/2}}\sqrt{nv/\log(A_4M_0/\sqrt{v})}$, *with* $A_4=\max(e,A_3^{1/\alpha})$.

*Then, for some absolute constant $A_5$,*

$$\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\right\|\right]\leq A_5\sqrt{\alpha nv\log(A_4M_0/\sqrt{v})}.$$

## A.2 Deviation results

We first introduce some notations that will be used throughout Sections A.2 to B. In the following, $\varphi_{\boldsymbol{\theta}}$ is a function indexed by $\boldsymbol{\theta} = (\sigma, \gamma)^t$ denoting either $\phi(\cdot, \boldsymbol{\theta})$, $\partial_\sigma \phi(\cdot, \boldsymbol{\theta})$, or $\partial_\gamma \phi(\cdot, \boldsymbol{\theta})$.

We consider in the following the class of functions $\mathfrak{F}$ defined as

$$\mathfrak{F} = \{ y \mapsto \varphi_{\boldsymbol{\theta}}(y-u)\mathbf{1}_{y \geq u}\mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ u \in [u_{\min}; u_{\max}], \ell = 1, ..., K \}. \tag{A.1}$$

By Lemma 11, the functions $y \mapsto \partial_\sigma \phi(y-u, \boldsymbol{\theta})$ and $y \mapsto \partial_\gamma \phi(y-u, \boldsymbol{\theta})$ are uniformly bounded (eventually up to some multiplication by a constant) by $\Phi(y) = \log(1+wy)$, where $w = \gamma_{\max}/\sigma_{\min}$. On the other hand, $y \mapsto \phi(y-u, \boldsymbol{\theta})$ is bounded by $\log \sigma_n + \Phi(y) = O(\log(k_n)) + \Phi(y)$.

Next, for $\ell = 1, \ldots, K$, and $\boldsymbol{\theta} = (\sigma, \gamma)^t \in \boldsymbol{\Theta}$, let

$$L_n^\ell(\boldsymbol{\theta}, u) = \frac{1}{k_n} \sum_{i=1}^{n} \phi(Y_i - u, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

be the (normalized) negative GP log-likelihood associated with the leaf $\ell$ of a tree $T_K(u)$ with set of $K$ leaves $(\mathcal{T}_\ell)_{\ell=1,\ldots,K}$. Let $L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}[L_n^\ell(\boldsymbol{\theta}, u)]$. The key results behind Theorems 1 and 3 relies on studying the deviations of the processes, indexed by $\boldsymbol{\theta}$, $u$ and $\ell$,

$$\begin{aligned} \mathcal{W}_0^\ell(\boldsymbol{\theta}, u) &= L_n^\ell(\boldsymbol{\theta}, u) - L^\ell(\boldsymbol{\theta}, u), \\ \mathcal{W}_1^\ell(\boldsymbol{\theta}, u) &= \nabla_{\boldsymbol{\theta}} L_n^\ell(\boldsymbol{\theta}, u) - \nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u). \end{aligned}$$

Let $M_n = \beta \log k_n \leq \beta a_1 \log(n)$ with $\beta > 0$ and $a_1 > 0$ (with $a_1$ defined in Assumption 1). We study the deviations of these processes by decomposing $\mathcal{W}_i^\ell(\boldsymbol{\theta}, u)$, for $i = 0, 1$, (which is a sum of i.i.d. observations) into two sums.

- the first one gathers observations smaller than some bound (more precisely, such that $\Phi(Y_i) \leq M_n$), which is considered in Theorem 6. Since these observations are bounded (even if this bound in fact depends on $n$ and can tend to infinity when $n$ grows), we can apply a concentration inequality such as the one of Section A.1. Let us stress that $\sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \|\varphi_{\boldsymbol{\theta}}(y) \mathbf{1}_{\Phi(y) \leq M_n}\|_\infty \leq M_n$;
- in the second one (Theorem 7), we consider the observations larger than this bound, and control them through the fact that the function $\Phi$ has finite exponential moments (see Lemma 11).

Corollary 8, which provides deviation bounds for estimation errors in the leaves of the tree, is then a direct consequence.

**Theorem 6.** *Let*

$$\underline{\mathcal{Z}}(M_n) = \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \left| \frac{1}{k_n} \sum_{i=1}^{n} \left( \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) \leq M_n} - \mathbb{E}\left[ \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) \leq M_n} \right] \right) \right|.$$

*If $k_n = O(n^{a_1})$ with $a_1 > 0$ (Assumption 1), then, for $t \geq \mathfrak{c}_1 (\log k_n)^{1/2} k_n^{-1/2}$,*

$$\mathbb{P}\left( \underline{\mathcal{Z}}(M_n) \geq t \right) \leq 2 \left( \exp\left( -\frac{C_1 k_n t^2}{\beta^2 (\log k_n)^2} \right) + \exp\left( -\frac{C_2 k_n t}{\beta \log k_n} \right) \right). \tag{A.2}$$

*Proof.* From Proposition 4,

$$\begin{aligned} &\mathbb{P}\left( \underline{\mathcal{Z}}(M_n) \geq A_1 \left\{ \mathbb{E}\left[ \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) \leq M_n} \varepsilon_i \right| \right] + t \right\} \right) \qquad \text{(A.3)} \\ \leq \;& 2 \left( \exp\left( -\frac{A_2 k_n^2 t^2}{n v_{\mathfrak{F}}} \right) + \exp\left( -\frac{A_2 k_n t}{M_n} \right) \right), \end{aligned}$$

with $v_{\mathfrak{F}} = \sup_{\varphi \in \mathfrak{F}} (|\varphi(Y)|)$. From Lemma 12, $v_{\mathfrak{F}} \leq M_n^2 k_n n^{-1}$, which shows that the first exponential term on the right-hand side of (A.3) is smaller than

$$\exp\left( -\frac{A_2 k_n t^2}{M_n^2} \right). \tag{A.4}$$

We can now apply Proposition 5 (combined with Lemma 10) to this class of functions with $v = M_n^2 k_n n^{-1}$ and $M_0 = M_n$. Hence,

$$\mathbb{E}\left[ \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) \leq M_n} \varepsilon_i \right| \right] \leq \frac{A_6}{k_n} \sqrt{n v \mathfrak{s}_n} = A_6 \frac{\mathfrak{s}_n^{1/2}}{k_n^{1/2}},$$

where $A_6' > 0$ and $\mathfrak{s}_n = \log(\sigma_n^{\alpha} K^{4(d+1)(d+2)} n / k_n)$ ($\alpha > 0$ being defined in Lemma 10). From Assumption 1, we see that $\mathfrak{s}_n = O(\log(k_n))$ (let us recall that $K$ is necessarily less than $n$). Whence, if $\mathfrak{c}_1 = 2 A_1 A_6'$, for $t \geq \mathfrak{c}_1 \left\{ \log(k_n) \right\}^{1/2} k_n^{-1/2}$,

$$\mathbb{P}\left( \underline{\mathcal{Z}}(M_n) \geq t \right) \leq \mathbb{P}\left( \underline{\mathcal{Z}}(M_n) \geq A_1 \left\{ \mathbb{E}\left[ \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) \leq M_n} \varepsilon_i \right| \right] + \frac{t}{2A_1} \right\} \right).$$

Equation (A.2) follows from (A.3) and (A.4) with $C_1 = A_2 A_1^{-2}/4$ and $C_2 = A_2 A_1^{-1}/2$. □

**Theorem 7.** *Let*

$$\overline{\mathcal{Z}}(M_n) = \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \left| \frac{1}{k_n} \sum_{i=1}^{n} \left( f(Y_i) \mathbf{1}_{\Phi(Y_i) > M_n} \right) - \mathbb{E}\left[ \varphi_{\boldsymbol{\theta}}(Y_i) \mathbf{1}_{\Phi(Y_i) > M_n} \right] \right|.$$

*If $k_n = O(a_1)$ with $a_1 > 0$ (Assumption 1), then there exists $\rho_0 > 0$ (Lemma 11) such that for $\beta a_1 \geq 10/\rho_0$, and $t \geq \mathfrak{e}_2 k_n^{-1/2}$,*

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geq t\right) \leq \frac{C_3}{k_n^{5/2} t^3}. \tag{A.5}$$

*Proof.* Let $\beta' = \beta a_2$. $\overline{\mathcal{Z}}(M_n)$ is upper-bounded by

$$\frac{1}{k_n} \sum_{i=1}^{n} \left\{ \Phi(Y_i) \mathbf{1}_{\Phi(Y_i) \geq M_n} \mathbf{1}_{Y_i \geq u_{\min}} + \mathbb{E}\left[ \Phi(Y) \mathbf{1}_{\Phi(Y) \geq M_n} \mathbf{1}_{Y \geq u_{\min}} \right] \right\}.$$

A bound for $E_{1,n} = \mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geq M_n}\mathbf{1}_{Y\geq u_{\min}}\right]$ is obtained from Lemma 13, and $nE_{1,n}/k_n \leq \mathfrak{e}_1 k_n^{-1/2}$ if $\beta' \geq 2/\rho_0$.

Next, from Markov inequality,

$$\begin{aligned} t^3 \mathbb{P}\left( \frac{1}{k_n} \sum_{i=1}^{n} \Phi(Y_i) \mathbf{1}_{\Phi(Y_i)\geq M_n} \mathbf{1}_{Y_i \geq u_{\min}} \geq t \right) \quad &\leq \quad \frac{nE_{3,n}}{k_n^3} + \frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} \\ &\quad + \frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3}. \end{aligned}$$

From Lemma 13, we get

$$\begin{aligned} \frac{nE_{3,n}}{k_n^3} &\leq \frac{\mathfrak{e}_3 n^{-(\rho_0\beta'/4 - 1/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} &\leq \frac{\mathfrak{e}_2\mathfrak{e}_1 n^{-(\rho_0\beta'/2 - 3/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3} &\leq \frac{\mathfrak{e}_1^3 n^{-(\rho_0\beta'/4 - 5/2)}}{k_n^{5/2}}. \end{aligned}$$

Each of these terms is bounded by $\max(\mathfrak{e}_3, \mathfrak{e}_2\mathfrak{e}_1, \mathfrak{e}_1^3) k_n^{-5/2}$ for $\beta' \geq 10/\rho_0$. Thus, for $t \geq 2\mathfrak{e}_1 k_n^{-1/2}$ and $\beta' \geq 10/\rho_0$,

$$\begin{aligned} &\mathbb{P}\left(\overline{\mathcal{Z}}_n \geq t\right) \\ &\leq \quad \mathbb{P}\left( \frac{1}{k_n} \sum_{i=1}^{n} \Phi(Y_i) \mathbf{1}_{\Phi(Y_i)\geq M_n} \mathbf{1}_{Y_i \geq u_{\min}} \geq \frac{t}{2} \right) + \mathbb{P}\left( \mathbb{E}\left[ \Phi(Y) \mathbf{1}_{\Phi(Y)\geq M_n} \mathbf{1}_{Y \geq u_{\min}} \right] \geq \frac{t}{2} \right) \\ &\leq \quad \frac{8 \max(\mathfrak{e}_3, \mathfrak{e}_2\mathfrak{c}_1, \mathfrak{e}_1^3)}{t^3 k_n^{5/2}} \end{aligned}$$

□

We now apply these results to deduce deviation bounds on the estimators $\widehat{\boldsymbol{\theta}}_\ell$ in the leaves of the tree.

**Corollary 8.** *Under the assumptions of Theorems 6 and 7 and Assumption 2, for* $t \geq \mathfrak{c}_3(\log k_n)^{1/2}k_n^{-1/2}$,

$$\mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty \geq t\right) \leq 2\left(\exp\left(-\frac{C_4 k_n t^2}{\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{C_5 k_n t}{\beta \log k_n}\right)\right) + \frac{C_6}{k_n^{5/2}t^3}.$$

*Proof.* For $1 \leq \ell \leq K$ and $u_{\min} \leq u \leq u_{\max}$, let $\boldsymbol{\theta} = (s,\gamma)^t$ and, for $\ell = 1,\ldots,K$, $\boldsymbol{\theta}_\ell^{*K}(u) = (s_\ell^{*K}(u), \gamma_\ell^{*K}(u))^t$, and let

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}\left[\begin{pmatrix} \partial_\sigma \phi(Y-u,\boldsymbol{\theta}) \\ \partial_\gamma \phi(Y-u,\boldsymbol{\theta}) \end{pmatrix} \mathbf{1}_{Y\geq u}\mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell}\right].$$

From Taylor series,

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}\left[H^\ell_{(\tilde{\sigma}_1,\gamma_1),(\sigma_1,\tilde{\gamma}_1),(\tilde{\sigma}_2,\gamma_2),(\sigma_2,\tilde{\gamma}_2)}(Y-u)\mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell}\right](\boldsymbol{\theta} - \boldsymbol{\theta}_\ell^{*K}(u))^t,$$

for some parameters $\tilde{\sigma}_j$ (resp. $\tilde{\gamma}_j$) between $\sigma$ and $\sigma_\ell^{*K}(u)$ (resp. $\gamma$ and $\gamma_\ell^{*K}(u)$). From Assumption 2, we get, for all $\ell = 1,\ldots,K$,

$$\frac{n}{k_n}\|\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u)\|_\infty \geq \mathfrak{C}_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty.$$

Hence, for all $\ell = 1,\ldots,K$,

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty \geq t\right) \leq \mathbb{P}\left(\frac{n}{k_n}\|\nabla_{\boldsymbol{\theta}} L^\ell(\widehat{\boldsymbol{\theta}}^K, u)\|_\infty \geq \mathfrak{C}_1 t\right).$$

Since for all $\ell = 1,\ldots,K$, $\nabla_{\boldsymbol{\theta}} L_n^\ell(\widehat{\boldsymbol{\theta}}^K) = 0$, $\mathcal{W}_1^\ell(\widehat{\boldsymbol{\theta}}^K(u), u) = -\frac{n}{k_n}\nabla_{\boldsymbol{\theta}} L^\ell(\widehat{\boldsymbol{\theta}}^K, u)$. Hence,

$$\mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_l^{*K}(u)\|_\infty \geq t\right) \leq \mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\mathcal{W}_1^\ell(\widehat{\boldsymbol{\theta}}^K(u), u)\|_\infty \geq \mathfrak{C}_1 t\right),$$

and the right-hand side is bounded by

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geq \frac{\mathfrak{C}_1 t}{2}\right) + \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq \frac{\mathfrak{C}_1 t}{2}\right).$$

The result follows from Theorem 6 and 7. □

## A.3 Proof of Theorem 1

The proof of the first part of Theorem 1 then consists in gathering the results on the leaves obtained in Corollary 8. Let $u_{\min} \leq u \leq u_{\max}$,

$$\|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \leq \sum_{\ell=1}^{K} \|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty^2 \leq K \sup_{\ell=1,\ldots,K} \|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty^2.$$

Hence

$$\begin{aligned}
&\mathbb{P}\left( \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \geq t \right) \\
&\leq \mathbb{P}\left( \sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K(u) - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty \geq t^{1/2} K^{-1/2} \right).
\end{aligned}$$

The results follows from Corollary 8, and from the assumption on $K \leq K_{\max} = O(k_n^3)$ (Assumption 1).

To prove the second part of Theorem 1, write

$$\mathbb{E}\left[ \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \right] = \int_0^\infty \mathbb{P}( \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \geq t) dt.$$

Let $t_n = c_1 K (\log k_n) k_n^{-1}$, then

$$\begin{aligned}
&\int_0^\infty \mathbb{P}( \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \geq t) dt \\
&\leq \; t_n + \int_{t_n}^\infty \mathbb{P}( \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \geq t) dt.
\end{aligned}$$

We now use Theorem 1 to bound the integral on the right-hand side. Since $\int_0^\infty \exp(-at)dt = \frac{1}{a}$, $\int_0^\infty \exp(-a^{1/2}t^{1/2})dt = \frac{2}{a}$, and $\int_1^\infty t^{-3/2}dt = 2$, we get

$$\begin{aligned}
\mathbb{E}\left[ \sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K(u) - T_K^*(u)\|_2^2 \right] &\leq t_n + \frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n} + \frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n} + \frac{2\mathcal{C}_3 K}{k_n^{5/2}} \\
&\leq \frac{c_1 K \log k_n}{k_n} + \frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n} \\
&\quad + \frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n} + \frac{2\mathcal{C}_3 K}{k_n^{5/2}} \\
&\leq \frac{\mathcal{C}_4 K (\log k_n)^2}{k_n}.
\end{aligned}$$

## A.4 Proof of Proposition 2

For all $\mathbf{x}$,

$$\|\boldsymbol{\theta}^*(\mathbf{x}) - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty = \|\sum_{\ell=1}^{K_{\max}} (\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_0(\mathbf{x}))\, \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\|_\infty \leq \sum_{\ell=1}^{K_{\max}} \|\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\,.$$

Now, from Taylor series, for $\ell = 1, \ldots, K$, conditionally on $\mathbf{X} \in \mathcal{T}_\ell$,

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}_0(\mathbf{X}), u) \;=\; \mathbb{E}\left[H^\ell_{(\tilde{\sigma}_1,\gamma_1),(\sigma_1,\tilde{\gamma}_1),(\tilde{\sigma}_2,\gamma_2),(\sigma_2,\tilde{\gamma}_2)}(Y-u) \mid \mathbf{X}\in\mathcal{T}_\ell\right](\boldsymbol{\theta}_0(\mathbf{X}) - \boldsymbol{\theta}_\ell^*)^t,$$

for some parameters $\tilde{\sigma}_j$ (resp. $\tilde{\gamma}_j$) between $\sigma_0(\mathbf{X})$ and $\sigma_\ell^{*K}(u)$ (resp. $\gamma_0(\mathbf{X})$ and $\gamma_\ell^{*K}(u)$).

Thus, under Assumption 2,

$$\begin{aligned}
&\|\boldsymbol{\theta}_0(\mathbf{X}) - \boldsymbol{\theta}_\ell^*\|_\infty \\
&\leq \frac{1}{\mathfrak{C}_1}\|\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}_0(\mathbf{X}), u)\|_\infty \\
&\leq \frac{1}{\mathfrak{C}_1}\frac{k_n}{n}\max\left(|\mathbb{E}\left[\partial_\sigma \phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]|, |\mathbb{E}\left[\partial_\gamma \phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]|\right)\,,
\end{aligned}$$

where $Z$ is a random variable distributed according to the distribution $F_u$ defined in Section 2.1 with $\sigma_0(\mathbf{X}) = u\gamma_0(\mathbf{X})$ and with

$$\begin{aligned}
\mathbb{E}\left[\partial_\sigma \phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= -\frac{1}{u\gamma_0(\mathbf{X})} + \frac{1}{u^2\gamma_0(\mathbf{X})}\left(1+\frac{1}{\gamma_0(\mathbf{X})}\right)\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right] \\
\mathbb{E}\left[\partial_\gamma \phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= -\frac{1}{\gamma_0(\mathbf{X})^2}\mathbb{E}\left[\log(1+Z/u)\mid \mathbf{X}\in\mathcal{T}_\ell\right] \\
&\quad + \frac{1}{u\gamma_0(\mathbf{x})}\left(1+\frac{1}{\gamma_0(\mathbf{X})}\right)\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right]\,.
\end{aligned}$$

Under Assumption 3, we have

$$\overline{F}_u(z) = \left(1+\frac{z}{u}\right)^{-1/\gamma_0(\mathbf{X})}\left\{1 + c\psi(u)\int_1^{1+z/u} v^{\rho-1}\mathrm{d}v + o(\psi(u))\right\}\,.$$

$$\begin{aligned}
\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= \int_0^u \overline{F}_u\left(\frac{t}{1-t/u}\right)\mathrm{d}t \\
&= \frac{u}{1+1/\gamma_0(\mathbf{X})}\left(1+\frac{c\psi(u)}{1+1/\gamma_0(\mathbf{X})-\rho} + o(\psi(u))\right) \\
&\leq u\left(1 + c\gamma_0(\mathbf{X})\psi(u) + o(\psi(u))\right)
\end{aligned}$$

and then

$$\begin{aligned}\mathbb{E}\left[\log(1+Z/u) \mid \mathbf{X}\in\mathcal{T}_\ell\right] &= \int_0^u \mathbb{P}\left[Z \geq u(\mathrm{e}^t-1) \mid \mathbf{X}\in\mathcal{T}_\ell\right] \mathrm{d}t \\ &= \gamma_0(\mathbf{X})\left(1+\frac{c\psi(u)}{1/\gamma_0(\mathbf{X})-\rho}+o(\psi(u))\right) \\ &\leq \gamma_0(\mathbf{X})\left(1+c\gamma_0(\mathbf{X})\psi(\mathbf{X})(u)+o(\psi(u))\right) .\end{aligned}$$

Consequently,

$$\left|\mathbb{E}\left[\partial_\sigma\phi(Z,\boldsymbol{\theta}_0(\mathbf{X})) \mid \mathbf{X}\in\mathcal{T}_\ell\right]\right| \leq \frac{1}{\gamma_{\min}}\left(1+\frac{1}{u}\left(1+\frac{1}{\gamma_{\min}}\right)\right)\left(1+c\gamma_0(\mathbf{X})\psi(u)+o(\psi(u))\right)$$

and

$$\left|\mathbb{E}\left[\partial_\gamma\phi(Z,\boldsymbol{\theta}_0(\mathbf{X})) \mid \mathbf{X}\in\mathcal{T}_\ell\right]\right| \leq \frac{1}{\gamma_{\min}}\left(1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)\left(1+c\gamma_0(\mathbf{X})\psi(u)+o(\psi(u))\right) .$$

Hence, conditionally on $\mathbf{X}\in\mathcal{T}_\ell$,

$$\|\boldsymbol{\theta}_0(\mathbf{X})-\boldsymbol{\theta}_\ell^*\|_\infty \leq \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right) ,$$

where $\mathfrak{C}_2(u)=\frac{1}{\mathfrak{c}_1}\frac{1}{\gamma_{\min}}\max\left(1+\frac{1}{u}+\frac{1}{u\gamma_{\min}}, 1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)$.

Finally, for all $\mathbf{x}$,

$$\begin{aligned}\|\boldsymbol{\theta}^*()-\boldsymbol{\theta}_0(\mathbf{x})\|_\infty &\leq \sum_{\ell=1}^{K_{\max}}\|\boldsymbol{\theta}_\ell^*-\boldsymbol{\theta}_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\ &\leq \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right)\sum_{\ell=1}^{K_{\max}}\mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\ &\leq \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right) .\end{aligned}$$

## A.5 Proof of Theorem 3

First, let us introduce some notations that are needed in the proof.

Define the log-likelihood $L_n(T_K,u)$ associated with a tree $T_K(u)$ with $K$ leaves $(\mathcal{T}_\ell)_{\ell=1,\ldots,K}$ and with parameters $\boldsymbol{\theta}(u)=\left(\boldsymbol{\theta}_\ell^K(u)\right)_{\ell=1,\ldots,K}$

$$L_n(T_K,u)=\sum_{\ell=1}^K L_n^\ell(\boldsymbol{\theta}_\ell^K,u)=\frac{1}{k_n}\sum_{\ell=1}^K\sum_{i=1}^n\phi(Y_i-u,\boldsymbol{\theta}_\ell^K)\mathbf{1}_{Y_i>u}\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell} ,$$

and $L(T_K,u)=\mathbb{E}[L_n(T_K,u)]$. Finally, for two trees $T$ and $T'$, $\Delta L_n(T,T')=L_n(T,u)-L_n(T',u)$ and similarly, $\Delta L(T,S)=L(T,u)-L(T',u)$.

The following lemma will be needed to prove Theorem 3.

**Lemma 9.** *Let* $\mathfrak{D} = \inf_u \inf_{K<K^*(u)} \Delta L(T^*(u), T_K^*(u))$ *and* $u \in [u_{\min}, u_{\max}]$ *fixed. Suppose that there exists a constant* $c_2 > 0$ *such that the penalization constant* $\lambda$ *satisfies*

$$c_2\{\log k_n\}^{1/2} k_n^{-1/2} \leq \lambda \leq (\mathfrak{D} - 2c_2\{\log(k_n)\}^{1/2} k_n^{-1/2}) k_n^{-1},$$

*then, under Assumptions 1 and 2, for* $K > K^*(u)$,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leq 2\left(\exp\left(-\frac{C_1 k_n \lambda^2 (K - K^*(u))^2}{\beta^2 (\log k_n)^2}\right) + \exp\left(-\frac{C_2 k_n \lambda (K - K^*(u)))}{\beta \log k_n}\right)\right) \\ &\quad + \frac{C_3}{k_n^{5/2} \lambda^3 (K - K^*(u))^3},\end{aligned}$$

*and, for* $K < K^*(u)$,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leq 4\exp\left(-\frac{C_1 k_n \{\mathfrak{D} - \lambda(K^*(u) - K)\}^2}{\beta^2 (\log k_n)^2}\right) \\ &\quad + 4\exp\left(-\frac{C_2 k_n \{\mathfrak{D} - \lambda(K^*(u) - K)\}}{\beta \log k_n}\right) \\ &\quad + \frac{2C_3}{k_n^{5/2} \{\mathfrak{D} - \lambda(K^*(u) - K)\}^3}.\end{aligned}$$

*Proof.* Let $u \in [u_{\min}, u_{\max}]$ fixed. If $\widehat{K}(u) = K$, this means that

$$\Delta L_n(T_K(u), T_{K^*(u)}(u)) := L_n(T_K, u) - L_n(T_{K^*(u)}, u) > \lambda(K - K^*(u)).$$

Decompose

$$\begin{aligned}\Delta L_n(T_K(u), T_{K_0}(u)) &= \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\} \\ &\quad + \{L_n(T^*, u) - L_n(T_{K^*(u)}, u)\}.\end{aligned}$$

Since $L_n(T^*, u) - L_n(T_{K^*(u)}, u) < 0$,

$$\Delta L_n(T_K(u), T_{K^*(u)}(u)) \leq \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\}.$$

For $K > K^*(u)$, $T_K^*(u) = T^*(u)$, hence,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leq \mathbb{P}\left(\Delta L_n(T_K(u), T_K^*(u)) > \lambda(K - K^*(u))\right) \\ &\leq \mathbb{P}\left(|\Delta L_n(T_K(u), T_K^*(u)) - \Delta L(T_K(u), T_K^*(u))| > \lambda(K - K^*(u))\right).\end{aligned}$$

For $K > K^*(u)$, a bound is then obtained from Theorems 6 and 7 if $\lambda(K - K^*(u)) \geq c_1\{\log(k_n)\}^{1/2} k_n^{-1/2}$, that is $\lambda \geq c_1\{\log k_n\}^{1/2} k_n^{-1/2}$.

Now, for $K < K^*(u)$,

$$\begin{aligned}\Delta L_n(T_K^*(u), T^*(u)) &\leq |\Delta L_n(T_K^*(u), T^*(u)) - \Delta L(T_K^*(u), T^*(u))| + \Delta L(T_K^*(u), T^*(u)) \\ &\leq |\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| - \mathfrak{D}(K^*(u), K).\end{aligned}$$

where $\mathfrak{D} = \inf_{K<K^*(u), u\in[u_{\min}, u_{\max}]} \mathfrak{D}(K^*(u), K)$, Hence,

$$\begin{aligned}&\mathbb{P}(\widehat{K}(u) = K) \\ &\leq \mathbb{P}\left(\Delta L_n(T_K(u), T_K^*(u)) \geq \frac{\mathfrak{D} - \lambda(K^*(u) - K)}{2}\right) \\ &\quad + \mathbb{P}\left(|\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| \geq \frac{\mathfrak{D} - \lambda(K^*(u) - K)}{2}\right) \\ &\leq \mathbb{P}\left(|\Delta L_n(T_K(u), T_K^*(u)) - \Delta L(T_K(u), T_K^*(u))| \geq \frac{\mathfrak{D} - \lambda(K^*(u) - K)}{2}\right) \\ &\quad + \mathbb{P}\left(|\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| \geq \frac{\mathfrak{D} - \lambda(K^*(u) - K)}{2}\right).\end{aligned}$$

These two probabilities can be bounded using Theorems 6 and 7 provided that, for all $K < K^*(u)$,

$$\frac{\mathfrak{D} - \lambda(K^*(u) - K)}{2} \geq \mathfrak{c}_1\{\log(k_n)\}^{1/2} k_n^{-1/2},$$

that is,

$$\lambda \leq \mathfrak{D} - 2\mathfrak{c}_1\{\log(k_n)\}^{1/2} k_n^{-1/2}.$$

□

We are now ready to prove Theorem 3. Let $u \in [u_{\min}, u_{\max}]$ fixed.

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{T}(u) - T^*(u)\|_2^2\right] &= \sum_{K=1}^{K_{\max}} \mathbb{E}\left[\|T_K(u) - T^*(u)\|_2^2 \mathbf{1}_{\widehat{K}(u)=K}\right] \\
&\leq \mathbb{E}\left[\|T_{K^*(u)}(u) - T^*(u)\|_2^2\right] + \sum_{K=1, K\neq K^*(u)}^{K_{\max}} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + \sum_{K=1, K\neq K^*(u)}^{K_{\max}} \mathbb{E}\left[\|T_K(u) - T^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K} \mathbf{1}_{\widehat{K}(u)=K}\right] \\
&\leq \mathbb{E}\left[\|T_{K^*(u)}(u) - T^*(u)\|_2^2\right] + \sum_{K=1}^{K^*(u)-1} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + \sum_{K=K^*(u)+1}^{K_{\max}} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + 2\sum_{K=1, K\neq K^*(u)}^{K_{\max}} \mathbb{E}\left[\|T_K(u) - T_K^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K}\right] \\
&\quad + 2\sum_{K=1, K\neq K^*(u)}^{K_{\max}} \mathbb{P}(\widehat{K}(u) = K)\|T^*(u) - T_K^*(u)\|_2^2.
\end{aligned}$$

Firstly, from Theorem 1,

$$\begin{aligned}
&\mathbb{E}\left[\|T_K(u) - T_K^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K}\right] \\
&= K\mathbb{P}\left(\|T_K(u) - T_K^*(u)\|_2^2 > K\right) + \int_K^\infty \mathbb{P}\left(\|T_K(u) - T_K^*(u)\|_2^2 > t\right) \mathrm{d}t \\
&\leq 2K\left(1 + \frac{\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n}\right) \exp\left(-\frac{\mathcal{C}_1 k_n}{\beta^2(\log k_n)^2}\right) \\
&\quad + 2K\left(1 + \frac{2\beta(\log k_n)}{\mathcal{C}_2 k_n} + \frac{2\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n^2}\right) \exp\left(-\frac{\mathcal{C}_2 k_n}{\beta(\log k_n)}\right) + \frac{2\mathcal{C}_3 K^{1/2}}{k_n^{5/2}}\,.
\end{aligned}$$

Secondly, recall that

$$\|T_K^*(u) - T^*(u)\|_2^2 = \int \|\boldsymbol{\theta}^{*K}(\mathbf{x}) - \boldsymbol{\theta}^*(\mathbf{x})\|_\infty^2 \mathrm{d}P_{\mathbf{X}}(\mathbf{x}) \leq K_{\max} \sum_{\ell=1}^{K_{\max}} \mu(\mathcal{T}_\ell)\|\boldsymbol{\theta}_\ell^{*K} - \boldsymbol{\theta}_\ell^*\|_\infty^2\,,$$

where $\mu(\mathcal{T}_\ell) = \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell)$. Following the same idea as in the proof of Proposition 2, from Taylor series , under Assumptions 2 and 3,

$$\|\boldsymbol{\theta}_\ell^{*K} - \boldsymbol{\theta}_\ell^*\|_\infty^2 \leq \mathfrak{C}_2^2(u)\frac{k_n^2}{n^2}\left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right)^2\,.$$

Hence,

$$\begin{aligned}\|T_K^*(u) - T^*(u)\|_2^2 &\leq \mathfrak{C}_2^2(u)\frac{k_n^2}{n^2}(1 + c\gamma_{\max}\psi(u) + o(\psi(u)))^2 \sum_{\ell=1}^{K_{\max}} \mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell} \\ &\leq \mathfrak{C}_3(u)\frac{k_n^2}{n^2}\,.\end{aligned}$$

Finally,

$$\mathbb{E}\left[\|\widehat{T}(u) - T^*(u)\|_2^2\right] \leq \frac{\mathcal{C}_5 K^*(u)(\log k_n)^2}{k_n},$$

for some constant $\mathcal{C}_5$. .

# B Covering numbers

**Lemma 10.** *Following the notations of the proof of Theorem 6, the class of functions* $\mathfrak{F}$ *satisfies*

$$\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) \leq \frac{\mathfrak{C}_4 K^{4(d+1)(d+2)} \|\Phi\|_2^{\alpha_1} \sigma_n^\alpha}{\varepsilon^\alpha},$$

*for some constants* $\mathfrak{C}_4 > 0$ *and* $\alpha > 0$ *(not depending on* $n$ *nor* $K$*).*

*Proof.* Let

$$\begin{aligned} g_{\boldsymbol{\theta}}(z) &= -\frac{1}{\sigma} + \left(\frac{1}{\gamma} + 1\right)\frac{\gamma z}{\sigma^2(1 + \frac{z\gamma}{\sigma})}, \\ h_{\boldsymbol{\theta}}(z) &= -\frac{1}{\gamma^2}\log\left(1 + \frac{z\gamma}{\sigma}\right) + \frac{\left(\frac{1}{\gamma} + 1\right) z}{\sigma + z\gamma}, \end{aligned}$$

for $z > 0$. For $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in $\mathcal{S} \times \Gamma$, we have (from a straightforward Taylor expansion),

$$|g_{\boldsymbol{\theta}}(y - u) - g_{\boldsymbol{\theta}'}(y - u)| \leq C|\gamma - \gamma'| + C'|\sigma - \sigma'|,$$

for some constants $C$ and $C'$. More precisely, one can take

$$\begin{aligned} C &= \frac{6}{\gamma_{\min}^2 \sigma_{\min}}, \\ C' &= \frac{1}{\sigma_{\min}^2}\left(1 + 3\left\{1 + \frac{1}{\gamma_{\min}}\right\}\right). \end{aligned}$$

Next, observe that

$$|g_{\boldsymbol{\theta}'}(y - u) - g_{\boldsymbol{\theta}'}(y - u')| \leq C''|u - u'|,$$

where $C'' = 4\gamma_{\max}^2/[\gamma_{\min}\sigma^3]$. Which leads to

$$|g_{\boldsymbol{\theta}}(y-u) - g_{\boldsymbol{\theta}'}(y-u')| \leq C_g \max(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_\infty, |u-u'|),$$

for some constant $C_g > 0$. Similarly,

$$|h_{\boldsymbol{\theta}}(y-u) - h_{\boldsymbol{\theta}'}(y-u)| \leq C_1(4+\log(1+wy))|\gamma-\gamma'| + C_2|\sigma-\sigma'|,$$

Next,

$$|h_{\boldsymbol{\theta}'}(y-u) - h_{\boldsymbol{\theta}'}(y-u')| \leq C_7|u-u'|,$$

where $C_7 = 5/(\gamma_{\min}\sigma_{\min})$, leading to, for some $C_h > 0$,

$$|h_{\boldsymbol{\theta}}(y-u) - h_{\boldsymbol{\theta}'}(y-u')| \leq C_h \max(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_\infty, |u-u'|).$$

On the other hand,

$$|\phi(y-u,\boldsymbol{\theta}) - \phi(y-u,\boldsymbol{\theta}')| \leq \frac{1}{\gamma_{\min}^2}(2+\log(1+wy))|\gamma-\gamma'| + \frac{3}{\gamma_{\min}\sigma_{\min}}|\sigma-\sigma'|,$$

and

$$|\phi(y-u,\boldsymbol{\theta}') - \phi(y-u',\boldsymbol{\theta}')| \leq \frac{1}{\sigma_{\min}}|u-u'|.$$

Define $\mathfrak{F}_1 = \{g_{\boldsymbol{\theta}}(\cdot - u) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, $\mathfrak{F}_2 = \{h_{\boldsymbol{\theta}}(\cdot - u) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, and $\mathfrak{F}_3 = \{\phi(\cdot - u, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$. From [van der Vaart, 1998, Example 19.7], we get, for $i = 1, ..., 3$,

$$N(\varepsilon, \mathfrak{F}_i) \leq \varphi_i \|\Phi\|_2^{\alpha_1} \sigma_n^{\alpha_1} \varepsilon^{-\alpha_1},$$

for some $\alpha > 0$ and constants $\varphi_i$.

On the other hand, let

$$\mathfrak{F}_4 = \{\mathbf{x} \mapsto \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \colon \ell = 1, \ldots, K\},$$

and

$$\mathfrak{F}_5 = \{y \mapsto \mathbf{1}_{y>u} \colon u \in \mathcal{U}\}.$$

From Lemma 4 in [Lopez et al., 2016], we have $N(\varepsilon, \mathfrak{F}_4) \leq m^k K^{\alpha_2} \varepsilon^{-\alpha_2}$, where $\alpha_2 = 4(d+1)(d+2)$, and where $k$ is the number of discrete components taking at most $m$ modalities. On the other hand, from Example 19.6 in [van der Vaart, 1998], $N(\varepsilon, \mathfrak{F}_5) \leq 2\varepsilon^{-2}$.

From [Einmahl et al., 2005, Lemma A.1], we get, for $i = 1, \ldots, 3$,

$$N(\varepsilon, \mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5) \leq \frac{4m^k K^{\alpha_2} \max(C_g, C_h) \|\Phi\|_2^{\alpha_1} \sigma_n^{\alpha_1}}{\varepsilon^{\alpha_1+\alpha_2+\alpha_3}}.$$

Multiplying $\mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5$ by a single indicator function $\mathbf{1}_{\Phi(Y_i)\leq M_n}$ does not change the covering number, and the result follows. □

# C Technical Lemmas

**Lemma 11.** *1. The derivatives of the functions $y \to \phi(y-u, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are uniformly bounded by*

$$\Phi(y) = C(1 + \log(1 + wy)),$$

*where $C$ is a constant (not depending on $n$), and $w = \gamma_{\max}/\sigma_{\min}$.*

*2. There exists a certain $\rho_0 > 0$ such that*

$$m_{\rho_0} := \mathbb{E}\left[\exp(\rho_0 \Phi(Y))\right] < \infty.$$

*Proof.* To proof point 1, it is sufficient to derive the GP likelihood and see that they can be upper-bounded by $\Phi$.

Now, for point 2, note that for all $\mathbf{x}$, $\gamma(\mathbf{x}) \geq \gamma_{\min} > 0$, $Y$ is heavy-tailed random variable, then $\log(Y)$, and thus $\Phi(Y)$, is a light-tailed random variable. Thus $\Phi(Y)$ has finite exponential moments. □

**Lemma 12.** *With $v_{\mathfrak{F}}$ defined in Proposition 4,*

$$v_{\mathfrak{F}} \leq \frac{M_n^2 k_n}{n}.$$

*Proof.* We have

$$\begin{aligned} v_{\mathfrak{F}} &\leq \mathbb{E}\left[\Phi(Y)^2 \mathbf{1}_{Y \geq u_{\min}} \mathbf{1}_{\Phi(Y) \leq M_n}\right] \\ &\leq M_n^2 \mathbb{P}(Y \geq u_{\min}) = \frac{M_n^2 k_n}{n}. \end{aligned}$$

□

**Lemma 13.** *Define, for $j = 1, 2, 3$,*

$$E_{j,n} = \mathbb{E}\left[\Phi(Y)^j \mathbf{1}_{\Phi(Y) \geq M_n} \mathbf{1}_{Y \geq u_{\min}}\right].$$

*Under the assumptions of Theorem 7,*

$$E_{j,n} \leq \frac{\mathfrak{e}_j k_n^{1/2}}{n^{1/2} n^{\rho_0 \beta a_2/4}}.$$

*Proof.* Applying twice Cauchy-Schwarz inequality leads to

$$E_{j,n} \leq \mathbb{P}(Y \geq u_{\min})^{1/2} \mathbb{E}[\Phi(Y)^{2j} \mathbf{1}_{\Phi(Y) \geq M_n}]^{1/2} \leq \frac{k_n^{1/2}}{n^{1/2}} \mathbb{E}[\Phi(Y)^{4j}]^{1/4} \mathbb{P}(\Phi(Y) \geq M_n)^{1/4}.$$

Next, from Chernoff inequality,

$$\mathbb{P}(\Phi(Y) \geq M_n) \leq \exp(-\rho_0 M_n) \mathbb{E}[\exp(\rho_0 \Phi(Y))] \leq \frac{m_{\rho_0}}{n^{\rho_0 \beta a_2}}.$$

□

**R codes:** Tthe `R` codes are publicly available at https://github.com/antoine-heranval/Generalized-Pareto-Regression-Trees-for-extreme-event-analysis.

**Ethical Approval and Consent to participate** All the authors approve and consent to participate.

**Consent for publication** All the authors consent for publication.

**Human and Animal Ethics** Not applicable.

**Availability of supporting data** Since the data were provided by a private partnership with the Mission Risques Naturels, the data are not publicly available.

**Competing interests** The authors have no competing interests.

**Funding** Not applicable.

**Authors' contributions** All the authors wrote the main manuscript text, and the supplementary material All the authors prepared the all figures and tables All the authors reviewed the manuscript.

**Acknowledgments** The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project).

# References

Catastrophe naturelle, assurance et prévention. Technical report, Mission Risques Naturels, 2016. URL `https://www.mrn.asso.fr`.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974. doi: https://doi.org/10.1080/00401706.1974.10489157.

M. Allouche, S. Girard, and E. Gobet. Estimation of extreme quantiles from heavy-tailed distributions with neural networks. working paper or preprint, 2022. URL `https://hal.science/hal-03751980`.

A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974. doi: https://doi.org/10.1214/aop/1176996548.

A. M. Barlow, E. Mackay, E. Eastoe, and P. Jonathan. A penalised piecewise-linear model for non-stationary extreme value analysis of peaks over threshold. *Ocean Engineering*, 267:113265, 2023.

J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for Pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118, 2004. doi: https://doi.org/10.1016/S0047-259X(03)00125-8.

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: Theory and Applications*. John Wiley & Sons, 2004. ISBN 978-0-471-97647-9.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

J. Carreau and M. Vrac. Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47(10), 2011.

A. Charpentier, L. Barry, and M. R. James. Insurance against natural catastrophes: balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance - Issues and Practice*, May 2021. ISSN 1018-5895, 1468-0440. doi: https://doi.org/10.1057/s41288-021-00233-7.

P. Chaudhuri. Asymptotic consistency of median regression trees. *Journal of statistical planning and inference*, 91(2):229–238, 2000. doi: https://doi.org/10.1016/S0378-3758(00)00180-4.

P. Chaudhuri and W.-Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, pages 561–576, 2002.

V. Chavez-Demoulin, P. Embrechts, and M. Hofert. An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2015. doi: https://doi.org/10.1111/jori.12059.

V. Chernozhukov. Extremal quantile regression. *The Annals of Statistics*, 33(2):806–839, 2005.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990. doi: https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

G. De'ath and K. E. Fabricius. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000. doi: https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

U. Einmahl, D. M. Mason, et al. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005. doi: https://doi.org/10.1214/009053605000000129.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021. doi: https://doi.org/10.1016/j.insmatheco.2021.02.009.

L. Gardes and G. Stupfler. An integrated functional weissman estimator for conditional extreme quantiles. *REVSTAT-Statistical Journal*, 17(1):109–144, 2019.

S. Gey and E. Nedelec. Model selection for cart regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005. doi: https://doi.org/10.1109/TIT.2004.840903.

N. Gnecco, E. M. Terefe, and S. Engelke. Extremal random forests. *arXiv preprint arXiv:2201.12865*, 2022.

C. González, J. Mira-McWilliams, and I. Juárez. Important variable assessment and electricity price forecasting based on regression tree models: Classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission Distribution*, 9 (11):1120–1128, 2015. doi: https://doi.org/10.1049/iet-gtd.2014.0655.

W. K. Huang, D. W. Nychka, and H. Zhang. Estimating precipitation extremes using the log-histospline. *Environmetrics*, 30(4):e2543, 2019.

R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002. doi: https://doi.org/10.1016/S0309-1708(02)00056-8.

W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011. doi: https://doi.org/10.1002/widm.8.

W.-Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014. doi: https://doi.org/10.1111/insr.12016.

O. Lopez, X. Milhaud, and P.-E. Thérond. Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10(2):2685–2716, 2016. doi: https://doi.org/10.1214/16-EJS1189.

O. C. Pasche and S. Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*, 2022.

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3 (1):119–131, 1975.

J. Richards and R. Huser. A unifying partially-interpretable framework for neural network-based extreme quantile regression. *arXiv preprint arXiv:2208.07581*, 2022.

T. Rietsch, P. Naveau, N. Gilardi, and A. Guillou. Network design for heavy rainfall analysis. *Journal of Geophysical Research: Atmospheres*, 118(23):13–075, 2013.

V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818, 2015. doi: https://doi.org/10.1016/j.oregeorev.2015.01.001.

E. Ross, S. Sam, D. Randell, G. Feld, and P. Jonathan. Estimating surge in extreme north sea storms. *Ocean Engineering*, 154:430–444, 2018.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, 10(1):33–60, 2012.

R. L. Smith. Threshold methods for sample extremes. In *Statistical extremes and applications*, pages 621–638. Springer, 1984.

R. L. Smith. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377, 1989.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

X. Su, M. Wang, and J. Fan. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13(3):586–598, 2004. doi: https://doi.org/10.1198/106186004X2165.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

P. Tencaliec, A.-C. Favre, P. Naveau, C. Prieur, and G. Nicolet. Flexible semiparametric generalized pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31 (2):e2582, 2020.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

J. Velthoen, J.-J. Cai, G. Jongbloed, and M. Schmeits. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019. doi: https://doi.org/10.1007/s10687-019-00355-1.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*, 2021.

H. J. Wang, D. Li, and X. He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464, 2012. doi: https://doi.org/10.1080/01621459.2012.716382.

B. D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for us wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019.

# Article C

# Semiparametric copula models applied to the decomposition of claim costs

# Semiparametric copula models applied to the decomposition of claim amounts.

Sébastien Farkas[1], Olivier Lopez[1].

**Abstract**

In this paper, we develop a conditional copula model to analyze the distribution of a claim that generates different types of costs and/or simultaneously impact several guarantees. Our methodology is adapted to taking into account the particular structure of our data, since observations are subject to right-censoring. Right-censoring occurs since payment of a claim is not made instantaneously, and therefore unsettled claims only provide a partial information on the phenomenon that one wishes to model. The new methodology that we develop is supported by theoretical results that show the asymptotic normality of our estimators. A simulation study and a real data analysis illustrate the method.

**Key words:** Conditional copula; Right-censoring; Claim reserving; Insurance.

**Short title:** Semiparametric copulas for the decomposition of claim amounts.

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France.

E-mails: sebastien.farkas@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr

# 1 Introduction

Analyzing and predicting the evolution of claims is a challenging aspect of risk management, especially in some branches where the volatility of the final amount may be important. The question of computing appropriate reserve is, of course, crucial, but such type of analysis also enable to take actions before the settlement of a claim, in order to reduce its impact once a difficulty has been identified. In this paper, we consider the particular situation where the final cost of a claim is decomposed between several categories of costs. The example we have in mind is the study of cyber-insurance contracts. Due to the relative novelty of insurance products against this type of risk, the various types of situations that are covered, and the innovations in the products (different type of technical or legal assistance may be included in such policies), a clear analysis of what generates the expanses for the company is required. This problematic is also present for many types of claims, for which, for example, several guarantees are simultaneously activated.

With the increase of available data to analyze claims, many techniques have been proposed recently to perform an accurate evaluation of RBNS (Reported But Not Settled) claims. Traditional aggregated methods like chain-ladder, see Mack (1993), Merz et al. (2013) or Saluz et al. (2014), can be modified to incorporate additional informations. For example, Wüthrich (2017) or Wüthrich (2016) considered the introduction of covariates in the development factors, allowing to use machine learning techniques to increase the precision of the reserve computation. Alternatively, micro-level reserving methods directly consider the prediction of the evolution of a claim based on its characteristics. Such type of methods have been proposed for example by Norberg (1993), Norberg (1999), Antonio and Plat (2010), Antonio et al. (2016), or Pigeon et al. (2014) in a dynamic setting (the time phenomena being modeled by Poisson processes). Lopez et al. (2016) proposed regression tree techniques to predict the amount of a claim based on information available at its occurrence, while Sabban et al. (2020) used deep learning methods to deduce from insurance reports, the outcome of severe claims. A comparison between micro-level and macro-level methods can be found for example in Jin and Frees (2013).

In the present paper, we develop a methodology that is close to Lopez (2019). Our model to predict the outcome of a claim is based on incomplete data: settled claims but also unsettled one. This phenomenon is related to right-censoring, which is classical in survival analysis, see Fleming and Harrington (2011) for example. The general idea is the following: settled claims are, in average, closed faster than the one that are still

open at the extraction of the database. Consequently, calibrating a model based on the settled claims solely typically tends to lead to an under-evaluation of the reserve. Hence an adapted methodology should be developed to correct this bias. The second axis of our methodology is to deal with an outcome of a claim which is multivariate, since the cost is decomposed into several lines of businesses or types of expanses. Hence, it is natural to rely on copula theory (see for example Nelsen (2006)) to model the dependence structure of each component of the vector of losses.

An advantage of the copula approach is the possibility to use models of various types to describe the distribution of each margin. Since the distribution type of each type of expanse may be quite different (typically some may be heavy-tailed, some may not), copula theory allows to perform the analysis of these marginal distributions separately, while the dependence structure is, in a second step, done through the fitting of a parametric copula function. This explains the popularity of such techniques, see Zhao and Zhou (2010), Bouyé et al. (2000) or Jaworski et al. (2010) for examples of applications. Due to the presence of covariates describing the circumstances and nature of the claim, the dependence structure may not be the same for all claims, leading to a conditional copula modeling, see Fermanian and Wegkamp (2004) or Veraverbeke et al. (2011). Our approach is then close to the semiparametric model developed by Abegaz et al. (2012), but with an adaptation to the particular structure of our data. Apart from the bias caused by censoring, a difficulty arises since, among the covariates that may have impact on the dependence structure, one of them is unavailable for open claims (namely, the time before settlement). Hence, when it comes to prediction, an evaluation of this time before settlement must be combined with the conditional copula model we develop.

The rest of the paper is organized as follows. In Section 2, we describe the general model we develop to analyse the joint distribution of the loss vector. Theoretical validity of this approach is provided in Section 3. A simulation study and a real data analysis demonstrate the practical feasibility of the method in Section 4.

# 2 A model for the decomposition of the claim cost

This section is devoted to the description of the model used to describe the cost of a claim, and to the techniques we use to calibrate its parameters. Section 2.1 describes the structure of our data, with the description of the right-censoring phenomenon. Correction of the bias caused by right-censoring is considered in Section 2.2. As we already mentioned,

our approach is based on a separation between the margins, for which models are proposed in Section 2.3, and the dependence structure via a conditional copula model described in Section 2.4. The method to predict an open claim, once the model is fitted, is summarized in Section 2.5.

## 2.1 Model and observations

In some situations, an insurance claim can trigger several guarantees and generate varied additional expanses, like expert cost, legal fees and so on. In Section 4.2, we give an example in the case of medical malpractice claims, but many other fields may be affected by such a decomposition of the costs. The cost of a claim is decomposed into $\mathbf{L} = (L^{(1)}, ..., L^{(d)})$, the total cost being $L_{tot} = \sum_{k=1}^{d} L^{(k)}$. It is expected that these partial costs may not be independent from each other, since related to the same claim. Moreover, their distributions may be quite different since they are not of the same nature (and not affected with the same limits). Hence, a joint modeling of the different components of the random vector $\mathbf{L}$ may be delicate.

Copula analysis is a convenient way to deal with such difficulties. Sklar's Theorem, see Sklar (1959), is at the core of the copula approach, and states that

$$F(l_1, ..., l_d) = \mathfrak{C}(F_1(l_1), ..., F_d(l_d)),$$

where $F(l_1, ..., l_d) = \mathbb{P}(L_1 \leq l_1, ..., L_d \leq l_d)$, $F^{(k)}(l) = \mathbb{P}(L^{(k)} \leq l)$, and $\mathfrak{C}$ is a copula function, that is a function from $[0,1]^d \to [0,1]$ which is the distribution of a $d-$dimensional random vector whose margins are uniformly distributed over $[0,1]$. The copula function $\mathfrak{C}$ is unique when the margins are continuous (which will be our assumption throughout this paper), hence this object characterizes the dependence structure of the random vector $\mathbf{L}$ (informations on the marginal distributions are contained in the one-dimensional cumulative distribution functions $F^{(k)}$).

In our case, covariates are present since one has informations on the circumstances of the claim and on the characteristics of the policyholder. This covariates $\mathbf{X} \in \mathbb{R}^p$ have impact on the marginal distribution, but potentially also on the dependence structure. Let $F(l_1, ..., l_d|\mathbf{x}) = \mathbb{P}(L^{(k)} \leq l_1, ..., L^{(d)} \leq l_d|\mathbf{x})$, and $F^{(k)}(l|\mathbf{x}) = \mathbb{P}(L^{(k)} \leq l_k|\mathbf{X} = \mathbf{x})$. Then, the conditional copula of $\mathbf{L}$ conditionally to $\mathbf{X} = \mathbf{x}$ (see e.g. Veraverbeke et al. (2011)) is the copula function $\mathfrak{C}^{(\mathbf{x})}$ such that

$$F(l_1, ..., l_d|\mathbf{x}) = \mathfrak{C}^{(\mathbf{x})}(F^{(1)}(l_1|\mathbf{x}), ..., F^{(d)}(l_d|\mathbf{x})).$$

Standard regression models can be used to estimate each of the marginal conditional distribution functions, see examples in Section 2.3 below. Our main purpose is to focus on the estimation of the dependence structure.

Introducing a parametric copula family $\mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^m$ is finite dimensional. We assume that, for all possible values of $x$, $\mathfrak{C}^{(\mathbf{x})} \in \mathcal{C}$. Let $\theta(\mathbf{x})$ denote the function such that $\mathfrak{C}^{(\mathbf{x})} = \mathfrak{C}_{\theta(\mathbf{x})}$. Our aim is to retrieve this function, either using a parametric model or a semiparametric model, as described in Section 2.4 below.

To estimate this function, we rely on observations of a set of $n$ past claims. These claims $(\mathbf{L}_i, \mathbf{X}_i)_{1\leq i\leq n}$ are assumed to be i.i.d. In addition, let $T_i$ denote the time required to solve claim $i$, that is the difference between the date of occurrence of the claim and the date of its settlement. In the database used to calibrate the model, all claims are not closed. For ongoing ones, $T_i$ is unknown. This is a classical right-censoring situation, see Fleming and Harrington (2011): let $Y_i = \inf(T_i, C_i)$ and $\delta_i = \mathbf{1}_{T_i\leq C_i}$, where $C_i$ is a random censoring variable. If $\delta_i = 1$ the claim is closed, and one observes $Y_i = T_i$. In the opposite situation, the claim is still open since $Y_i = C_i$, $C_i$ being the difference between the date of occurrence and the date at which one looses track on the claim (because data has been extracted before its closure, or because the claim is part of a retroceded portfolio).

The reason for not calibrating the model only on claims such that $\delta_i = 1$ is explained in detail in Lopez (2019). Typically, a positive (and potentially strong) correlation is expected between $T_i$ and the total loss $L_{tot}$, based on the heuristic that "the longer it takes for a claim to be solved, the higher it costs". Calibrating a model only on closed claims is likely to tend to underestimate the typical values taken by $L_{tot}$, since the population of closed claims is characterized by an overrepresentation of claims with small final amount.

This leads to the following set of observations, which are i.i.d. replications $(\mathbf{M}_i, \mathbf{X}_i, Y_i, \delta_i)_{1\leq i\leq n}$ of $(\mathbf{M}, \mathbf{X}, Y, \delta)$ where

$$\left\{\begin{array}{rcl} Y & = & \inf(T, C), \\ \delta & = & \mathbf{1}_{T\leq C}, \\ \mathbf{M} & = & (M^{(1)}, ..., M^{(d)}), \end{array}\right.$$

with $M^{(k)} = L^{(k)}$ if $\delta = 1$, and $M^{(k)} = L^{(k)}$ otherwise. $M^{(k)}$ represents partial payments done on the $k-$th line at the end of the observation of the claim. Each of these variables can be understood as a right-censored variable, that is $M^{(k)} = \inf(L^{(k)}, D^{(k)})$, where $D^{(k)} > L^{(k)}$ only when $\delta = 1$.

## 2.2 Inverse probability of censoring weighting (IPCW)

Since a duration phenomenon is present in the data acquisition process (one must wait until the settlement of a claim to know its final state), censoring would introduce some bias if no correction is performed. The Inverse Probability of Censoring Weighting methodology (IPCW), see for example Van der Laan and Robins (2003), is a simple way to proceed. It consists of determining an appropriate weight to put on each observation to asymptotically cancel the bias caused by the censoring. Only the uncensored observations are affected with a non-zero weight (since these observations are complete), but the partial information contained in the censored one is used to determine these weights.

The core of such an approach is the following result. In the rest of the paper, we assume that $(T, \mathbf{L}, \mathbf{X})$ is independent from the censoring mechanism $C$. Under this assumption, for any function $\phi$ such that $E[|\phi(T, \mathbf{L}, \mathbf{X})|] < \infty$, and such that $\phi(y, \mathbf{m}, \mathbf{x}) = 0$ if $y$ is not in the support of the distribution of $Y$,

$$E\left[\frac{\delta\phi(Y, \mathbf{M}, \mathbf{X})}{S_C(Y)}\right] = E\left[\phi(T, \mathbf{L}, \mathbf{X})\right], \tag{2.1}$$

where $S_C(y) = \mathbb{P}(C \geq y)$. In the following, we assume to simplify that $T$ and $C$ have the same support, which guarantees that $\inf\{t : \mathbb{P}(T \geq t) = 0\} = \inf\{t : \mathbb{P}(C \geq t) = 0\}$, so that (2.1) holds for all function $\phi$ with first order finite moment. In the general case, this would lead to consider truncated versions of functions whose support is not compact, introducing some bias which can not be removed without some additional parametric assumption (since one part of the distribution is not observed in this case).

Equation (2.1) implies that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(Y_i, \mathbf{M}_i, \mathbf{X_i})}{S_C(Y_i)} \rightarrow_{n\to\infty} E[\phi(T, \mathbf{L}, \mathbf{X})], \ \textit{almost surely}, \tag{2.2}$$

from the Strong Law of Large Numbers. The left-hand side quantity has the advantage to be computable from observed quantities, while the limit is an expectation with respect to the variables we are actually interested in ($T$, $\mathbf{L}$ and $\mathbf{X}$). Nevertheless, the function $S_C$ is in general unknown and hard to model, due to the lack of visibility on the censoring process.

A Kaplan-Meier estimator (see Kaplan and Meier (1958)) can be used for $S_C$, that is, if $T$ is continuous,

$$\hat{S}_C(t) = \prod_{Y_i \leq t}\left(1 - \frac{\delta_i}{\sum_{j=1}^{n} Y_j \mathbf{1}_{Y_j \geq t}}\right),$$

a more general expression can be found in  that covers also the discrete case. Defining

$$W_{i,n} = \frac{1}{n}\frac{\delta_i}{\hat{S}^C(Y_i)}, \tag{2.3}$$

the IPCW approach consists in estimating any quantity of the type $E[\phi(T, \mathbf{L}, \mathbf{X})]$ by

$$\sum_{i=1}^{n} W_{i,n}\phi(Y_i, \mathbf{M}_i, \mathbf{X_i}).$$

These computable weighted sums will replace every empirical mean that we would use in the case of complete (uncensored) data.

## 2.3 Regression model for the margins

Since the margins are expected to have heterogeneous behaviors (some lines may be very volatile while some others are not, some of them may have a low probability of activation...), the models that are used to study the distribution of each $L^{(k)}$ may be of different types. Typically we distinguish between fully parametric models, and semi-parametric or non-parametric ones.

**Fully parametric models**. In a fully parametric model, one assumes that $F^{(k)}(l|\mathbf{x}, t) = F_\beta(l|\mathbf{x}, t)$, where $\{F_\beta(\cdot|\mathbf{x}, t), \mathbf{x} \in \mathcal{X}, t \geq 0, \beta \in \mathcal{B}\}$ is a parametric set of distribution functions (here, $\mathcal{X}$ is the support of the random vector $\mathbf{X}$). Typically, this is the framework of the Generalized Linear Model (see Nelder and Baker (1972)). According to this model, the distribution of $L^{(k)}|\mathbf{X} = \mathbf{x}, T = t$ has density $f_{\beta(\mathbf{x},t)}(l)$, where $\{f_\beta : \beta \in \mathcal{B}\}$ is the collection of densities from a given exponential-type distribution. Moreover, it is assumed that

$$g(E[L^{(k)}|\mathbf{X} = \mathbf{x}, T = t]) = h(\beta(\mathbf{x}, t)) = \alpha'\mathbf{z},$$

where $\mathbf{z} = (\mathbf{x}, t)$, $g$ is a fixed monotonic function ($h$ is deduced from $g$) and $\alpha$ a finite dimensional parameter (here, the $'$ symbol denotes the transpose of a vector). From an estimator $\hat{\alpha}$ of $\alpha$, one deduce an estimator of the conditional distribution function $\hat{F}^{(k)}(l|\mathbf{x}, t) = \int_0^l f_{g^{-1}(\alpha'(\mathbf{x},t)')}(u)du$. Since $L^{(k)}$ is subject to censoring, a specific adaptation of the log-likelihood estimation is required. Consistent estimators of $\hat{\alpha}$ can be used, such as Stute (1999) (see also Lopez (2009)).

**Semi-parametric or nonparametric models.** More elaborate models, coming from example from machine-learning field, consist in decomposing $L^{(k)} = m^{(k)}(\mathbf{X}, T)+\varepsilon^{(k)}$, where $m$ is a function belonging to a potentially infinite dimension class, and $\varepsilon$ is a residual. For example, in the case where $m^{(k)}(\mathbf{x}) = E[L^{(k)}|\mathbf{X} = \mathbf{x}, T = t]$, $E[\varepsilon^{(k)}|\mathbf{X} = \mathbf{x}, T = t] = 0$,

while in the case of median regression, $m^{(k)}(\mathbf{x})$ is the conditional median and $\varepsilon$ is such that $\mathbb{P}(\varepsilon^{(k)} \geq 0|\mathbf{X} = \mathbf{x}, T = t) = 1/2$. If one assumes that $\varepsilon$ does not depend on $\mathbf{X}$ or $T$, $F^{(k)}(l|\mathbf{x}nt) = F^{(\varepsilon^{(k)})}(l - m(\mathbf{x}, t))$. Based on an estimator $\hat{m}^{(k)}$ of $m^{(k)}$, one can define $\hat{\varepsilon}_i^{(k)} = L_i^{(k)} - \hat{m}^{(k)}(\mathbf{X}_i, T_i)$, and

$$\hat{F}^{(\varepsilon^{(k)})}(e) = \sum_{i=1}^{n} W_{i,n} \mathbf{1}_{\hat{\varepsilon}_i^{(k)} \leq e},$$

and $\hat{F}^{(k)}(l|\mathbf{x}, t) = \hat{F}^{(\varepsilon^{(k)})}(l - \hat{m}^{(k)}(\mathbf{x}, t))$. Various machine learning techniques have been proposed for censoring models, like regression trees (see Bou-Hamad et al. (2011)), random forests (see Ishwaran et al. (2008), Gerber et al. (2020)) or neural networks (see Sabban et al. (2020)).

## 2.4 Conditional copula model estimation

Let $\mathbf{U}_i = (U_i^{(1)}, ..., U_i^{(d)})$ where $U_i^{(k)} = F^{(k)}(L_i^{(k)}|\mathbf{X}_i, T_i)$. The variables $U_i^{(k)}$ are uniformly distributed over $[0, 1]$. Moreover, the conditional density of $\mathbf{U}_i$ conditionally to $(\mathbf{X}_i, T_i)$ is $\mathfrak{C}_{\theta(\mathbf{X}_i, T_i)}$.

In the following, we consider two cases:

- the dependence structure does not depend on the covariates, that is $\theta(\mathbf{x}) = \theta_0$ (which is called by several authors the "simplifying assumption", see Portier and Segers (2018) or Derumigny and Fermanian (2017));
- $\theta(\mathbf{x})$ is estimated using a nonparametric estimator.

The first case is of course the most straightforward. If we had the ability to observe $(\mathbf{U}_i)_{1 \leq i \leq n}$, the maximum likelihood estimator $\theta_0$ would achieve the maximum of $\sum_{i=1}^{n} \log \mathfrak{c}_\theta(\mathbf{U}_i)$, where $\mathfrak{c}_\theta$ is the copula density associated with $\mathfrak{C}_\theta$, that is $\mathfrak{c}_\theta(\mathbf{u}) = \partial^d \mathfrak{C}_\theta(\mathbf{u}) / \partial u^{(1)} ... \partial u^{(d)}$. In our case, the vectors $(\mathbf{U}_i)_{1 \leq i \leq n}$ are not directly observed, but one can compute

$$\hat{U}_i^{(k)} = \hat{F}^{(k)}(M_i^{(k)}|\mathbf{X}_i, Y_i),$$

from the marginal distribution estimators of Section 2.3. Let us note that $\hat{U}_i^{(k)}$ is supposed to be close to $U_i^{(k)}$ only if $\delta_i = 1$ (uncensored observations), that is when $M_i^{(k)} = L_i^{(k)}$ and $Y_i = T_i$. As we will see in the following, we only need to estimate $\hat{U}_i^{(k)}$ when $\delta_i = 1$ to perform our estimation of the association parameter. Next, to correct the bias caused

by the censoring, we rely on the IPCW technique of Section 2.2, which leads to the pseudo-likelihood estimator

$$\hat{\theta}(\mathbf{x},t) = \hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} W_{i,n} \log \mathfrak{c}_\theta(\hat{\mathbf{U}}_i). \tag{2.4}$$

Let us recall that $W_{i,n} = 0$ if $\delta_i = 0$. This estimator is close to the one of Lopez (2019), and its behavior can be deduced from this work up to some small modifications that we mention in Section 6.3 and 6.4.

If we do not consider the simplifying assumption, the log-likelihood must be modified to obtained a conditional version. Introducing a kernel function $K$ (that is function from $\mathbb{R}^p$ to $\mathbb{R}$ such that $\int K(\mathbf{z})d\mathbf{z} = 1$, with $\int \mathbf{z}K(\mathbf{z})d\mathbf{z} = 0$), we define

$$\hat{\theta}(\mathbf{x},t) = \arg\max_{\theta\in\Theta} \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_\theta(\hat{\mathbf{U}}_i), \tag{2.5}$$

for $h > 0$, and $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$. The estimator (2.5) can be seen as an adaptation of the estimator of Abegaz et al. (2012). To simplify, we chose to apply the same bandwidth for each component of $\mathbf{Z} = (\mathbf{X}, T)$ to simplify. In full generality, $h$ is a square matrix of size $p + 1$.

## 2.5 Prediction of an open claim

Let us consider an open claim with characteristics $\mathbf{x}$. The claim is open since a duration $y$. The conditional copula model and the marginal regression models described above allows to determine the conditional distribution of $\mathbf{L}$ conditionally to $\mathbf{X} = \mathbf{x}$ and $T = t$. When it comes to predicting an open claim, they can not be used directly, since $T$ is unknown (one only knows that $T \geq y$). A possibility is to rely on a predictor of $T$, say $\hat{T}$. Let

$$\hat{p}(\mathbf{x},t) = \int_{\mathbb{R}^d} \mathbf{l} \times \mathfrak{c}_{\hat{\theta}(\mathbf{x},t)}(\hat{F}^{(1)}(l^{(1)}|\mathbf{x}), ..., \hat{F}^{(d)}(l^{(d)}|\mathbf{x}))d\hat{F}^{(1)}(l^{(1)}|\mathbf{x},t)...d\hat{F}^{(d)}(l^{(d)}|\mathbf{x},t),$$

which is an estimator of $p(\mathbf{x},t) = E[\mathbf{L}|\mathbf{X} = \mathbf{x}, T = t]$. Based on $\hat{T}$, one can use $\hat{p}(\mathbf{x},\hat{T})$ to predict the final state of the claim. This only gives the "central scenario". The fitted conditional copula density and the marginal models can be used more generally to simulate the conditional distribution of $\mathbf{L}$, and thus to get an analysis on the volatility of this prediction.

The crucial question is hence to determine a proper prediction method $\hat{T}$. This can be based on a standard regression model on the censored variable $T$. An accelerated

failure-time model (see Wei (1992)) can for example be fitted to obtain an estimator of $F^T(t|\mathbf{x}) = \mathbb{P}(T \leq t|\mathbf{X} = \mathbf{x})$, or a Cox model (see Cox (1975)). Other semiparametric models, like Stute (1999) or Lopez (2009) can also be used. Machine learning methods, like for example survival forests (see Ishwaran et al. (2008)) are also available. In each case, one obtains an estimator $\hat{F}^T(t|\mathbf{x})$ of $F^T$, from which a predictor of $T$ can be obtained.

For an open claim with characteristics $\mathbf{x}$, open since $y$, the idea is to estimate $E[T|\mathbf{X} = \mathbf{x}, T \geq y]$. This leads to

$$\hat{T} = \frac{\int_y^\infty t d\hat{F}^T(t|\mathbf{x})}{\hat{S}^T(y|\mathbf{x})}, \tag{2.6}$$

where $\hat{S}^T(y|\mathbf{x}) = 1 - \hat{F}^T(y - |\mathbf{x})$. Clearly, the quality of the prediction will rely on the regression model on $T$.

**Remark 2.1** *It is important to use a predictor $\hat{T}$ of the form (2.6), that is an estimator of $E[T|\mathbf{X} = \mathbf{x}, T \geq y]$, instead of a more simple estimator of $E[T|\mathbf{X} = \mathbf{x}]$ : this alternative method does not take into account all the available information on $T$, and could lead to predictions for which $\hat{T} < y$.*

# 3 Consistency of the approach

To study the theoretical behavior of the method, we first describe in Section 3.1 the assumptions required to obtain consistency of the estimation of the dependence structure. Our asymptotic results are gathered in Section 3.2.

## 3.1 List of assumptions and discussion

We distinguish between three types of assumptions required to obtain the theoretical results: on the copula family, on the estimation of the margins, and regularity assumptions (required only if the simplifying assumption does not hold, and that one considers the kernel estimator (2.5)). The assumptions are of the same type as the one used by Tsukahara (2005) to study the behavior of semiparametric estimators of copulas (that is the copula structure is parametric, but the margins are estimated through a nonparametric estimator, namely the empirical distribution function). In our case, we add assumptions on how the margins are estimated, since we want to be able to consider various types of models for the margins.

**Assumptions on the copula family.**

Before stating the assumptions, we need to introduce some notations. We use the bracketing number (see, for example, Chapter 19 in Van der Vaart (2000)) to define the richness of a class of functions. A bracket $[\mathfrak{u}, \mathfrak{l}]$, where $\mathfrak{u}$ and $\mathfrak{l}$ are functions such that $\mathfrak{u} \leq \mathfrak{l}$, is the set of functions $\mathfrak{f}$ such that $\mathfrak{u} \leq \mathfrak{f} \leq \mathfrak{l}$. A $\varepsilon-$bracket is such that $E\left[(\mathfrak{u}(\mathbf{U}) - \mathfrak{l}(\mathbf{U}))^2\right] \leq \varepsilon^2$. $N_{[]}(\varepsilon, \mathcal{F})$ denotes the number of $\varepsilon-$brackets required to cover a class of functions $\mathcal{F}$. How fast $N_{[]}(\varepsilon, \mathcal{F})$ explodes when $\varepsilon$ tends to zero is an indication about the complexity of the class of functions $\mathcal{F}$.

**Assumption 1** *Let* $\mathcal{F} = \{\mathbf{u} \to \log \mathfrak{c}_\theta(\mathbf{u}) : \theta \in \Theta\}$, *with* $\log \mathfrak{c}_\theta(\mathbf{u}) \leq \log \mathfrak{C}(\mathbf{u})$ *for all* $\theta$ *and* $\mathbf{u}$ *with* $E[\log \mathfrak{C}(\mathbf{u})^2] < \infty$. *Assume that* $N_{[]}(\varepsilon, \mathcal{F}) \leq A\varepsilon^{-k}$, *for some* $A$ *and* $k > 0$.

This assumption is relatively easy to fulfill for a parametric copula family. Typically, a polynomial bound $A\varepsilon^{-k}$ for the covering number is obtained when the class $\mathcal{F}$ is regular enough (typically, when this class is Lipschitz with respect to the parameter $\theta$, see Van der Vaart (2000)).

Next, we need to dominate the class of copula functions and some of their derivatives, with an assumption which is close to the one present in Tsukahara (2005). The only difference stands in the presence of censoring in our case, which strengthens the assumptions. Right-censoring induces potentially erratic behavior when studying the right-tail of the distribution. Typically, this explains the introduction of $S^C$ and a function $\mathfrak{K}$ in the last two moment conditions of Assumption 2, where we recall that $S^C$ is the survival function of the censoring, and we define

$$\mathfrak{K}(t) = \left[ -\int_\infty^t \frac{dS^C(s)}{S^C(s)^2 S^T(s)} \right]^{-1},$$

where $S^T$ is the survival function of $T$. If these two functions decreased too fast (compared to the tail of the distribution of $T$), the proper convergence rate can not be achieved. Truncation of the highest observations is required, leading to some bias that can not be cancelled even asymptotically. This type of conditions is classical in presence of censoring, see for example Gill (1983).

Let us introduce additional notations. A function $r : (0,1) \to (0;\infty)$ is said $u-$shaped if $r$ is symmetric around $1/2$ and increasing on (0,1/2]. For a $u-$shaped function $r$ and for $0 < \beta < 1$, define

$$r_\beta(t) = r(\beta t)\mathbf{1}_{0<t\leq 1/2} + r(1 - \beta(1-t))\mathbf{1}_{1/2<t\leq 1}.$$

A reproducing $u-$shaped function is the $u-$shaped function such that, for all $\beta$, there exists $M_\beta \geq 0$ such that $r_\beta \leq M_\beta r$. We use the notation $\mathcal{R}$ for the set of reproducing

$u-$shaped functions. We also consider $\mathcal{Q}$ the set of continuous functions $q$ on $(0,1)$ that are $u-$shaped, and such that $\int_0^1 q(t)^{-2}dt < \infty$.

**Assumption 2** *Let*

$$\begin{aligned}\Phi(\mathbf{u}) &= \frac{\nabla_\theta \mathfrak{c}_{\theta(\mathbf{z})}(\mathbf{u})}{\mathfrak{c}_{\theta(\mathbf{z})}},\\ \Phi_\theta^{(j,k)}(\mathbf{u}) &= \frac{\partial^2 \log \mathfrak{c}_\theta(\mathbf{u})}{\partial\theta^{(j)}\partial\theta^{(k)}}.\end{aligned}$$

*Let* $\dot{\Phi} = (\dot{\Phi}^{(1)}, ..., \dot{\Phi}^{(d)})$ *(resp.* $\dot{\Phi}_\theta^{(j,k)} = (\dot{\Phi}_\theta^{(j,k),1}, ..., \dot{\Phi}_\theta^{(j,k),d})$*) denote the vectors of partial derivatives of* $\Phi$ *(resp.* $\Phi_\theta^{(j,k)}$*). Assume that there exists functions* $\mathfrak{q}^{(k)} \in \mathcal{Q}$, $\mathfrak{r}^{(k)} \in \mathcal{R}$, $\tilde{\mathfrak{r}}^{(k)} \in \mathcal{R}$ *and* $\bar{\mathfrak{r}}^{(k)} \in \mathcal{R}$ *such that*

$$\begin{aligned}|\Phi(\mathbf{u})| + \sup_{j,k,\theta}|\Phi_\theta^{(j,k)}(\mathbf{u})| &\leq \prod_{k=1}^d \mathfrak{r}^{(k)}(u^{(k)}),\\ |\dot{\Phi}^{(k)}(\mathbf{u})| &\leq \tilde{\mathfrak{r}}^{(k)}(u^{(k)})\prod_{j\neq k}\mathfrak{r}^{(j)}(u^{(j)}),\\ \sup_{j,k,\theta}|\dot{\Phi}_\theta^{(j,k),l}(\mathbf{u})| &\leq \bar{\mathfrak{r}}^{(l)}(u^{(l)})\prod_{j\neq l}\mathfrak{r}^{(j)}(u^{(j)}),\end{aligned}$$

*with, for some* $\iota > 0$,

$$\begin{aligned}E\left[\frac{\left\{\prod_{k=1}^d \mathfrak{r}^{(k)}(U^{(k)})\right\}^2}{S_C(T)}\right] + E\left[\frac{\left\{\prod_{k=1}^d \mathfrak{r}^{(k)}(U^{(k)})\right\}}{\mathfrak{K}(T)^{1/2+\iota}}\right] &< \infty,\\ E\left[\frac{\mathfrak{q}^{(k)}(u^{(k)})\tilde{\mathfrak{r}}^{(k)}(U^{(k)})\prod_{j\neq k}\mathfrak{r}^{(j)}(U^{(j)})}{\mathfrak{K}(T)^{1/2+\iota}}\right] &< \infty,\\ E\left[\frac{\mathfrak{q}^{(k)}(u^{(k)})\bar{\mathfrak{r}}^{(k)}(U^{(k)})\prod_{j\neq k}\mathfrak{r}^{(j)}(U^{(j)})}{S^C(T)^{\iota}}\right] &< \infty.\end{aligned}$$

The functions $\mathfrak{q}^{(k)}$ vanish near the boundaries of the unit square, so that they allow to lower the moment conditions (for typical copula families, the explosion of the partial derivatives of the copula density increases with the order of these derivatives). Nevertheless, there is a compromise: these functions should not decrease to fast, since they are related on the next assumption on the margins.

**Assumptions on the estimation of the margins.**

The margins estimation should satisfy the following assumption.

**Assumption 3** *Assume that*

$$\sup_{i=1,\ldots,n}\sup_{k=1,\ldots,d}\left|\frac{U_i^{(k)}}{\hat{U}_i^{(k)}}+\frac{1-U_i^{(k)}}{1-\hat{U}_i^{(k)}}\right| = O_P(1), \tag{3.1}$$

$$\sup_{i=1,\ldots,n}\sup_{k=1,\ldots,d}\left|\frac{U_i^{(k)}-\hat{U}_i^{(k)}}{\mathfrak{q}^{(k)}(U_i^{(k)})}\right| = O_P(\varepsilon_n), \tag{3.2}$$

*for some sequence $\varepsilon_n$ tending to zero.*

This conditions would be fulfilled if one would consider the empirical distribution function to estimate the margins, as long as $\mathfrak{q}^{(k)}(u) \geq [u(1-u)]^\alpha$ with $\alpha < 1/2$. These condition also hold for parametric estimation of the margins, as long as the model is regular enough.

Additionally, we need an asymptotic i.i.d. representation of these pseudo-observations to obtain asymptotic normality under the simplifying assumption. This assumption will not be required when dealing with a kernel estimator. Indeed, under the simplifying assumption, the expected convergence rate is $n^{-1/2}$ (maximum likelihood convergence rate), and one may not expect to beat this convergence rate when estimating the margins. On the other hand, the sequence $\varepsilon_n$ defined in Assumption 3 may be faster than the nonparametric obtained for the conditional estimator.

**Assumption 4** *Under the simplifying assumption, assume that*

$$\hat{U}_i^{(k)} - U_i^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\eta^{(k)}_{\mathbf{L}_i^{(k)},\mathbf{Z}_i}(\mathbf{L}_j^{(k)},\mathbf{Z}_j,\delta_j) + r_{i,n}^{(k)},$$

*where* $\sup_{i,k}|r_{i,n}^{(k)}| = o_P(n^{-1/2})$, *with* $E[\eta^{(k)}_{\mathbf{l},\mathbf{z}}(\mathbf{L}^{(k)},\mathbf{Z}_j,\delta)] = 0$, *and* $E[\eta^{(k)}_{\mathbf{L}_1^{(k)},\mathbf{Z}_1}(\mathbf{L}_2^{(k)},\mathbf{Z}_2,\delta_2)^2] < \infty$.

**Assumptions on the kernel and on the regularity of the copula regression model.**

Let us recall that this last set of assumptions is only required if the simplifying assumption does not hold.

**Assumption 5** *The kernel function $K : \mathbb{R}^{p+1} \to \mathbb{R}$, has the following properties:*

$$\int K(\mathbf{u})d\mathbf{u} = 1,$$

$$\int \mathbf{u}K(\mathbf{u})d\mathbf{u} = 0,$$

$$\int \mathbf{u}^2|K(\mathbf{u})|d\mathbf{u} < \infty.$$

Additionally, we need a regularity assumption on the conditional distribution of $\mathbf{U}$ given $\mathbf{Z}$.

**Assumption 6** *Let $f_{\mathbf{U},\mathbf{Z}}(\mathbf{u},\mathbf{z})$ denote the joint density of $(\mathbf{U},\mathbf{Z})$ computed at point $(\mathbf{u},\mathbf{z})$. Assume that, for all $\mathbf{z}\in\mathcal{Z}$,*

$$f_{\mathbf{U},\mathbf{Z}}(\mathbf{u},\mathbf{z}) \leq \mathfrak{s}_1(\mathbf{u}), \tag{3.3}$$

$$\forall \mathbf{v},\ \mathbf{v}'\nabla^2_{\mathbf{z}} f_{\mathbf{U},\mathbf{Z}}(\mathbf{u},\mathbf{z})\mathbf{v} \leq \mathbf{v}'\mathfrak{s}_2(\mathbf{u})\mathbf{v}, \tag{3.4}$$

*with*

$$\begin{aligned} \|E\left[\Phi(\mathbf{U})\mathfrak{s}_2(\mathbf{U})\right]\|_\infty &< \infty, \\ \|E\left[\dot{\Phi}(\mathbf{U})\mathfrak{s}_1(\mathbf{U})\right]\|_\infty &< \infty. \end{aligned}$$

## 3.2 Asymptotic behavior of the estimator of the association parameter

We now state our main theoretical result on the asymptotic behavior of the estimates of the association parameter. We obtain two asymptotic representations, one under the simplifying assumption, and one in the case where the association parameter depends on the covariates.

**Theorem 3.1** *Let $\theta_0 = \arg\max_\theta E\left[\log \mathfrak{c}_\theta(\mathbf{U})\right]$, and $\theta_h^*(\mathbf{z}) = \arg\max_\theta E\left[K\left(\frac{\mathbf{Z}-\mathbf{z}}{h}\right)\log \mathfrak{c}_\theta(\mathbf{U})\right]$.*

1. *Under Assumptions 1 to 4,*

$$(\hat{\theta}-\theta_0) = -\Sigma^{-1}\left[\frac{1}{n}\sum_{i=1}^n\left\{\nu_i^\Phi + \sum_{k=1}^d \Lambda_i^{(k)}\right\}\right] + o_P(n^{-1/2}), \tag{3.5}$$

*where $\nu^\Phi$ is defined in Proposition 6.5, and*

$$\Lambda_i^{(k)} = E\left[\phi^{(k)}(\mathbf{U}^{(j)},\theta_0)\eta_{\mathbf{L},Y,\delta}(\mathbf{L}_1,\mathbf{Z}_1)|\mathbf{L}_1 = L_i, \mathbf{Z}_1 = \mathbf{Z}_i\right].$$

*Hence*

$$\hat{\theta}-\theta_0 \Longrightarrow \mathcal{N}\left(\mathbf{0}, \Sigma^{-1}V\Sigma\right), \tag{3.6}$$

*where $V$ is the covariance matrix of $\nu_1(\theta_0) + \sum_{j=1}^d \eta_{\mathbf{L}_1,\mathbf{Z}_1,\delta_1}(\theta_0)E\left[\phi^{(j)}(\mathbf{U}^{(j)},\theta_0)\right]$, and where $\Sigma = E\left[\nabla^2_\theta \log \mathfrak{c}_{\theta_0}(\mathbf{U})\right]$ is supposed to be invertible.*

2. *Under Assumptions 1 to 3, and under Assumption 5 and 6,*

$$\hat{\theta}(\mathbf{z}) - \theta_h^*(\mathbf{z}) = -\Sigma(\mathbf{z})^{-1} \left\{ \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \Phi(\mathbf{U}_i) \right\} + o_P(n^{-1/2} h^{(p+1)/2}),$$

*with* $\Sigma(\mathbf{z}) = \left(\sigma^{(j,k)}(\mathbf{z})\right)_{j,k}$, *is supposed to be invertible, with*

$$\sigma^{(j,k)}(\mathbf{z}) = E\left[\Phi_{\theta(\mathbf{z})}^{(j,k)}(\mathbf{U_i}) | \mathbf{Z} = \mathbf{z}\right].$$

*Hence,*

$$n^{1/2} h^{(p+1)/2} \{\hat{\theta}(\mathbf{z}) - \theta_h^*(\mathbf{z})\} \Longrightarrow \mathcal{N}\left(\mathbf{0}, \Sigma(\mathbf{z})^{-1} S(\mathbf{z}) S(\mathbf{z})' \Sigma(\mathbf{z})^{-1})\right),$$

*where* $S(\mathbf{z}) = E\left[\Phi(\mathbf{U}) | \mathbf{Z} = \mathbf{z}\right]$, *and the transpose of a matrix* $A$ *is denoted by* $A'$.

The convergence rate is not the same in (3.5) and (3.6), as expected. Indeed, under the simplifying assumption, no kernel smoothing is used, allowing to achieve $n^{1/2}-$consistency. On the other hand, the asymptotic representation (3.5) shows that the asymptotic behavior of $\hat{\theta}$ depends on the way the margins are estimated (the presence of the term $\sum_{k=1}^{d} \Lambda_i^{(k)}$. It also depends on the correction used to correct the bias caused by censoring (see Proposition 6.5, that shows that $\nu_i^{\Phi}$ depends on the distribution of the censoring).

For the case where the association parameter depends on $\mathbf{z}$, the asymptotic representation is simpler: asymptotically speaking, it is as if smoothing was performed of the true $\mathbf{U}_i$ and as if the censoring distribution were known (recall that $W_{i,n}^*$ is the weight that we would use if we exactly knew the distribution of the censoring). The asymptotic distribution does not depend on the way the margins are estimated. This last property is due to the fact that the rate of estimation of the margins is faster than the convergence rate of $\hat{\theta}(\mathbf{z})$.

Finally, let us note that (3.6) only concern the stochastic part of the error, and does not include the bias term, that is the difference between $\theta_j^*(\mathbf{z})$ and $\theta(\mathbf{z})$. The bias term is covered by the following Proposition 3.2.

**Proposition 3.2** *Under Assumptions 2, 5 and 6,*

$$\sup_{\mathbf{z}} \|\theta_h^*(z) - \theta(\mathbf{z})\| = O(h^2).$$

All proofs are postponed to the appendix section 6.

# 4 Empirical evidence

## 4.1 Simulation setting

### 4.1.1 Framework

In this section, we illustrate the estimation performences of the methodology regarding $\theta(\mathbf{x},t)$ or the corresponding Kendall's tau $\tau(\mathbf{x},t)$, before identifying the resulting improvments in predicting individual claims and portfolio losses distributions. We proceed by designing $B$ run-off portfolios of $N$ claims, with covariates $\mathbf{X}_i^b = \left(X_i^{(1),b}, X_i^{(2),b}, X_i^{(3),b}\right) \overset{i.i.d}{\sim} \mathcal{U}\left([0;1]^3\right)$, for $1 \leq i \leq N$ and $1 \leq b \leq B$. Regarding claims closing delay, we let it follows a log-normal distribution $\log(T_i^b) \overset{i.i.d}{\sim} \mathcal{N}(\alpha_0 + \alpha \mathbf{X}_i^b, 1)$, and faces a Weibull distributed censoring phenomena $C_i^b$, with density $f_{(k,\lambda)}(x) = k\lambda^{-k}x^{k-1}\exp(-(x/\lambda)^k)$, independent of $\left(T_i^b, \mathbf{X}_i^b\right)$. Claim costs $L_i^{(j),b}$ are set as the exponential of a GLM model with Gamma density $f_{(\mu,\nu)}(x) = \Gamma(\nu)^{-1}x^{\nu-1}\nu^\nu\mu^{-\nu}\exp(-x\nu/\mu)$ and log link function, given $\left(T_i^b, \mathbf{X}_i^b\right)$ and independently from $C_i^b$. In other words, for $1 \leq j \leq 2$, $\log(L_i^{(j),b})$ is Gamma distributed with parameters $\mu_i^{(j),b} = \mathbb{E}[\log(L_i^{(j),b})] = \exp(\mu_0^{(j)} + \mu^{(j)}\left(\mathbf{X}_i^b, T_i^b\right))$, and $\nu^{(j)}$. We finally consider Clayton, Frank and Gumbel dependance structures among the margins $(L_i^{(1),b}, L_i^{(2),b})$, according to the same Kendall's tau $\tau((\mathbf{X}_i^b, T_i^b)) = w_3 \tanh(w_2 \tanh(w_1(\mathbf{X}_i, T_i) + b_1) + b_2) + b_3$.

The set of parameters are set as follows :

- regarding claims closing delay, we set $\alpha_0 = 0.2$ and $\alpha = (0.02, 0.04, -0.03)$,
- regarding censoring, we set $k = 1$ and $\lambda = 2$,
- regarding claim costs, we set $\begin{cases} \nu^{(1)} = 5, \mu_0^{(1)} = 0.2, & \mu^{(1)} = (0.02, 0.04, -0.03, 0.005), \\ \nu^{(2)} = 3, \mu_0^{(2)} = 0.3, & \mu^{(2)} = (-0.03, 0.05, 0.01, 0.002), \end{cases}$
- regarding the dependance structure, we set :

$$w_1 = \begin{pmatrix} -5 & 0 & 0 & 0 \\ 0 & -5 & 0 & 0 \\ 0 & 0 & 0 & -5 \end{pmatrix}, w_2 = \begin{pmatrix} 5 & 0 & 0 \\ -5 & 5 & 0 \\ -5 & -5 & 5 \\ -5 & -5 & -5 \end{pmatrix}, \begin{cases} w_3 &= \left(0.25, \frac{1}{12}, \frac{5}{6}, \frac{2}{3}\right), \\ b_1 &= (-0.3, -0.5, 0, -0.5), \\ b_2 &= (0, -1, -2, -2), \\ b_3 &= \frac{13}{24}, \end{cases}$$

Based on those simulation framework settings, we train two estimators of the dependence structure $\tau((\mathbf{X}_i^b, T_i^b))$:

- firstly a parametric estimator $\hat{\tau}_s^b$ that assume the simplified hypothesis, in other words that $\tau((\mathbf{X}_i^b, T_i^b))$ is constant. This Kendall's tau estimator being bijectively linked with the copula parameter estimator $\hat{\theta}_s^b$,
- secondly a semi parametric estimator $\hat{\tau}^b(\mathbf{x}, t)$ that assume the regularity considered in assumptions of this paper. This semi-parametric Kendall's tau estimator being bijectively linked with the copula semi-parameter estimator $\hat{\theta}^b(\mathbf{x}, t)$.

The performances metrics at the root of their relative comparison are introduced in the next Section.

### 4.1.2 Performance metrics

We focus on main expectations regarding the estimators performance metrics definitions : firstly regarding the dependance structure and secondly its impact on claims assessment. The dependance structure of the framework being characterized by either the Kendall's tau and the copula parameter, we introduce the following $L_2$ metrics :

- with regards to the copula parameter :

$$e_{sp}^{\theta} = \frac{1}{B}\sum_{i=1}^{B} ||\hat{\theta}^b(\mathbf{x}, t) - \theta(\mathbf{x}, t)||_2, \; e_s^{\theta} = \frac{1}{B}\sum_{i=1}^{B} ||\hat{\theta}_s^b - \theta(\mathbf{x}, t)||_2, \text{ and } g^{\theta} = \frac{e_{sp}^{\theta} - e_s^{\theta}}{e_s^{\theta}},$$

- regarding the Kendall's tau :

$$e_{sp}^{\tau} = \frac{1}{B}\sum_{i=1}^{B} ||\hat{\tau}^b(\mathbf{x}, t) - \tau(\mathbf{x}, t)||_2, \; e_s^{\tau} = \frac{1}{B}\sum_{i=1}^{B} ||\hat{\tau}_s^b - \tau(\mathbf{x}, t)||_2, \text{ and } g^{\tau} = \frac{e_{sp}^{\tau} - e_s^{\tau}}{e_s^{\tau}}.$$

In addition, we focus on quantile estimation of open claims with $\delta_i^b = 0$, at both claims individual level and at portfolio level. We follow the approach developped in Section 2.5 to simulate realizations according to the estimated and the true setting. Concretely, we begin to sample P realizations $T_i^{(b,p)}$ according to the estimated conditional duration density - conditional because $T_i^{(b,p)}$ must be higher than $C_i^b$. Then, conditionnally on those realizations, we simulate samples of $\hat{\mathbf{L}}_i^{(b,p)}$, from the estimated margins and copula. We therefore deduce samples of total costs $\hat{L}_i^{(1),b} + \hat{L}_i^{(2),b}$, that we write either $L_{s,i}^{(b,p)}$ for the simplified hypothesis copula estimation, either $L_{sp,i}^{(b,p)}$ for the semi parametric copula

estimation. Next, we deduce the respective empirical quantiles of level $\alpha$, $q_\alpha(L^b_{s,i})$ and $q_\alpha(L^b_{sp,i})$ over $p$ for $1 \leq p \leq P$. We finally compare their relevance with regards to the simulated costs at both claims and portfolio levels :

- at the individual level we compute absolute error of quantile predictions :

$$e_s^{(\alpha,r)} = \left| (1-\alpha) - \frac{1}{\sum(1-\delta_i^b)} \sum_{b=1}^{B} \sum_{i=1}^{N} \mathbb{1}_{\left\{\delta_i^b=0\right\} \cap \left\{L^b_{tot,i} < q_\alpha(L^b_{s,i})\right\}} \right|,$$

$$e_{sp}^{(\alpha,r)} = \left| (1-\alpha) - \frac{1}{\sum(1-\delta_i^b)} \sum_{b=1}^{B} \sum_{i=1}^{N} \mathbb{1}_{\left\{\delta_i^b=0\right\} \cap \left\{L^b_{tot,i} < q_\alpha(L^b_{sp,i})\right\}} \right|,$$

$$\text{and obtain the semi-parametric gain } g^{(\alpha,r)} = \frac{e_{sp}^{(\alpha,r)} - e_s^{(\alpha,r)}}{e_s^{(\alpha,r)}},$$

- at the portfolio level we compute absolute error of quantile predictions :

$$E_s^{\alpha} = \left| (1-\alpha) - \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\left\{\sum_{i=1}^{N}(1-\delta_i^b)L^b_{tot,i} < q_\alpha\left(\sum_{i=1}^{N}(1-\delta_i^b)L^b_{s,i}\right)\right\}} \right|,$$

$$E_{sp}^{\alpha} = \left| (1-\alpha) - \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\left\{\sum_{i=1}^{N}(1-\delta_i^b)L^b_{tot,i} < q_\alpha\left(\sum_{i=1}^{N}(1-\delta_i^b)L^b_{sp,i}\right)\right\}} \right|,$$

$$\text{and obtain the semi-parametric gain } G^{(\alpha,r)} = \frac{E_{sp}^{(\alpha,r)} - E_s^{(\alpha,r)}}{E_s^{(\alpha,r)}},$$

where the gains account for the evolution of the quantile error : a reduction if negative - as we expect - and an increase if positive.

### 4.1.3 Results

We simulate data follwing the framework and considering a number of runoff policies of $N = 10000$, a number of portfolios of $B = 500$ and a number of predictions of $P = 1000$. We study two situations of censoring :

- 30% of censoring, for $k = 1$,
- 50% of censoring, for $k = 0.25$.

| Copula | Simplified | | Semi-parametric | | Gains | |
|---|---|---|---|---|---|---|
| | $e_s^\theta$ | $e_s^\tau$ | $e_{sp}^\theta$ | $e_{sp}^\tau$ | $g^\theta$ | $g^\tau$ |
| Clayton | 49 | $9.4\ 10^{-2}$ | 18 | $3.3\ 10^{-3}$ | $-64\%$ | $-96\%$ |
| Frank | 201 | $9.0\ 10^{-2}$ | 59 | $3.1\ 10^{-3}$ | $-71\%$ | $-97\%$ |
| Gumbel | 12 | $9.0\ 10^{-2}$ | 4 | $2.9\ 10^{-3}$ | $-69\%$ | $-97\%$ |

Table 1: Comparison of performances regarding the dependence structure with 30% of censoring ($k = 1$)

| Copula | Simplified | | Semi-parametric | | Gains | |
|---|---|---|---|---|---|---|
| | $e_s^\theta$ | $e_s^\tau$ | $e_{sp}^\theta$ | $e_{sp}^\tau$ | $g^\theta$ | $g^\tau$ |
| Clayton | 49 | $9.4\ 10^{-2}$ | 18 | $3.4\ 10^{-3}$ | $-64\%$ | $-96\%$ |
| Frank | 201 | $9.0\ 10^{-2}$ | 59 | $3.2\ 10^{-3}$ | $-71\%$ | $-96\%$ |
| Gumbel | 12 | $9.0\ 10^{-2}$ | 4 | $3.2\ 10^{-3}$ | $-69\%$ | $-96\%$ |

Table 2: Comparison of performences regarding the dependence structure with 50% of censoring ($k = 0.25$)

Then, we tune the semi-parametric estimation with a bandwidth $h = 0.15$.

The performances regarding the dependence sutructure estimation are gathered in Table 1 for $k = 1$ and in Table 2 for $k = 0.25$. In addition, the prediction performances of the 75% quantile are gathered in Table 3.

The result firstly highlight the relevance of the semi-parametric estimator in the framework, with high reduction of $L_2$ errors for all copula structures, ranging from $-64\%$ to $-97\%$. We remark that the reduction with regards to the copula parameter vary from a copula to another from $-64\%$ to $-71\%$, because the parameters values differs to obtain same Kendall's tau. As expected, this is not the case for $L_2$ Kendall's tau errors with either reductions of $-96\%$ or $-97\%$. We only identify a slight impact of the level of cen-

| Copula | $k = 1$ | | $k = 0.25$ | |
|---|---|---|---|---|
| | $g^{(\alpha,r)}$ | $G^{(\alpha,r)}$ | $g^{(\alpha,r)}$ | $G^{(\alpha,r)}$ |
| Clayton | $-67\%$ | $-2\%$ | $-99\%$ | $+0\%$ |
| Frank | $-99\%$ | $-3\%$ | $-98\%$ | $+10\%$ |
| Gumbel | $-15\%$ | $-2\%$ | $-16\%$ | $+6\%$ |

Table 3: Comparison of performances regarding the prediction of quantile of level $\alpha = 75\%$

| Copula | $h = 0.05$ | $h = 0.10$ | $h = 0.15$ | $h = 0.20$ |
|---|---|---|---|---|
| Clayton | $-87\%$ | $-92\%$ | $-64\%$ | $-40\%$ |
| Frank | $-92\%$ | $-94\%$ | $-71\%$ | $-49\%$ |
| Gumbel | $-83\%$ | $-93\%$ | $-69\%$ | $-46\%$ |

Table 4: Comparison of performences regarding the bandwidth of $g^{\theta}$.

| Copula | $h = 0.05$ | $h = 0.10$ | $h = 0.15$ | $h = 0.20$ |
|---|---|---|---|---|
| Clayton | $-75\%$ | $-67\%$ | $-63\%$ | $-17\%$ |
| Frank | $-91\%$ | $-99\%$ | $-81\%$ | $-58\%$ |
| Gumbel | $-99\%$ | $-15\%$ | $-10\%$ | $-10\%$ |

Table 5: Comparison of performences regarding the bandwidth of $g^{(\alpha,r)}$ with $\alpha = 70\%$.

soring on the performances of the estimators, with a decrease of the performances with the raise from 30% to 50% of the level of censoring.

Finally, the analysis of the 75% quantile prediction performances highlights the relevance of the semi-parametric estimator at the claims level that is neutralized at the portfolio level.

#### 4.1.4 Bandwidth sensibility

We also study the impact of the bandwidth on both the dependance structure and the quantile prediction in Tables 4 and 5. It highlights the sensibility of the semi-parameteric tuning on the errors reduction. Globally the higher performances are obtained with $h = 0.10$.

### 4.2 Real data example

As the methodology demonstrates its legitimity on simulations, we illustrate its predictions on a real data exemple. We consider a dataset hosted by the Texas Departmment of Insurance, that gathers 40 868 injury claims, closed between 2007 and 2012. The dataset is publicly available at `https://www.tdi.texas.gov/`. The cost of each claim $L_{(tot,i)}$ is composed by the indemnity for the injured party $L_i^{(1)}$ and the expenses linked to trials $L_i^{(2)}$. For simplicity, we focus on claims with impacts on both margins, resulting to 21 680 claims. For a global approach, the following models may be plugged to a mixture model, in order to incorporate models of claims with univariate costs. Moreover, we restrict our-

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $T$ | $L^{(1)}$ | $L^{(2)}$ |
|---|---|---|---|---|---|---|---|
| Mean | 176 | 43 | 56 882 | 9 944 | 744 | 163 397 | 38 379 |
| $q_{0.25}$ | 1 | 38 | 5 000 | 115 | 418 | 25 000 | 2 000 |
| $q_{0.5}$ | 6 | 43 | 14 875 | 2 500 | 682 | 60 000 | 12 843 |
| $q_{0.75}$ | 133 | 49 | 42 375 | 10 000 | 1009 | 175 500 | 41 000 |

Table 6: Descriptive statistics on closed claims.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ | $T$ | $L^{(1)}$ | $L^{(2)}$ |
|---|---|---|---|---|---|---|---|---|
| Mean | 169 | 44 | 68 297 | 11 880 | 598 | 933 | 221 684 | 53 486 |
| $q_{0.25}$ | 1 | 40 | 6 000 | 103 | 254 | 621 | 50 000 | 7 000 |
| $q_{0.5}$ | 6 | 43 | 16 000 | 5 000 | 511 | 857 | 100 000 | 22 698 |
| $q_{0.75}$ | 138 | 49 | 50 000 | 12 993 | 844 | 1167 | 247 500 | 56 000 |

Table 7: Descriptive statistics on open claims. The informations of $T$, $L^{(1)}$ and $L^{(2)}$ are assumed to be unknown on the 31/12/2010.

selves to information available at the 31/12/2010, corresponding to 15 717 closed claims and 4 917 open claims, with the aim to assess the costs of open claims.

We rely on 4 covariates to predict open claims costs : the delay between the injury and the insurance report $X_1$, the age of the injured party $X_2$, the initial estimation of the indemnity $X_3$, and the initial estimation of the expenses $X_4$. Descriptive statistics of the dataset are given in Tables 7 and 6. As expected, the closing delay $T$ is globally inferior on closed claims than on open claims. The distributions of covariates $X_1$ and $X_2$ are quite similar whereas the first estimations, as the final costs, are higher for closed claims than open ones. In addition, the Kendall's tau between $L^{(1)}$ and $L^{(2)}$ is higher on closed claims (0.40) than for open claims (0.30).

Regarding the margins, we begin to calibrate models on the logarithm of both $T$, $L^{(1)}$ and $L^{(2)}$. We considered several models, and their analysis highlighted in Section 6.5, lead to the following choices. For the duration $T$, we consider a GLM with Weibull distributions defined by the density $f_{(\lambda,k)}(x) = kx^{k-1}\lambda^{-k}e^{-(x/\lambda)^k}$, under a fixed shape parameter $k = \alpha_k$, letting $\mathbb{E}[\log(T_i)] = \lambda_i\Gamma(1 + 1/\alpha_k) = \exp(\alpha_0 + \alpha\mathbf{X}_i)$. Regarding $L^{(1)}$, we chose a GLM model with Gamma distribution with $\mathbb{E}[\log(L_i^{(1)})] = \exp(\beta_0^{(1)} + \beta^{(j)}(\mathbf{X}_i, Y_i))$, and $\mu = \beta_\mu$. Finally, for $L^{(2)}$, we select a GLM with Weibull distributions under a fixed shape parameter $k = \gamma_k$, letting $\mathbb{E}[\log(L^{(2)})] = \lambda_i\Gamma(1 + 1/\gamma_k) = \exp(\gamma_0 + \gamma\mathbf{X}_i)$. The estimated parameters, given in Table 8, allow to obtain bivariate pseudo observations of $\left(L^{(1)}, L^{(2)}\right)$ illustrated

| | | | |
|---|---|---|---|
| $T$ | $\alpha_0 = 1.864$ | $\alpha = (3.973\ 10^{-3}, 1.652\ 10^{-4}, -4.373\ 10^{-4}, 2.273\ 10^{-3})$ | $\alpha_k = 11.77$ |
| $L^{(1)}$ | $\beta_0 = 2.201$ | $\beta = (3.308\ 10^{-3}, -1.652\ 10^{-4}, 1.650\ 10^{-2}, 3.226\ 10^{-3}, 3.884\ 10^{-5})$ | $\beta_\nu = 4.459$ |
| $L^{(2)}$ | $\alpha_0 = 1.892$ | $\alpha = (1.672\ 10^{-2}, -2.272\ 10^{-4}, 8.342\ 10^{-3}, 1.104\ 10^{-2})$ | $\gamma_k = 5.195$ |

Table 8: Estimated parameters for univariate models for $T$, $L^{(1)}$ and $L^{(2)}$

in Figure 1.

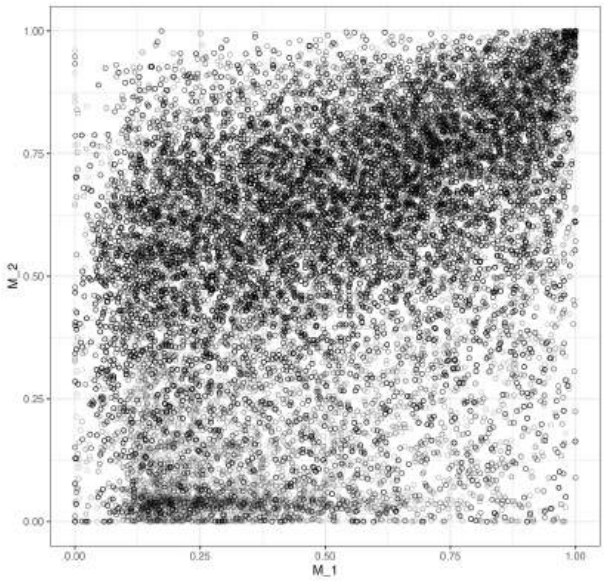
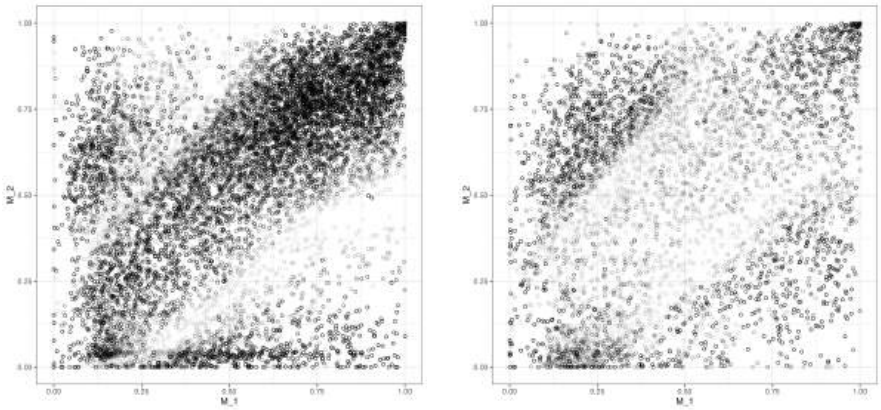



Figure 1: Pseudo observations $\left(L^{(1)}, L^{(2)}\right)$, with IPCW weighted transparency.

Then, based on pseudo observations $(\mathbf{U}_i)_{1\leq i\leq n}$, we calibrate Clayton, Frank and Gumbel copulas $\mathfrak{C}_\theta$. For each family, we considered both the simplified $\theta_0$ and the semiparametric $\theta(\mathbf{x})$ estimation. For the latter, we consider the multivariate Gaussian density for the Kernel, with a variance $h = 0.1$, selected using $3-$folds cross validation as highlighted in Section 6.5.2. We compare estimators regarding their log-likelihood performances. The obtained statitics are gathered in Table 9. We also illustrate the log-likelihood differences on pseudo observations in Figure 2, to understand dependence areas involved in the semi parametric flexibility. Globally, the the semi parametric framework demonstrates its flexibility and the Clayton copula turns out to be the more appropriate family to combine with the framework.

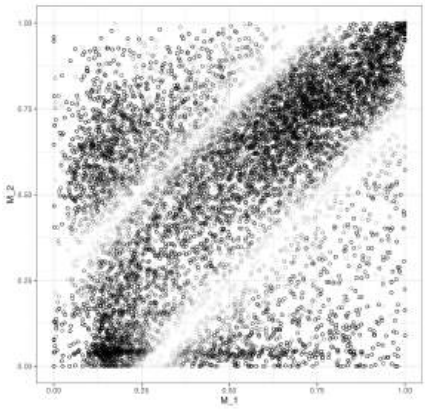
Finally, we assess the distribution of open claims costs as follows. We begin by sampling $P = 1000$ closing delays for each claim $i$, according to the law $\frac{\int_{Y_i}^{\infty} d\hat{F}^T(t|\mathbf{x})}{\hat{S}^T(Y_i|\mathbf{x})}$. Indeed, a partial information about the claim $i$ rely in the closing delay lower bound $Y_i$. Then,



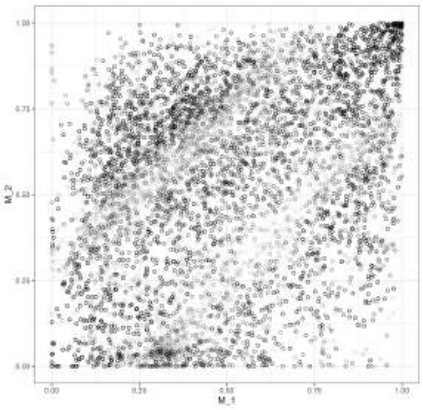
(a) Clayton '+'



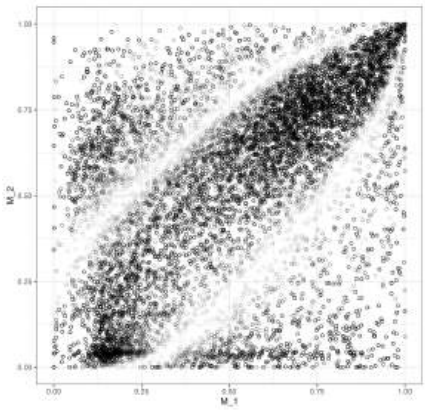
(b) Clayton '-'



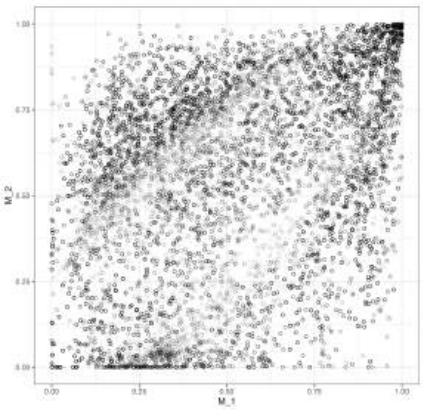
(c) Frank '+'



(d) Frank '-'



(e) Gumbel '+'



(f) Gumbel '-'

Figure 2: Log-likelihood differences between semi parametric and simplified models on pseudo observations for each copula family. Graphics at the left hand side gather pseudo observations associated with higher likelihood for semi parametric models than for simplified models. At the opposite at the right hand side, graphics gather pseudo observations associated with lower likelihood for semi parametric models than for simplified models. Transparency is higher for high differences, and IPCW weighted.

| | Simplified | Semi parametric |
|---|---|---|
| Clayton | $4.335\ 10^{-2}$ | $\mathbf{1.991\ 10^{-1}}$ |
| Frank | $1.108\ 10^{-1}$ | $1.516\ 10^{-1}$ |
| Gumbel | $\mathbf{1.354\ 10^{-1}}$ | $1.715\ 10^{-1}$ |

Table 9: Comparison of simplified and semi-parametric models for each copula family, in terms of empirical log-likelihood.

| Copula | 25% | 50% | 75% | 90% | 95% | 99% |
|---|---|---|---|---|---|---|
| Clayton s | $4.68\ 10^{9}$ | $7.41\ 10^{9}$ | $1.35\ 10^{10}$ | $4.94\ 10^{10}$ | $5.23\ 10^{11}$ | $6.18\ 10^{13}$ |
| Clayton sp | $5.09\ 10^{9}$ | $8.11\ 10^{9}$ | $2.42\ 10^{10}$ | $6.68\ 10^{10}$ | $1.18\ 10^{11}$ | $1.56\ 10^{13}$ |
| Frank s | $5.18\ 10^{9}$ | $7.20\ 10^{9}$ | $1.62\ 10^{10}$ | $6.88\ 10^{10}$ | $1.45\ 10^{11}$ | $9.26\ 10^{11}$ |
| Frank sp | $4.75\ 10^{9}$ | $7.06\ 10^{9}$ | $1.57\ 10^{10}$ | $4.18\ 10^{10}$ | $1.88\ 10^{11}$ | $2.76\ 10^{15}$ |
| Gumbel s | $4.62\ 10^{9}$ | $8.32\ 10^{9}$ | $2.37\ 10^{10}$ | $9.04\ 10^{10}$ | $2.51\ 10^{11}$ | $1.72\ 10^{13}$ |
| Gumbel sp | $5.07\ 10^{9}$ | $6.91\ 10^{9}$ | $1.43\ 10^{10}$ | $4.13\ 10^{10}$ | $1.39\ 10^{11}$ | $1.13\ 10^{12}$ |

Table 10: Quantiles predictions of the total cost of open claims, depending on copula famyly and simplified or semi-parametric framework. The empirical total cost is equal to $1.35\ 10^{9}$.

conditionnally on $(\mathbf{X}_i, T_{i,p})$ , we simulate costs $\left(L_{i,p}^{(1)}, L_{i,p}^{(2)}\right)$ according to estimated models, both regarding margins and the dependance structure. Therefore, we deduce a distribution of total costs at the individual level $L_{tot,i}$ and regarding the portfolio $L_{tot}$. Resulting statistics given in Table 10 must be compared to the observed cost of open claims equal to $1.35\ 10^{9}$. Clearly all models lead to higher estimations but the context of high volatility nuance those disappointing prediction performances. In the same time, the confidence intervals resulting from the model allow to rationalize this volatility.

In addition, we assess the performance of the model at the individual level of claims, withing the following metrics :

$$g_s^{(\alpha,a)} = \frac{1}{\sum(1-\delta_i^b)} \sum_{i=1}^{B} \sum_{j=1}^{N} \mathbb{1}_{\{\delta_i^b=0\}} \left( |L_i^{(1),b} - q_{50\%}(L_{s,i}^{(1),b})| + |L_i^{(2),b} - q_{50\%}(L_{s,i}^{(2),b})| \right),$$

$$g_{sp}^{(\alpha,a)} = \frac{1}{\sum(1-\delta_i^b)} \sum_{i=1}^{B} \sum_{j=1}^{N} \mathbb{1}_{\{\delta_i^b=0\}} \left( |L_i^{(1),b} - q_{50\%}(L_{sp,i}^{(1),b})| + |L_i^{(2),b} - q_{50\%}(L_{sp,i}^{(2),b})| \right).$$

| Copula | $g_s^{(\alpha,a)}$ | $g_{sp}^{(\alpha,a)}$ | $g_s^{(\alpha,r)}$ | $g_{sp}^{(\alpha,r)}$ |
|---|---|---|---|---|
| Clayton | $2.03\ 10^5$ | $8.01\ 10^4$ | $6.26\ 10^{-1}$ | $6.24\ 10^{-1}$ |
| Frank | $2.02\ 10^5$ | $8.05\ 10^4$ | $6.22\ 10^{-1}$ | $6.27\ 10^{-1}$ |
| Gumbel | $2.03\ 10^5$ | $7.92\ 10^4$ | $6.32\ 10^{-1}$ | $6.24\ 10^{-1}$ |

Table 11: Comparison of prediction performences at the individual level regarding $g^{(\alpha,a)}$ and $g^{(\alpha,r)}$ for the simplified and the semi parametric framework.

$$g_s^{(\alpha,r)} = \frac{1}{\sum(1-\delta_i^b)} \sum_{i=1}^{B} \sum_{j=1}^{N} \mathbb{1}_{\{\delta_i^b=0\}} \left( \frac{|L_i^{(1),b} - q_{50\%}(L_{s,i}^{(1),b})| + |L_i^{(2),b} - q_{50\%}(L_{s,i}^{(2),b})|}{L_i^{(1),b} + L_i^{(2),b}} \right),$$

$$g_{sp}^{(\alpha,r)} = \frac{1}{\sum(1-\delta_i^b)} \sum_{i=1}^{B} \sum_{j=1}^{N} \mathbb{1}_{\{\delta_i^b=0\}} \left( \frac{|L_i^{(1),b} - q_{50\%}(L_{sp,i}^{(1),b})| + |L_i^{(2),b} - q_{50\%}(L_{sp,i}^{(2),b})|}{L_i^{(1),b} + L_i^{(2),b}} \right).$$

The resulting statistics are given in Table 11. The absolute errors at the claims level turn out to be lower under the semi-parametric estimations, which is approximately neutralized with the relative point of view. The flexibility of the semi-parametric and its ability to fit the dependance between the prejudice and the fees therefore allow to better assess the losses for serious claims. This highlight the potential of claims monitoring applications at the micro level.

# 5 Conclusion

In this paper, we considered the case of analyzing the cost of claims in the case where this cost is decomposed in several lines of payment. In the framework we develop, the interaction between these lines of payment is modeled using copulas. Moreover, we take into account the link between the "lifetime of the claim" (time before settlement) and its final cost, with the correction of the bias caused by censoring. This decomposition can be useful to better understand what drives the severity of a claim, and potentially to improve risk management by winning on some lines (like legal fees or expert costs) that can be more easily controlled. Let us also mention another possible application to the proper design of insurance products, through the example of cyber insurance. Due to the variety of situations in cyber risk (data theft, business interruption, legal dispute after a data breach, corporal prejudice...), an important question is to define which will be

covered or not by an insurance product. **?** show some analysis of some products available on the US market, showing that their nature may be strongly differ. Carefully analyzing the different components of the loss is essential to better understand which (and at which level) some types of losses can be covered or not.

# 6 Appendix

The Appendix section is organized as follows. In Section 6.1, we show the convergence of the bias term of our kernel estimator. Preliminary results to show Theorem 3.1 are gathered in Section 6.2. Asymptotic representations of Kaplan-Meier intergrals with pseudo-observations, which are required to study the estimator used under the simplifying assumption, are given in Section 6.3. The proof of Theorem 3.1 is then given in Section 6.4. Finally, Section 6.5 motivates the model choices assumed for the real data analysis performed in Section 4.2.

## 6.1 Proof of Proposition 3.2

Let

$$M_{\mathbf{z}}^*(\theta) = E\left[\log \mathfrak{c}_\theta(\mathbf{U}) K\left(\frac{\mathbf{Z}-\mathbf{z}}{h}\right)\right].$$

We have $\theta_h^*(\mathbf{z}) = \arg\max_\theta M_{\mathbf{z}}^*(\theta)$. Since $\nabla_\theta M_{\mathbf{z}}^*(\theta(\mathbf{z})) = 0$, a second order Taylor expansion yields

$$\theta_h^*(\mathbf{z}) - \theta(\mathbf{z}) = -\nabla_\theta^2 M_{\mathbf{z}}^*(\tilde{\theta})^{-1} \nabla_\theta M_{\mathbf{z}}^*(\theta(\mathbf{z})), \tag{6.1}$$

where each component of $\tilde{\theta}$ is between the corresponding components of $\theta_h^*(\mathbf{z})$ and $\theta(\mathbf{z})$.

On the other hand,

$$\begin{aligned}
\nabla_\theta M_{\mathbf{z}}^*(\theta) &= E\left[\frac{\nabla_\theta \mathfrak{c}_\theta(\mathbf{U})}{\mathfrak{c}_\theta(\mathbf{U})} K\left(\frac{\mathbf{Z}-\mathbf{z}}{h}\right)\right] \\
&= \int\int \frac{\nabla_\theta \mathfrak{c}_\theta(\mathbf{u})}{\mathfrak{c}_\theta(\mathbf{u})} f_{\mathbf{U},\mathbf{Z}}(\mathbf{u}, \mathbf{z}+vh) K(v) dv d\mathbf{u} \\
&= hE\left[\frac{\nabla_\theta \mathfrak{c}_\theta(\mathbf{U})}{\mathfrak{c}_\theta(\mathbf{U})} | \mathbf{Z} = \mathbf{z}\right] f_{\mathbf{Z}}(\mathbf{z}) + h^3 \int\int \frac{\nabla_\theta \mathfrak{c}_\theta(\mathbf{u}) \mathbf{v}' \nabla_{\mathbf{z}}^2 f_{\mathbf{U},\mathbf{Z}}(\mathbf{u}, \tilde{\mathbf{z}}(\mathbf{v})) \mathbf{v}}{\mathfrak{c}_\theta(\mathbf{u})} K(\mathbf{v}) d\mathbf{v} d\mathbf{u},
\end{aligned}$$

where $f_{\mathbf{Z}}$ denotes the density of $\mathbf{Z}$. From (3.4) in Assumption 6, and the fact that

$$E\left[\frac{\nabla_\theta \mathfrak{c}_{\theta(\mathbf{z})}(\mathbf{U})}{\mathfrak{c}_{\theta(\mathbf{z})}(\mathbf{U})} | \mathbf{Z} = \mathbf{z}\right] = 0,$$

we get $\sup_{\mathbf{z}} \|\nabla_\theta M_{\mathbf{z}}^*(\theta(\mathbf{z}))\|_\infty = O(h^3)$.

Moreover,

$$\nabla_\theta^2 M_{\mathbf{z}}^*(\theta) = h \int \dot{\Phi}(\mathbf{u}) f_{\mathbf{U},\mathbf{Z}}(\mathbf{u}, \mathbf{z} + vh) K(v) dv d\mathbf{u}.$$

From (3.3) in Assumption, we get $\sup_{\mathbf{z},\theta\in\Theta} \|\nabla_\theta^2 M_{\mathbf{z}}^*(\theta)\|_\infty = O(h)$. The result follows from (6.1).

## 6.2 Technical results

The following three propositions are required to prove Theorem 3.1. They are stated in the context of kernel estimators, but similar results can be obtained for non-kernel estimators (used if one is working under the simplifying assumption), replacing $1/h^{p+1}$ and $K$ by the constant 1.

**Proposition 6.1** *Let*

$$\Delta^\chi I_n(\psi, \mathbf{z}) = \frac{1}{h^{p+1}} \sum_{i=1}^n W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \chi(Y_i) \left\{\psi(\hat{\mathbf{U}}_i) - \psi(\mathbf{U}_i)\right\},$$

*for $\psi$ in a set of functions $\mathcal{F}$, and let $\dot{\psi}$ denote the gradient vector of $\psi$. Assume that the $k-$th component $\dot{\psi}^{(k)}$ of $\dot{\psi}$ is bounded by,*

$$\dot{\psi}^{(k)}(\mathbf{u}) \leq \tilde{\mathfrak{p}}^{(k)}(u^{(k)}) \prod_{j\neq k} \mathfrak{p}^{(j)}(u^{(j)}), \tag{6.2}$$

*with $\tilde{\mathfrak{p}}^{(k)}$ and $\mathfrak{p}^{(k)} \in \mathcal{R}$ for all $k = 1, ..., d$. Let*

$$\mathfrak{q}(\mathbf{u}) = \begin{pmatrix} \mathfrak{q}^{(1)}(u^{(1)}) \\ \vdots \\ \mathfrak{q}^{(d)}(u^{(d)}) \end{pmatrix}, \quad \bar{\mathfrak{q}}(\mathbf{u}) = \begin{pmatrix} 1/\mathfrak{q}^{(1)}(u^{(1)}) \\ \vdots \\ 1/\mathfrak{q}^{(d)}(u^{(d)}) \end{pmatrix},$$

*where $\mathfrak{q}^{(k)} \in \mathcal{Q}$ for $k = 1, ..., d$. Assume that*

$$\sup_{1\leq i\leq n} \|\bar{\mathfrak{q}}(\mathbf{U}_i)'(\hat{\mathbf{U}}_i - \mathbf{U}_i)\|_\infty = O_P(\varepsilon_n), \tag{6.3}$$

*for some sequence $\varepsilon_n$ tending to 0, and that*

$$\sup_{k=1,...,d} E\left[\chi(T)\mathfrak{q}^{(k)}(U^{(k)})\tilde{\mathfrak{p}}^{(k)}(U^{(k)}) \prod_{j\neq k} \mathfrak{p}^{(j)}(U^{(j)})\right] < \infty. \tag{6.4}$$

*then, under Assumption 3 equation (3.1),*

$$\sup_{\psi\in\mathcal{F}} \Delta^\chi I_n(\psi, \mathbf{z}) = O_P(\varepsilon_n).$$

**Proof.** From a first order Taylor expansion,

$$\Delta^{\chi} I_n(\psi, \mathbf{z}) = \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^{*} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \chi(Y_i) \dot{\psi}(\tilde{\mathbf{U}}_i)'[\hat{\mathbf{U}}_i - \mathbf{U}_i],$$

where the $k-$th component of the vector $\tilde{\mathbf{U}}_i$ is between $\hat{U}_i^{(k)}$ and $U_i^{(k)}$. For $M \geq 0$, let

$$E_{M,n} = \bigcap_{k=1}^{d} \left\{ \inf_{1 \leq i \leq n} \inf \left( \frac{\hat{U}_i^{(k)}}{U_i^{(k)}}, \frac{1 - \hat{U}_i^{(k)}}{1 - U_i^{(k)}} \right) \geq \frac{1}{M} \right\}.$$

On the event $E_{M,n}$,

$$\|\mathfrak{q}(\mathbf{U}_i)' \dot{\psi}(\tilde{\mathbf{U}}_i)\|_1 \leq M^d \sum_{k=1,\ldots,d} \mathfrak{q}^{(k)}(U_i^{(k)}) \tilde{\mathfrak{p}}^{(k)}(U_i^{(k)}) \prod_{j \neq k} \mathfrak{p}^{(j)}(U_i^{(j)}), \tag{6.5}$$

where $\|\cdot\|_1$ denotes the sum of the absolute values of the coefficients of a vector. Moreover,

$$\frac{|\Delta^{\chi} I_n(\psi, \mathbf{z})|}{\sup_{1 \leq i \leq n} \|\bar{\mathfrak{q}}(\mathbf{U}_i)'(\hat{\mathbf{U}}_i - \mathbf{U}_i)\|_{\infty}} \leq \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^{*} |K| \left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \chi(Y_i) \|\mathfrak{q}(\mathbf{U}_i)' \dot{\psi}(\tilde{\mathbf{U}}_i)\|_1. \tag{6.6}$$

It follows from (6.5), (6.4), and the fact that $\dot{\psi} \in \mathcal{R}$, that the term on the right-hand side of (6.6) has its expectation bounded by $M^d \times M'$ where $M' \geq 0$ is a constant independent from $M$ and $n$. Since $\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}(E_{M,n}) = 0$ from Assumption 3 equation (3.1), we get

$$\sup_{\psi} \frac{|\Delta^{\chi} I_n(\psi, \mathbf{z})|}{\sup_{1 \leq i \leq n} \|\bar{\mathfrak{q}}(\mathbf{U}_i)'(\hat{\mathbf{U}}_i - \mathbf{U}_i)\|_{\infty}} = O_P(1).$$

The result follows from (6.3). ■

**Proposition 6.2** *Let $\mathcal{F}$ denote a class of functions with envelope $\Psi$ such that*

$$N_{[]}(\varepsilon, \mathcal{F}) \leq \frac{A}{\varepsilon^k},$$

*and with*

$$E\left[\frac{\Psi(\mathbf{U}_i)^2}{S^C(T_i)}\right] < \infty, \tag{6.7}$$

$$E\left[\frac{\Psi(\mathbf{U}_i)^2}{\mathfrak{K}(T_i)^{1/2+\iota}}\right] < \infty. \tag{6.8}$$

*Then, under the Assumptions required to obtain (6.9) in Proposition 6.3,*

$$\sup_{\psi \in \mathcal{F}} \left| \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \psi(\hat{\mathbf{U}}_i) - E\left[K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \psi(\mathbf{U}_i)\right] \right| = O_P\left(\varepsilon_n + \frac{[\log n]^{1/2}}{n^{-1/2} h^{-(p+1)/2}}\right).$$

**Proof.** Write

$$\frac{1}{h^{p+1}}\sum_{i=1}^{n} W_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right)\psi(\mathbf{U}_i) = \frac{1}{h^{p+1}}\sum_{i=1}^{n} W^*_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right)\psi(\mathbf{U}_i) + \Delta^1 I_n(\psi, \mathbf{z}) + R_n(\psi, \mathbf{z}).$$

From Proposition 6.1, $\sup_{\psi\in\mathcal{F}} |\Delta^1 I_n(\psi, \mathbf{z})| = O_P(\varepsilon_n)$, while, from (6.9) in Proposition 6.3, $\sup_{\psi\in\mathcal{F}} |R_n(\psi, \mathbf{z})| = O_P(n^{-1/2})$. The result follows from Theorem 4 in Einmahl et al. (2005). ■

**Proposition 6.3** *Let*

$$R_n(\psi, \mathbf{z}) = \frac{1}{h^{p+1}}\sum_{i=1}^{n} (W_{i,n} - W^*_{i,n}) K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right)\psi(\hat{\mathbf{U}}_i),$$

*for $\psi \in \mathcal{F}$. Assume that the conditions of Proposition 6.1 hold for $\psi$ and $\chi(t) = 1/\mathfrak{K}(t)^{1/2+\iota}$, for some $\iota > 0$. Then*

$$\sup_{\psi\in\mathcal{F}} |R_n(\psi, \mathbf{z})| = O_P(n^{-1/2}). \tag{6.9}$$

*If the conditions of Proposition 6.1 hold for $\psi$ and $\chi(t) = S^C(t)^{-\iota}$, then*

$$\sup_{\psi\in\mathcal{F}} |R_n(\psi, \mathbf{z})| = o_P(1). \tag{6.10}$$

**Proof.** Observe that

$$\begin{aligned} |R_n(\psi, \mathbf{z})| \;\leq\; & \sup_{t\leq Y_{(n)}} \left\{ \mathfrak{K}(t)^{1/2+\iota} \frac{|\hat{S}^C(t) - S^C(t)|}{S^C(t)} \right\} \times \sup_{t\leq Y_{(n)}} \left| \frac{S^C(t)}{\hat{S}^C(t)} \right| \\ & \times \frac{1}{h^{p+1}}\sum_{i=1}^{n} W^*_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \frac{\psi(\hat{\mathbf{U}}_i)}{\mathfrak{K}(Y_i)^{1/2+\iota}}. \end{aligned}$$

From Proposition 6.1,

$$\frac{1}{h^{p+1}}\sum_{i=1}^{n} W^*_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \frac{\psi(\hat{\mathbf{U}}_i)}{\mathfrak{K}(Y_i)^{1/2+\iota}} = \frac{1}{h^{p+1}}\sum_{i=1}^{n} W^*_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \frac{\psi(\mathbf{U}_i)}{\mathfrak{K}(Y_i)^{1/2+\iota}} + o_P(1).$$

From (6.2) and (6.4), the supremum (over $\psi \in \mathcal{F}$) of the first term in the right-hand side of the last display has a bounded expectation. This shows that

$$\sup_{\psi\in\mathcal{F}} \left| \frac{1}{h^{p+1}}\sum_{i=1}^{n} W^*_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \frac{\psi(\hat{\mathbf{U}}_i)}{\mathfrak{K}(Y_i)^{1/2+\iota}} \right| = O_P(1).$$

Equation (6.9) follows from Theorem 2.1 in Gill (1983), which gives

$$\sup_{t\leq Y_{(n)}} \left\{ \mathfrak{K}(t)^{1/2+\iota} \frac{|\hat{S}^C(t) - S^C(t)|}{S^C(t)} \right\} = O_P(n^{-1/2}),$$

and Lemma 2.6 in Gill (1983) which ensures that

$$\sup_{t\leq Y_{(n)}} \frac{S^C(t)}{\hat{S}^C(t)} = O_P(1).$$

To obtain (6.10), we use the same decomposition but with $\mathfrak{K}^{1/2+\iota}$ replaced by $(S^C)^\iota$. In which case, we use again Lemma 2.6 in Gill (1983) to obtain

$$\sup_{t\leq Y_{(n)}} \left\{ \frac{|\hat{S}^C(t) - S^C(t)|}{S^C(t)} \right\}^\iota = O_P(1),$$

while $\sup_{t\leq Y_{(n)}} |\hat{S}^C(t) - S^C(t)|^{1-\iota} = o_P(1)$, from Stute and Wang (1993), and the result follows.

**Remark 6.4** *Theorem 2.1 in Gill (1983) shows that*

$$\sup_{t\leq Y_{(n)}} \left\{ h(t) \frac{|\hat{S}^C(t) - S^C(t)|}{S^C(t)} \right\} = O_P(n^{-1/2}),$$

*provided that* $\int h^2(t) d(\mathfrak{K}^{-1}(t)) < \infty$. *Here,* $h(t) = \mathfrak{K}(t)^{1/2+\iota}$, *so that the condition holds. This the reason for introducing this function* $\mathfrak{K}$ *in the integrability conditions required to obtain our results.*

■

## 6.3 Asymptotic representation of weighted sums with pseudo-observations

**Proposition 6.5** *Let* $\phi$ *be a function such that*

$$E\left[ \frac{\phi(\mathbf{U}_i, T_i, \mathbf{X}_i)^2}{S_C(T_i)} \right] < \infty,$$

*and*

$$E\left[ |\phi(T)| \mathfrak{K}^{-1/2-\iota}(T) \right] < \infty,$$

*for some* $\iota > 0$. *Then, under Assumption 4,*

$$\sum_{i=1}^n W_{i,n} \phi(\hat{\mathbf{U}}_i) = \frac{1}{n} \sum_{i=1}^n \nu_i(\phi) + \frac{1}{n} \sum_{i=1}^n \tilde{\psi}^{(k)}(\mathbf{L}_i, \mathbf{Z}_i, \delta_i) + o_P(n^{-1/2}),$$

*where*

$$\nu_i(\phi) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \phi(\mathbf{U}_i, Y_i, \mathbf{X}_i)}{S_C(Y_i)} + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{(1-\delta_i)\bar{\phi}(Y_i)}{S_Y(Y_i)} - \int_{-\infty}^{Y_i} E\left[ \frac{\mathbf{1}_{C\leq t} \bar{\phi}(C)}{S_C(C) S_Y(C)} \right] \right\} dt,$$

*with* $S_Y(y) = \mathbb{P}(Y \geq y)$ *and*

$$\bar{\phi}(y) = E\left[\phi(\mathbf{U}, T, \mathbf{X})\mathbf{1}_{T \geq y}\right],$$

*and*

$$\tilde{\psi}^{(k)}(\mathbf{l}, \mathbf{z}, \mathfrak{d}) = E\left[\eta^{(k)}_{\mathbf{L}_1, \mathbf{Z}_1}(\mathbf{L}_2, \mathbf{Z}_2, \delta_2)\psi(\mathbf{U}_2)|\mathbf{L}_2 = \mathbf{l}, \mathbf{Z}_2 = \mathbf{z}, \delta_2 = \mathfrak{d}\right].$$

**Proof.** The proof is similar to the one of Proposition A.5 in Lopez (2019), but with some differences caused by the fact that we rely on Kaplan-Meier estimator. Essentially, the difference comes the fact that we do not have $\sup_t |\hat{S}_C(t) - S_C(t)| = O_P(n^{-1/2})$ due to the erratic behavior of the Kaplan-Meier near the tail (under the assumptions on the tail of the distribution that we consider here). The "trick" is to multiply $|\hat{S}_C(t) - S_C(t)|$ by a function tending to zero (namely $\mathfrak{K}(t)^{1/2+\iota}$ so that we can achieve the proper convergence rate.

The proof relies on the following decomposition,

$$\sum_{i=1}^{n} W_{i,n}\phi(\hat{\mathbf{U}}_i) = \sum_{i=1}^{n} W_{i,n}\phi(\mathbf{U}_i) + \sum_{j=1}^{d}\sum_{i=1}^{n} W^*_{i,n}[\hat{U}^{(j)}_i - U^{(j)}_i]\dot{\phi}^{(j)}(\mathbf{U}_i) \tag{6.11}$$

$$+ \sum_{k=1}^{d}\sum_{i=1}^{n}[W_{i,n} - W^*_{i,n}]\frac{[\hat{U}^{(k)}_i - U^{(k)}_i]}{\mathfrak{q}^{(k)}(U^{(k)}_i)}\mathfrak{q}^{(k)}(U^{(k)}_i)\dot{\phi}^{(k)}(\tilde{\mathbf{U}}^{(k)}_i) \tag{6.12}$$

$$+ \sum_{k=1}^{d}\sum_{i=1}^{n}[W_{i,n} - W^*_{i,n}]\frac{[\hat{U}^{(k)}_i - U^{(k)}_i]}{\mathfrak{q}^{(k)}(U^{(k)}_i)}\mathfrak{q}^{(k)}(U^{(k)}_i)\left\{\dot{\phi}^{(k)}(\tilde{\mathbf{U}}^{(k)}_i) - \dot{\phi}^{(k)}(\mathbf{U}_i)\right\}, \tag{6.13}$$

with $\tilde{\mathbf{U}}_{i,k}$ has the same components as $\mathbf{U}_i$ except for the $k-$th component which is between $U^{(k)}_i$ and $\hat{U}^{(k)}_i$.

The first two terms of (6.11) are covered by Lemma 6.6 and 6.7 respectively. The last two terms are higher order terms that are $o_P(n^{-1/2})$. To see that, note that

$$|W_{i,n} - W^*_{i,n}| = \frac{W^*_{i,n}}{\mathfrak{K}(Y_i)^{1/2-\iota}} \times \mathfrak{K}^{1/2-\iota}(Y_i)\left|\frac{\hat{S}_C(Y_i) - S_C(Y_i)}{\hat{S}_C(Y_i)}\right| = \frac{W^*_{i,n}}{\mathfrak{K}(Y_i)^{1/2-\iota}} \times O_P(n^{1/2}),$$

where the $O_P(n^{-1/2})$ holds uniformly over $i$, and where the $O_P$ rate comes from Theorem 2.1 in Gill (1983). Hence, we could show that (6.12) and (6.13) are $o_P(n^{-1/2})$ using Assumption 2, using the fact that $\{\dot{\phi}^{(k)}(\tilde{\mathbf{U}}_{i,k}) - \dot{\phi}^{(k)}(\mathbf{U}_i)\} = o_P(1)$ (using the same arguments as in the Proof of Proposition A.5 in Lopez (2019)). ■

**Lemma 6.6** *Let $\phi$ denote a function such that*

$$E\left[\frac{\phi(\mathbf{U}_i, T_i, \mathbf{X}_i)^2}{S_C(T_i)}\right] < \infty,$$

*and*

$$E\left[|\phi(T)|\mathfrak{K}^{-1/2}(T)\right] < \infty.$$

*Then,*

$$\begin{aligned}\sum_{i=1}^{n} W_{i,n}\phi(\mathbf{U}_i, Y_i, \mathbf{X}_i) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(\mathbf{U}_i, Y_i, \mathbf{X}_i)}{S_C(Y_i)} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1-\delta_i)\bar{\phi}(Y_i)}{S_Y(Y_i)}\right. \\ &\quad \left. - \int_{-\infty}^{Y_i} E\left[\frac{\mathbf{1}_{C\leq t}\bar{\phi}(C)}{S_C(C)S_Y(C)}\right]\right\} dt + o_P(n^{-1/2}),\end{aligned}$$

*with $S_Y(y) = \mathbb{P}(Y \geq y)$ and*

$$\bar{\phi}(y) = E\left[\phi(\mathbf{U}, T, \mathbf{X})\mathbf{1}_{T\geq y}\right].$$

**Proof.** This result is simply the asymptotic representation of Kaplan-Meier type integrals obtained by Stute (1996), noticing that $W_{i,n}$ is the jump of the Kaplan-Meier estimator at observation $i$ (see also Satten and Datta (2001)). ■

**Lemma 6.7** *Under Assumption 4,*

$$\sum_{i=1}^{n} W_{i,n}^{*}[\hat{U}_i^{(k)} - U_i^{(k)}]\psi(\mathbf{U}_i) = \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}^{(k)}(\mathbf{L}_i, \mathbf{Z}_i, \delta_i)$$

*with*

$$\tilde{\psi}^{(k)}(\mathbf{l}, \mathbf{z}, \mathfrak{d}) = E\left[\eta^{(k)}_{\mathbf{L}_1,\mathbf{Z}_1}(\mathbf{L}_2, \mathbf{Z}_2, \delta_2)\psi(\mathbf{U}_2)|\mathbf{L}_2 = \mathbf{l}, \mathbf{Z}_2 = \mathbf{z}, \delta_2 = \mathfrak{d}\right].$$

**Proof.** See the proof of Proposition A.5 in Lopez (2019). ■

## 6.4 Proof of Theorem 3.1

To prove Theorem 3.1, we mainly develop the representation (3.6), since it follows the same path as the proof for (3.5) but with the more delicate point of dealing with kernel estimators. At the end of this section, we go back to (3.5) to explain how the asymptotic representation of the pseudo-observations (Assumption 4) is involved in the result.

Let us define

$$\begin{aligned}
\hat{M}_n(\theta, \mathbf{z}) &= \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_\theta(\hat{\mathbf{U}}_i), \\
M_n^*(\theta, \mathbf{z}) &= \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_\theta(\hat{\mathbf{U}}_i), \\
M(\theta, \mathbf{z}) &= E\left[\log \mathfrak{c}_\theta(\mathbf{U}) | \mathbf{Z} = \mathbf{z}\right] f_{\mathbf{Z}}(\mathbf{z}).
\end{aligned}$$

First observe that $\hat{\theta}(\mathbf{z}) - \theta(\mathbf{z}) = o_P(1)$, from Van der Vaart (2000), since

$$\sup_{\theta \in \Theta} |\hat{M}_n(\theta, \mathbf{z}) - M(\theta, \mathbf{z})| = o_P(1). \tag{6.14}$$

To obtain (6.14), we apply Proposition 6.1, and use the fact that,

$$\frac{1}{h^{p+1}} E\left[K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \log \mathfrak{c}_\theta(\mathbf{U}_i)\right] = M(\theta, \mathbf{z}) + h^2 \int\int \log \mathfrak{c}_\theta(\mathbf{u}) \mathbf{v}' \nabla_{\mathbf{z}}^2 f_{\mathbf{U},\mathbf{Z}}(\mathbf{u}, \tilde{\mathbf{z}}(\mathbf{v})) \mathbf{v} K(\mathbf{v}) d\mathbf{v} d\mathbf{u},$$

from Assumption 5, where $\tilde{z}(\mathbf{v})$ is such that each of its components is between the corresponding components of $\mathbf{z}$ and $\mathbf{z} + h\mathbf{v}$. From Assumption 1 and **??**,

$$\sup_{\theta} \left| \int\int \log \mathfrak{c}_\theta(\mathbf{u}) \mathbf{v}' \nabla_{\mathbf{z}}^2 f_{\mathbf{U},\mathbf{Z}}(\mathbf{u}, \tilde{\mathbf{z}}(\mathbf{v})) \mathbf{v} K(\mathbf{v}) d\mathbf{v} d\mathbf{u} \right| \leq \left| \int\int \log \mathfrak{C}(\mathbf{u}) \mathbf{v}' \mathfrak{s}_2(\mathbf{u}) \mathbf{v} K(\mathbf{v}) d\mathbf{v} d\mathbf{u} \right| = O(1).$$

Next, by definition, $\nabla_\theta \hat{M}_n(\hat{\theta}(\mathbf{z})) = 0$. Moreover, from a Taylor expansion,

$$\nabla_\theta \hat{M}_n(\hat{\theta}(\mathbf{z}), \mathbf{z}) = \nabla_\theta \hat{M}_n(\theta(\mathbf{z}), \mathbf{z}) + \nabla_\theta^2 \hat{M}_n(\tilde{\theta}(\mathbf{z}))(\hat{\theta}(\mathbf{z}) - \theta(\mathbf{z})),$$

where each component of $\tilde{\theta}(\mathbf{z})$ is between $\hat{\theta}(\mathbf{z})$ and $\theta(\mathbf{z})$. Note that $\tilde{\theta}(\mathbf{z}) - \theta(\mathbf{z}) = o_P(1)$ from the consistency of $\hat{\theta}(\mathbf{z})$.

$\nabla_\theta^2 \hat{M}_n(\tilde{\theta}(\mathbf{z}))$ is a matrix whose coefficient $(j,k)$ is

$$\sigma_n^{(j,k)} = \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n} K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \Phi_{\tilde{\theta}(\mathbf{z})}^{(j,k)}(\hat{\mathbf{U}}_i).$$

Write

$$\begin{aligned}
\nabla_\theta \hat{M}_n(\theta(\mathbf{z}), \mathbf{z}) &= \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \Phi(\mathbf{U}_i) + \Delta^1 I_n(\Phi, \mathbf{z}) + R_n(\Phi, \mathbf{z}), \\
\sigma_n^{(j,k)} &= \frac{1}{h^{p+1}} \sum_{i=1}^{n} W_{i,n}^* K\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h}\right) \Phi_{\tilde{\theta}(\mathbf{z})}^{(j,k)}(\mathbf{U}_i) + \Delta^1 I_n(\Phi_{\tilde{\theta}(\mathbf{z})}^{(j,k)}, \mathbf{z}) + R_n(\Phi_{\tilde{\theta}(\mathbf{z})}^{(j,k)}, \mathbf{z}).
\end{aligned}$$

From Proposition 6.2 and Assumptions 2 and 3, we get

$$|\Delta^1 I_n(\Phi, \mathbf{z})| + |\Delta^1 I_n(\Phi_{\tilde{\theta}(\mathbf{z})}^{(j,k)}, \mathbf{z})| = o_P(n^{-1/2} h^{-(p+1)/2}).$$

Next $R_n(\Phi, \mathbf{z}) = O_P(n^{-1/2})$ from (6.9) in Proposition 6.3, while $R_n(\Phi^{(j,k)}_{\tilde{\theta}(\mathbf{z})}, \mathbf{z}) = o_P(1)$ from (6.10) in Proposition 6.3. This yields

$$\nabla^2_\theta \hat{M}_n(\tilde{\theta}(\mathbf{z})) = \Sigma_h(\mathbf{z}) + o_P(1),$$

with

$$\Sigma_h(\mathbf{z}) = \Sigma(\mathbf{z}) + o(1).$$

This leads to (3.6).

Similarly, to obtain (3.5),

$$\nabla^2_\theta \hat{M}_n(\tilde{\theta})(\hat{\theta} - \theta_0) = -\nabla_\theta \hat{M}_n(\theta_0),$$

with $\nabla^2_\theta M_n(\tilde{\theta}) = \Sigma + o_P(1)$. From Proposition 6.5, we get that

$$\nabla_\theta \hat{M}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \nu_i(\Phi) + \tilde{\Phi}^{(k)}(\mathbf{L}_i, \mathbf{Z}_i, \delta_i) + o_P(n^{-1/2}),$$

and this shows (3.5).

## 6.5 Real data model choices and diagnostics

In this Section, we motivate the model choices assumed in Section 4.2, regarding firstly the margins in Section 6.5.1 and secondly the bandwith choice of the semi parametric copula estimator in Section 6.5.2.

### 6.5.1 Margins

Beginning with the duration $T$ studied on a log-scale, we consider Gaussian and Weibull distributions, in the Generalized Linear Model (GLM) framework. For the Weibull family, defined by the density $f_{(\lambda,k)}(x) = kx^{k-1}\lambda^{-k}e^{-(x/\lambda)^k}$, we fix the shape parameter $k = \alpha_k$, to let the family belong to exponential distributions. The corresponding GLM model is expressed as follows :

$$\mathbb{E}[\log(T_i)] = \lambda_i \Gamma(1 + 1/\alpha_k) = \exp(\alpha_0 + \alpha \mathbf{X}_i).$$

Then, for the losses $L^{(1)}$ and $L^{(2)}$, we consider Gamma and Weibull GLM family, also considering a fixed shape parameter for the Weibull as before. As for the Gamma family, defined by the density $f_{(\mu,\nu)}(x) = \Gamma(\nu)^{-1}x^{\nu-1}\nu^\nu\mu^{-\nu}\exp(-x\nu/\mu)$, the GLM for the loss $L^{(j)}$ consists in the following setting :

$$\mathbb{E}[\log(L_i^{(j),b})] = \mu_i^{(j),b} = \exp(\mu_0^{(j)} + \mu^{(j)}\left(\mathbf{X}_i^b, T_i^b\right)),$$

together with the estimation of parameter $\nu^{(j)}$.

For each margin $V$, we compare the obtained models $\hat{F}_{V|\mathbf{X}}$ relying on goodness of fit metrics, empirical log-likelihood and graphic diagnostics. In details, under the cumulative distribution function $F_{V|\mathbf{X}}$ that we assume to be continuous, $\left(F_{V|\mathbf{X}}(V_i|\mathbf{X}_i)\right)_{1\leq i\leq n}$ should be distributed as i.i.d. realizations of a uniform random variable on $[0;1]$. We therefore study the deviation of the estimated model $\hat{F}_{V|\mathbf{X}}$ from $F_{V|\mathbf{X}}$ by studying the deviation of $\left(\hat{F}_{V|\mathbf{X}}(V_i|\mathbf{X}_i)\right)_{1\leq i\leq n}$ from uniformity. We especially compare its IPCW weighted empirical cumulative distribution function $\hat{F}_U$, defined in 6.15, to the Identity, by computing Kolmogorov Smirnov 6.16 and Cramer Von Mises 6.17 statistics. Results are gathered in Table 12, and the graphical proximity of $\hat{F}_U$ from the Identity is highlighted in Figure 3 for $T$, in Figure 4 for $L^{(1)}$ and in Figure 5 for $L^{(2)}$.

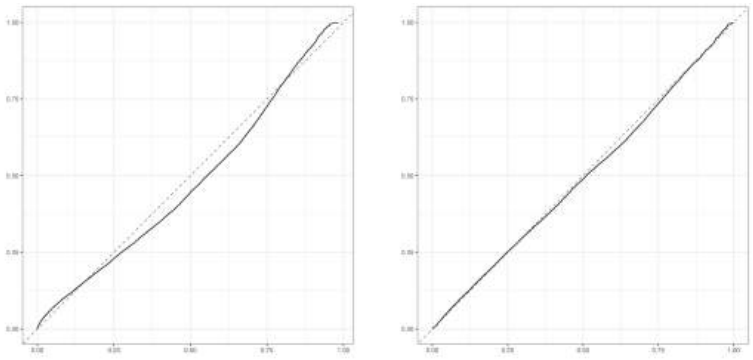
$$\hat{F}_U(x) = \frac{1}{\sum_{i=1}^n W_i} \sum_{i=1}^n W_i \mathbb{1}_{\hat{F}_{V|\mathbf{X}}(V_i|\mathbf{X}_i)\leq x} \tag{6.15}$$

$$d_{\text{KS}} = ||\hat{F}_U - \text{Id}||_\infty \tag{6.16}$$

$$d_{\text{CvM}} = \int_0^1 \left(\hat{F}_U(x) - x\right)^2 dx \tag{6.17}$$

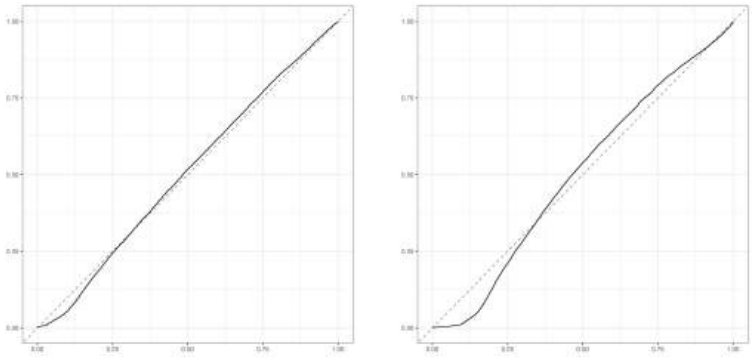
| Variable | Model | $d_{\text{KS}}$ | $d_{\text{CvM}}$ | $\mathcal{LL}$ |
|---|---|---|---|---|
| T | Log Gaussian GLM | 5.78 $10^{-2}$ | 1.33 $10^{-3}$ | -1.16 |
| | **Log Weibull GLM** | **2.65 $10^{-2}$** | **1.33 $10^{-4}$** | **$-1.09$** |
| $L^{(1)}$ | **Log Gamma GLM** | **4.55 $10^{-2}$** | **3.58 $10^{-4}$** | **$-1.72$** |
| | Log Weibull GLM | 1.00 $10^{-1}$ | 1.96 $10^{-3}$ | -1.79 |
| $L^{(2)}$ | Log Gamma GLM | 1.12 $10^{-1}$ | 4.32 $10^{-3}$ | -2.39 |
| | **Log Weibull GLM** | **1.21 $10^{-1}$** | **4.34 $10^{-3}$** | **$-2.24$** |

Table 12: Comparison of models for the margins $T$, $L^{(1)}$ and $L^{(2)}$. Chosen models and their statistics are in bold.



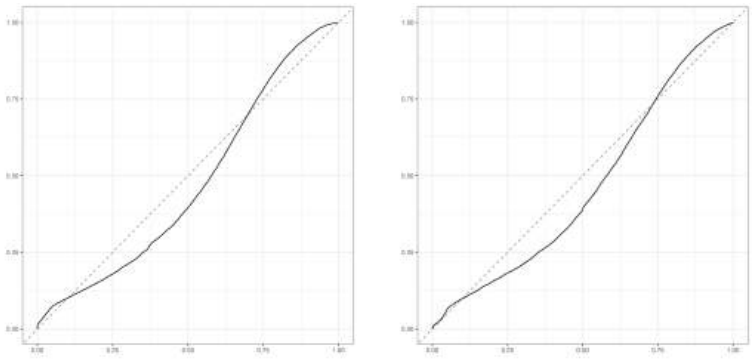
(a) Log normal GLM (b) Log Weibull GLM

Figure 3: Diagnostics of models for the variable $T$.



(a) Log Gamma GLM (b) Log Weibull GLM

Figure 4: Diagnostics of models for the variable $L^{(1)}$.



(a) Log Gamma GLM (b) Log Weibull GLM

Figure 5: Diagnostics of models for the variable $L^{(2)}$.

### 6.5.2 Copula

Regarding the semiparametric estimation of copulas on pseudo observations, the variance $h$ of the multivariate Gaussian density that plays the role of the kernel $K$ is a parameter that may influence the estimation performances. The optimal parameter $h^*$ is a trade of between overfitting, obtained with a low $h$, and oversmoothing, obtained with a high $h$. Indeed overfitting will prevent the estimated model to be relevant on new data while oversmoothing will lead to a quasi parametric model, both consequences being undesirable. To approach $h^*$, we rely on $k-$folds cross validation techniques. Consider a set of variances $(h_i)_{1\leq i\leq p}$ to be tested, for each one $h_i$, the cross validation approach consists in splitting the data in $k-$folds, before training for each fold $j$ a $h_i-$model based on the other folds, focusing on its performances on fold $j$. Suming the performances over the $k-$folds, we obtain a global assessment of the $h_i-$model performances. In our application, using $3-$folds cross validation for each copula family, we obtain in Table 13 the performences, expressed as log-likelihood. We finally select the $h$ associated with the higher performances for each copula family.

| | h=0.05 | h=0.1 | h=0.15 | h=0.2 |
|---|---|---|---|---|
| Clayton | 0.245 | **0.275** | 0.143 | 0.0866 |
| Frank | 0.0506 | **0.130** | 0.118 | 0.103 |
| Gumbel | 0.0281 | **0.123** | 0.110 | 0.0957 |

Table 13: $3-$folds cross validation performances, in terms of log-likelihood, depending on the parameter $h$ of the Kernel $K$. For each copula family, the highest performance is in bold, and lead to parameters used in Section 4.2 for semi-parametric copula estimation.

# References


Abegaz, F., Gijbels, I., and Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, 110:43 – 73. Special Issue on Copula Modeling and Dependence.

Antonio, K., Godecharle, E., and Van Oirbeek, R. (2016). A multi-state approach and flexible payment distributions for micro-level reserving in general insurance. *http://dx.doi.org/10.2139/ssrn.2777467*.

Antonio, K. and Plat, R. (2010). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al. (2011). A review of survival trees. *Statistics Surveys*, 5:44–71.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000). Copulas for finance-a reading guide and some applications. *Available at SSRN 1032533*.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Derumigny, A. and Fermanian, J.-D. (2017). About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, 5(1):154–197.

Einmahl, U., Mason, D. M., et al. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403.

Fermanian, J.-D. and Wegkamp, M. (2004). Time dependent copulas. *Preprint INSEE, Paris, France*.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Gerber, G., Le Faou, Y., Lopez, O., and Trupin, M. (2020). The impact of churn on client value in health insurance, evaluation using a random forest under various censoring mechanisms. *Journal of the American Statistical Association*, (just-accepted):1–22.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *The annals of statistics*, 11(1):49–58.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

Jaworski, P., Durante, F., Hardle, W. K., and Rychlik, T. (2010). *Copula theory and its applications*, volume 198. Springer.

Jin, X. and Frees, E. W. J. (2013). Comparing micro and macro level loss reserving models. *Preprint*.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Lopez, O. (2009). Single-index regression models with right-censored responses. *Journal of Statistical Planning and Inference*, 139(3):1082–1097.

Lopez, O. (2019). A censored copula model for micro-level claim reserving. *Insurance: Mathematics and Economics*, 87:1 – 14.

Lopez, O., Milhaud, X., and Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electron. J. Statist.*, 10(2):2685–2716.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *Astin bulletin*, 23(2):213–225.

Merz, M., Wüthrich, M. V., and Hashorva, E. (2013). Dependence modelling in multivariate claims run-off triangles. *Annals of Actuarial Science*, 7(1):3–25.

Nelder, J. A. and Baker, R. J. (1972). *Generalized linear models*. Wiley Online Library.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance 1. *ASTIN Bulletin: The Journal of the IAA*, 23(1):95–115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model variations and extensions. *ASTIN Bulletin*, 29(1):5–25.

Pigeon, M., Antonio, K., and Denuit, M. (2014). Individual loss reserving using paid–incurred data. *Insurance: Mathematics and Economics*, 58:121 – 131.

Portier, F. and Segers, J. (2018). On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166:160 – 181.

Sabban, I. C., LoPEz, O., and Mercuzot, Y. (2020). Automatic analysis of insurance reports through deep neural networks to identify severe claims.

Saluz, A., Bühlmann, H., Gisler, A., and Moriconi, F. (2014). Bornhuetter-Ferguson reserving method with repricing. *https://ssrn.com/abstract=2697167*.

Satten, G. A. and Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.

Sklar, M. (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian journal of statistics*, pages 461–471.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, pages 1089–1102.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *The Annals of statistics*, pages 1591–1607.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375.

Van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Veraverbeke, N., Omelka, M., and Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, 38(4):766–780.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879.

Wüthrich, M. V. (2016). Machine learning in individual claims reserving. *http://dx.doi.org/10.2139/ssrn.2867897*.

Wüthrich, M. V. (2017). Neural networks applied to chain-ladder reserving. *http://dx.doi.org/10.2139/ssrn.2966126*.

Zhao, X. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290 – 299.

# Article D

# Ransomware tweets analyzed by Hawkes process with application on insurance claims accumulation

# Ransomware tweets analyzed by Hawkes process with application on insurance claims accumulation

Sébastien Farkas[1], Caroline Hillairet[2], Olivier Lopez[1]

**Abstract**

In this paper we model insurance claims frequency by point processes. We especially focus on cyber risk and tackle its accumulation characteristic by estimating Hawkes processes. We develop a toy framework for focusing on assistance needs stochastic process in order to contribute to the sizing of assistance capacity. Indeed we provide theoretical bounds for the supremum of assistance needs stochastic process. We apply this methodology to ransomware tweets timelines analyzed as claims timelines. Our work highlights the strength of the Wannacry tweets propagation and specifies the level of assistance capacity than can mitigate cluster of ransomware attacks.

**Key words:** Insurance ; Cyber risk ; Ransomware ; Accumulation ; Hawkes processes.

**Short title:** Ransomware tweets analyzed by Hawkes process

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France.

E-mails: sebastien.farkas@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr

[2] Ensae Paris, Centre de Recherche en Economie et Statistique, CREST, 5 avenue Henry Le Chatelier, F-91120 Palaiseau, France.

E-mail: caroline.hillairet@ensae.fr

# 1 Introduction

Cyber risk is of a growing concern for our societies and their highly inter-connected people and economies. There is a need for organizations to protect themselves against malicious attack. In the same time, the cyber insurance can help to recover from unexpected financial impacts despite cautiauness. With an actuarial point of view, we propose in this paper to tackle the sustainability of cyber insurance regarding ransomwares, which are a type of malicious attack with an extorsion strategy. Indeed, cyber risk can question the risk pooling principle of insurance. For instance, information systems rely on few standardized kernels and a lot of economic actors may need assistance and financial compensation in the same time. Therefore, insurers are looking to a tradeoff between the insurance coverage growth and risk mitigation in order to contribute to cyber risk management at the society level and with rationalized reactions while remaining sustainable.

Regarding ransomwares, several approaches are pursued by hacker organizations : from massive spread attacks with automatic propagation, to big game hunting that focus on specific targets, by way of the sale of ransomware as a service. The rise in power of ransomwares shakes informatics gullibility by repositioning information systems as a critical asset for our economy, rather than an unlimited trusting tool. Because of the global dependence on cyber, its risk management needs a global approach, from the susceptible economic actors, with prevention and resilience, to the human root of ransomware, with legislation and investigation. We especially focus on the capacity of an insurer to absorb the assistance services needs in the case of a ransomware spreading.

The WannaCry ransomware that spread in May 2017 and paralyzed numerous information systems, is an illustration of cyber accumulation risk, that encourages in a global risk reduction, necessary to mitigate accumulation. Yet the ransomare phenomenon is globally rising, it seems very hard to track in the short term, from mild to storm, without physical explanation. Since it is a human made risk, its hazard part is likely the result of a game theory equilibrium between hackers and economic actors, but the statistical analysis of the resulting consequences remains key. In this paper, we investigate accumulation of assistance needs resulting by spreads of ransomwares. We rely on tweets data in order to calibrate those spreads. A broad range of cyber issues have been tackled by tweets data, from tweets behaviour after a cyber attack in Vogler Daniel (2020) to the tweets dynamics regarding 0-day vulnerabilities in Le et al. (2019). From an economic perspective, a cyber risk daily index has also been proposed in Stéphane Lhuissier (2021).

In the present paper, we propose an approach tailored for analyzing events occurrence data, with instability and clusters. We rely on Hawkes processes that allow us to capture self excitation phenomenon, that may result either from ransomware auto propagation, or by the evolving opportunities for hackers to spread their solutions across economic actors.

The rest of the paper is organized as follows. In Section 2, we introduce a framework to measure cyber assistance needs. Then, in Section 3, we introduce a bound for assistance needs with an underling Hawkes auto-excited process. Finally, in Section 4, we adjust Hawkes processes on ransomware tweets timelines. and apply the main result to built some scenarii of ransomware assistance needs.

# 2 Framework

This section introduces a framework for monitoring claims occurrence and assistance needs throught the portfolio of an insurer. In Section 2.1, we introduce the concept of point processes for encoding claims occurrence dynamics. Then, in Section 2.2 we introduce Hawkes processes that allow us to capture dependence in claims occurrence and especially to model auto-excited phenomenon. Finally, we introduce in Section 2.3 the stochastic process corresponding to assistance needs and the relative assumptions.

## 2.1 Modeling portfolio claims occurrence with point processes

The claims frequency of insurance portfolios is usually monitored by aggregated count data. For instance, one can study the monthly number of claims $(N_j)_{j\geq 1}$ thanks to a parametric and discrete distribution. Usually, those numbers are considered in front of an exposure metric, for instance the number of contracts or the number of coverages. This basically allows one to compute frequency metrics. In a straightforward application, that is assuming the $(N_j)_{j\geq 1}$ are i.i.d, this model does not allow one to capture neither the saisonality neither the dependence between claims occurrence.

Alternatively, the occurrence of claims can be modeled by a point process that is a stochastic process jumping by a unit each time a claim occured. Let us note this process $\mathfrak{I}(t)$ as we focus on ransomware infection related claims. The observation of a point process during an interval $[0;T]$ is therefore equivalent to the information of the relative claims date $\tau_i$ and the relative end of observation date, that is $T$. Indeed we have the

following point process definition,

$$\forall t \leq T, \mathfrak{I}(t) = \sum_{i \geq 1} \mathbb{1}_{\tau_i \leq t}.$$

This framework relies on more granular information and especially allows one to retrieve the more classical aggregated count data used in frequency studies. Indeed, let us note $t_j$ the relative end date of each month associated with the $N_j$ claims. Then, letting $t_0$ equal to zero, we have the following relation :

$$\forall j \geq 1, N_j = \mathfrak{I}(t_j) - \mathfrak{I}(t_{j-1}).$$

A point process can be characterized by its intensity, that informally accounts for the infinitesimal probability of realization of a claim at each time conditionnally on the past information, that is more formally a filtration $\mathcal{F}$ on which the process is predictible. Three families of point processes are of a particular interest :

- the homogeneous Poisson processes are characterized by a constant intensity $\lambda > 0$. More precisely $\mathfrak{I}$ is an homogeneous Poisson process if for all $t > s > 0$, we have that $\mathfrak{I}(t) - \mathfrak{I}(s)$ is a Poisson random variable with parameter $\lambda(t-s)$ independent of $\mathcal{F}_s$.

- the inhomogeneous Poisson processes are characterized by a deterministic function intensity $\lambda$. More precisely $\mathfrak{I}$ is an inhomogeneous Poisson process if for all $t > s > 0$, we have that $\mathfrak{I}(t) - \mathfrak{I}(s)$ is a Poisson random variable with parameter $\int_s^t \lambda(u)du$ independent of $\mathcal{F}_s$.

- the Cox processes are characterized by a stochastic intensity $\lambda$ predictible according to $\mathcal{F}$. More precisely $\mathfrak{I}$ is a Cox process if for all $t > s > 0$, we have that $\mathfrak{I}(t) - \mathfrak{I}(s)$ is a Poisson random variable with parameter $\int_s^t \lambda(u)du$. This random variable can be dependant according to $\mathcal{F}_s$.

Therefore, when the empirical number of claims $(N_j)_{j \geq 1}$ is modeled by a Poisson law, we can say that there is an underlying assumption that the claims ocurence follow an homogeneous Poisson process. Inhomogeneous Poisson process basically allows one to model claims saisonality and Cox processes to model claims dependence. In this paper, we consider Hawkes processes that are tailored for auto-excited phenomena.

## 2.2 Hawkes processess for ransomware propagation

Hawkes processes are tailored for analysis phenomenon with self-exciting succession of events. For instance it allows one to explain clusters by dependency between events occurrence, that is encoded as the kernel of the process, a function $h$ from $\mathbb{R}^+$ to $\mathbb{R}^+$. We focus on a subclass of Hawkes processes, called self exciting linear Hawkes processes. Let us consider a Hawkes process $\mathfrak{I}$ with an intensity $\lambda$ expressed as follows,

$$\lambda_{N^{(\mu,\theta)}}(t) = \mu + \int_0^t h(t-s)dN^{(\mu,h)}, \tag{2.1}$$

where $\mu > 0$ can be interpreted as an homogeneous Poisson intensity baseline. Simulating a Hawkes process can be done by simulating the baseline Poisson process and then by simulating iteratively for each event the new events caused by auto-exitation. To make the intensity $\lambda_{N^{(\mu,\theta)}}$ predictable, we can impose $h(0) = 0$ as pointed out in Example 14.3(c) of D. J. Daley (2008).

Hawkes analysis based on self-exciting observations arised in numerous fields, for instance in finance Hawkes (2018), in risk quantification Chavez-Demoulin and McGill (2012), and in neurology Bonnet et al. (2021). When self-inhibition is also observed, Hawkes processes remain an appropriate modeling tool to study the whole self-dependence structure. In this paper, we propose to apply this framework to the modelisation of ransomeware related claims dynamics.

## 2.3 Metrics for portfolio assistance needs

In this paper, we aim to quantify the assistance needs in an insurance portolio in order to contribute to the sizing of assistance capacities. We assume the claims occurrence follows a Hawkes process $\mathfrak{I}(t)$ with an intensity defined in equation (2.1). We further assume that each victim needs assistance during a constant time $\delta$ and that the crisis is mitigated after this time. Therefore, we introduce two stochatic process : $\mathfrak{R}(t)$ corresponding to the number of recovered victims and $\mathfrak{A}(t)$ corresponding to the number of victims in need of assistance.

$$\begin{cases} \mathfrak{I}(t) &= \sum_{i\geq 1} \mathbb{1}_{\{\tau_i \leq t\}} \\ \mathfrak{R}(t) &= \sum_{i\geq 1} \mathbb{1}_{\{\tau_i \leq t-\delta\}} \\ \mathfrak{A}(t) &= \sum_{i\geq 1} \mathbb{1}_{\{t-\delta < \tau_i \leq t\}} \end{cases} \tag{2.2}$$

We therefore focus in assessing the behaviour of assistance needs $\mathfrak{A}(t)$, and more specifically on its tail behaviour in order to quantify the probability of assistance capacity saturation.

**Remark 2.1.** *The metrics we propose do not take into account any portfolio exposure metric. Therefore, it must be extended to be applied to a portfolio with significant exposure changes. Moreover, we assume a deterministic and unique type of assistance need whereas it may depends on the victim and the attack specificities : for instance the size of the company, its sector, the type of ransomware and the claims scale. Our study may therefore be extended by cyber assistance dynamics.*

# 3 Scaling assistance capacity for accumulation scenarii

## 3.1 Analytical expression of the deviation

**Assumption 1.** *The $L^1$-norm of the kernel $h$ of the Hawkes process of the claims $\mathfrak{I}(t)$ is strictly lower than 1, that is $||h||_1 = \int_{\mathbb{R}^+} h(t)dt < 1$.*

Under this assumption, we can consider $H$ as follows,

$$H(t) = \sum_{k\geq 1} h^{*k}(t),$$

with $(h^{*k})_{k\geq 1}$ recursively defined by convolution products,

$$h^{*1}(t) = h(t), \text{ and for all } k \geq 2, h^{*k}(t) = \int_0^t h(t-s)h^{*(k-1)}(s)ds.$$

We especially have $||H||_1 = \int_{\mathbb{R}^+} H(t)dt = \frac{||h||_1}{1-||h||_1} < 1$.

**Proposition 3.1.** *Consider the stochastic process of assistance needs $\mathfrak{A}(t)$ defined in equation (2.2). Let $M_{\mathfrak{I}}$ be the compensated martingale of $\mathfrak{I}(t)$, that is $M_{\mathfrak{I}}(t) = \mathfrak{I}(t) - \Lambda_{\mathfrak{I}}(t)$, where $\Lambda_{\mathfrak{I}}(t) = \int_O^t \lambda_{\mathfrak{I}}(u)du$ is the compensator of $\mathfrak{I}(t)$. If Assumption 1 is verified, then $\mathfrak{A}(t)$ can be expressed as follows,*

$$\mathfrak{A}(t) = \mathbb{E}[\mathfrak{A}(t)] + \mathfrak{M}(t), \text{ with}$$

$$\mathbb{E}[\mathfrak{A}(t)] = \mu \int_0^t \mathbb{1}_{\{t-\delta<u\}} \left(1 + \int_0^u H(u-z)dz\right) du,$$

$$\mathfrak{M}(t) = \int_0^t \mathfrak{m}_t(z) dM_{\mathfrak{I}}(z),$$

$$\text{and } \mathfrak{m}_t(z) = \mathbb{1}_{\{t-\delta<z\}} + \int_z^t \mathbb{1}_{\{t-\delta<u\}} H(u-z)du.$$

This expression allows us to firstly focus on the bounding of the deviation $\mathfrak{M}(t)$ in Section 3.2 and then to obtain a bound for the assistance needs $\mathfrak{A}(t)$ in Section 3.3.

**Proof 3.2.** *The accumulation risk $\mathfrak{A}(t)$ can firstly be expressed as follows,*

$$\mathfrak{A}(t) = \sum_{i\geq 1} \mathbb{1}_{\{t-\delta<\tau_i\leq t\}} = \int_0^t \mathbb{1}_{\{t-\delta<u\}} d\mathfrak{I}(u) = \int_0^t \mathbb{1}_{\{t-\delta<u\}} dM_{\mathfrak{I}}(u) + \int_0^t \mathbb{1}_{\{t-\delta<u\}} \lambda(u)du.$$

*We then use an alternative expression of the intensity of a Hawkes process, developped in the proposition 2.1 of Jaisson and Rosenbaum (2015) and valid under Assumption 1,*

$$\lambda(s) = \mu\left(1 + \int_0^s H(s-u)du\right) + \int_0^s H(s-z)dM_{\mathfrak{I}}(z), \text{ therefore } \mathfrak{A}(t) \text{ is equal to}$$

$$\int_0^t \mathbb{1}_{\{t-\delta<u\}} dM_{\mathfrak{I}}(u) + \int_0^t \mathbb{1}_{\{t-\delta<u\}} \left[\mu\left(1 + \int_0^u H(u-z)dz\right) + \int_0^u H(u-z)dM_{\mathfrak{I}}(z)\right] du.$$

*Then, we focus on the term $\int_0^t \mathbb{1}_{\{t-\delta<u\}} \int_0^u H(u-z)dM_{\mathfrak{I}}(z)du$. We simplify it by applying the Fubini's theorem as follows :*

$$\int_{0\leq u\leq t} \int_{0\leq z\leq u} \mathbb{1}_{\{t-\delta<u\}} H(u-z)dM_{\mathfrak{I}}(z)du = \int_{0\leq z\leq t} \int_{z\leq u\leq t} \mathbb{1}_{\{t-\delta<u\}} H(u-z)dudM_{\mathfrak{I}}(z).$$

*Therefore, we obtain the following decomposition that leads to the result.*

$$\begin{aligned}
\mathfrak{A}(t) \quad = \quad & \mu \int_0^t \mathbb{1}_{\{t-\delta<u\}} \left(1 + \int_0^u H(u-z)dz\right) du \\
+ \quad & \int_0^t \mathbb{1}_{\{t-\delta<u\}} dM_{\mathfrak{I}}(u) \\
+ \quad & \int_0^t \int_z^t \mathbb{1}_{\{t-\delta<u\}} H(u-z)dudM_{\mathfrak{I}}(z).
\end{aligned}$$

□

## 3.2 Bounding the supremum of the deviation $\mathfrak{M}(t)$

Our aim is to bound the supremum over the interval $[0;T]$ of the deviation of $\mathfrak{A}(t)$ from its expectation, that is $\sup_{0\leq t\leq T}\mathfrak{M}(t)$. The result globally relies on the following decomposition,

$$\begin{array}{rcl}\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right) & = & \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\cap\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\\ & + & \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\cap\left\{\Lambda(T)\geq\mathfrak{c}_u^T T\right\}\right)\\ & \leq & \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\mid\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\mathbb{P}\left(\Lambda(T)<\mathfrak{c}_u^T T\right)\\ & + & \mathbb{P}\left(\Lambda(T)\geq\mathfrak{c}_u^T T\right)\\ & \leq & \mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\mid\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)\\ & + & \mathbb{P}\left(\Lambda(T)\geq\mathfrak{c}_u^T T\right)\end{array}$$

We begin to study in Section 3.2.1 the first term, applying a proposition of Guével (2021) that allows us to obtain a bound conditionnally on the event that the compensator on $T$ is bounded by $\mathfrak{c}_u^T T$. Then we bound the second term in Section 3.2.2 thanks to a proposition of Reynaud-Bouret and Roy (2007).

### 3.2.1 Bound of $\mathbb{P}\left(\left\{\sup_{0\leq t\leq T}\mathfrak{M}(t)>x\right\}\mid\left\{\Lambda(T)<\mathfrak{c}_u^T T\right\}\right)$

In this section, we check the necessary conditions of the Theorem 5 of Guével (2021) and we then apply the result with adapted notations. In detail, we check that the supremum of $\mathfrak{m}_t$ can be uniformly bounded.

**Proposition 3.3.** *If the Assumption 1 is verified we can uniformly bound the supremum of* $\mathfrak{m}_t$,

$$\sup_{t>0}\sup_{0\leq z\leq t}|\mathfrak{m}_t(z)|\leq\mathfrak{H}<+\infty,\ \textit{with}$$

- $\mathfrak{H}=1+\int_0^{+\infty}H(u)du$ *in full generality,*
- $\mathfrak{H}=1+\int_0^{\delta}H(u)du$ *in the case* $H$ *is decreasing.*

**Proof 3.4.** *For all* $T>0$, *for all* $t\in[0;T]$, *and for all* $z$ *between* $0$ *and* $t$, *we have :*

$$\begin{array}{rcl}\mathfrak{m}_t(z) & = & \mathbb{1}_{\{z+\delta>t\}}+\int_z^t\mathbb{1}_{\{u+\delta>t\}}H(u-z)du\\ & = & \mathbb{1}_{\{z+\delta>t\}}+\int_0^{t-z}\mathbb{1}_{\{v+z+\delta>t\}}H(v)dv\\ & = & \mathbb{1}_{\{t-z-\delta<0\}}+\int_0^{t-z}\mathbb{1}_{\{v>t-z-\delta\}}H(v)dv\\ & = & \mathbb{1}_{\{t-z-\delta<0\}}+\int_{\max(t-z-\delta,0)}^{t-z}H(v)dv,\end{array}$$

*We bound the first term by one and the second either by* $\int_0^{+\infty} H(u)du$ *in full generality or by* $\int_0^{\delta} H(u)du$ *in the case* $H$ *is decreasing, leading to the desired result.* □

**Corollary 3.5.** *If Assumption 1 is verified we can obtain the following bound,*

$$\int_0^T \sup_{0<t\leq T} \sup_{0\leq z\leq t} |\mathfrak{m}_t(u)| d\Lambda(u) \leq \mathfrak{c}_u^T T \mathfrak{H}^2 < +\infty.$$

The conditions are now gathered to apply the Theorem 5 of Guével (2021).

**Theorem 3.6** (Theorem 5 of Guével (2021))**.**
*If Assumption 1 is verified, then for all* $x \geq 0$,

$$\mathbb{P}\left(\left\{\sup_{0\leq t\leq T} \mathfrak{M}(t) > x\right\} \mid \left\{\Lambda(T) < \mathfrak{c}_u^T T\right\}\right) \leq \exp\left(-\frac{\mathfrak{c}_u^T T \mathfrak{H}^2}{\mathfrak{H}} I\left(\frac{\mathfrak{H}}{\mathfrak{c}_u^T T \mathfrak{H}^2} x\right)\right), \text{ with}$$

$$I : u \mapsto (1+u)\log(1+u) - u.$$

The bound can therefore be simplified by $\exp\left(-\mathfrak{c}_u^T T \mathfrak{H} I\left(\frac{x}{\mathfrak{c}_u^T T \mathfrak{H}}\right)\right)$.

### 3.2.2 Bound of $\mathbb{P}\left(\Lambda(T) \geq \mathfrak{c}_u^T T\right)$

In this section, we check the necessary conditions of the Theorem 3.4 of Reynaud-Bouret and Roy (2007) and we then apply the result with adapted notations. We especially make some assumptions regarding the kernel of the claims process $\mathfrak{I}(t)$.

**Assumption 2.** *The support of the kernel* $h$ *of the Hawkes process* $\mathfrak{I}$ *is assumed to be a compact :*

$$\exists \Delta < +\infty \text{ such that the support of } h \text{ is included in } [0;\Delta].$$

**Assumption 3.** *The kernel* $h$ *of the Hawkes process* $\mathfrak{I}$ *is assumed to be bounded :*

$$||h||_\infty := \sup_{s\in[0;\Delta]} h(s) < +\infty.$$

**Theorem 3.7** (Proposition 3.4 de Reynaud-Bouret and Roy (2007))**.**
*If Assumptions 1, 2 and 3 are verified, we have the following result regarding the compensator* $\Lambda(T)$.

$$\forall u > 0,\ \exists t_u > 0 \text{ such that for all } T \geq t_u,\ \mathbb{P}\left(\frac{\Lambda(T)}{T} \geq \mathfrak{c}_u^T\right) < \left(3 + \frac{\mu e}{u + \log(T)}\right) e^{-u},$$

$$\textit{with } \mathfrak{c}_u^T = \mu + ||H||_1 + a_T^{(\mu,h)}\sqrt{u} + b_T^{(\mu,h)}u + c_T^{(\mu,h)}u^2 + d_T^{(h)}u^3,$$

$$a_T^{(\mu,h)} = \sqrt{\frac{16\mu^2\Delta\log(T)}{T\mathfrak{h}}},\ b_T^{(\mu,h)} = \sqrt{\frac{16\mu^2\Delta}{T\mathfrak{h}}} + \frac{8\log(T)\left(\mu + 2||h||_\infty \mathfrak{d}_T^{(\mu,h)}\right)}{3T\mathfrak{h}},$$

$$c_T^{(\mu,h)} = \frac{8\left(\mu + 2||h||_\infty\left(\mathfrak{d}_T^{(\mu,h)} + \frac{\log(T)}{\mathfrak{h}}\right)\right)}{3T\mathfrak{h}},\ d_T^{(h)} = \frac{16||h||_\infty}{3T\mathfrak{h}^2},\ \mathfrak{h} = ||h||_1 - \log(||h||_1) - 1,$$

$$\mathfrak{d}_T^{(\mu,h)} = \frac{\mu\Delta\left(l_0(\mathfrak{h}) + l_1(\mathfrak{h})\right) + \log(T)}{\mathfrak{h}},\ l_0(v) = e^{g^{-1}(v)} - 1,\ l_1(v) = \frac{e^{g^{-1}(v)-v}}{1-e^{-v}} - 1,$$

$$g : z > 0 \mapsto z - ||h||_1(e^z - 1),\ \textit{that allow one to defined its inverse } g^{-1} \textit{ on } [0;\mathfrak{h}].$$

## 3.3 Bounding of assistance needs

In this section, we apply the results obtained in Sections 3.2.1 and 3.2.2 in order to bound the supremum of assistance needs $\mathfrak{A}(t)$. We first begin with a bound regarding the expectation of $\mathfrak{A}(t)$.

**Proposition 3.8.** *If Assumption 1 is verified, we can bound the expectation of* $\mathfrak{A}(t)$.

$$\mathbb{E}[\mathfrak{A}(t)] \leq \mu\delta\left(1 + ||H||_1\right).$$

**Proof 3.9.** *Let us recall the expression of the expectation of* $\mathfrak{A}(t)$.

$$\mathbb{E}[\mathfrak{A}(t)] = \mu\int_0^t \mathbb{1}_{\{u+\delta>t\}}\left(1 + \int_0^u H(u-z)dz\right)du,$$

*Therefore, for* $t \geq \delta$, *we have the following equality that is an upper bound for* $t < \delta$.

$$\mathbb{E}[\mathfrak{A}(t)] = \mu\int_{t-\delta}^t \left(1 + \int_0^u H(v)dv\right)du.$$

*Moreover since* $\int_0^u H(v)dv \leq ||H||_1$, *we obtain the desired bound.* □

**Theorem 3.10.** *If Assumptions 1, 2 and 3 are verified, we can bound in probability the supremum of* $\mathfrak{A}(t)$ *between* 0 *and* $T$ *with the following result with the notations previously introduced. In detail,* $\forall u > 0,\ \exists t_u > 0$ *such that for all* $T \geq t_u$*, we have,*

$$\mathbb{P}\left(\sup_{0\leq t\leq T} \mathfrak{A}(t) > x\right) \leq \exp\left(-\mathfrak{c}_u^T T\mathfrak{H} I\left(\frac{x-\mu\delta\left(1+||H||_1\right)}{\mathfrak{c}_u^T T\mathfrak{H}}\right)\right) + \left(3+\frac{\mu e}{u+\log(T)}\right)e^{-u}.$$

**Proof 3.11.**

$$\begin{array}{rcl}
\mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{A}(t) > x\right) & = & \mathbb{P}\left(\sup_{0\leq t\leq T}\left[\mathbb{E}[\mathfrak{A}(t)]+\mathfrak{M}(t)\right] > x\right) \\
& \leq & \mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{M}(t) > x - \sup_{0\leq t\leq T}\mathbb{E}[\mathfrak{A}(t)]\right) \\
& \leq & \mathbb{P}\left(\sup_{0\leq t\leq T}\mathfrak{M}(t) > x - \mu\delta\left(1+||H||_1\right)\right) \\
& \leq & \exp\left(-\mathfrak{c}_u^T T\mathfrak{H} I\left(\frac{x-\mu\delta(1+||H||_1)}{\mathfrak{c}_u^T T\mathfrak{H}}\right)\right) \\
& + & \left(3+\frac{\mu e}{u+\log(T)}\right)e^{-u}.
\end{array}$$

□

**Remark 3.12.** *In the obtained bound,* $u$ *is free in* $\mathbb{R}^{+*}$ *as soon as* $T > t_u$*. But* $t_u$ *itself is unknown and we do not investigate a way to approximate it. Assuming that* $T > t_u$ *for* $u \in [u_w; u_m]$*, the bound might be applied with an optimal* $u^*$ *that minimizes the bound on* $[u_w; u_m]$*. Unfortunately,* $u^*$ *can not be found analytically, but it can estimated by optimization algorithms in specified/numerical situations. Basically,* $u^*$ *will be the solution of a tradeoff between the left hand side of the bound, that increases to* 1 *for high* $u$*, and the right hand side of the bound, that increases to* 3 *for* $u$ *closed to 0. We however do not perform an analysis of this dependence on* $u$*.*

# 4 Application on ransomwares tweets

## 4.1 Analysis of tweets dynamics regarding some ransomwares

Twitter, recently renamed X, is a platform that allows people to publicly share and react on short messages. Under the Twitter Academic Research API, we can access to all tweets, beginning with the first tweet of March 2006. The extraction of tweets can be done by key words searches and timeline focus. We therefore identify and focus on 7 ransomwares mentionned in the report of the french governmental cyber security institution Agence nationale de la sécurité des systèmes d'information (ANSSI). The considered ransomwares cover the main types of ransomwares : self-propagation, ransomware as a service and big

game hunting, as detailed below. For each ransomware we perform an extraction in order to retrieve the related tweets propagation timeline. In details, we extract tweets containing both the ransomware name, for instance 'WannaCry', and the keyword 'ransom'. We exclude reactions information, basically retweets and replies. In addition we remove tweets simultaneoulty posted, firstly because under simple process framework, duplicates are almost surely impossible, and secondly because of the high likelihood that it comes from a robot, with similar tweets message. This is however not supposed to be a strain against messages posted by bots, which would require much more attention.

| Ransomware | Type | Apparition | Withdrawal |
|---|---|---|---|
| Wannacry | Self propagation | 2017-05-12 | - |
| Sodinokibi REvil | Ransomware as a service | 2019-04-17 | 2021-07-13 |
| Ryuk | Big Game Hunting | 2018-08 | - |
| SamSam | Big Game Hunting | 2016-01-06 | 2018-11 |
| GrandCrab | Ransomware as a service | 2018-01 | 2019-17-06 |
| Maze | Big Game Hunting | 2019-05-29 | 2020-12 |
| Dharma | Ransomware as a service | 2016-11-16 | - |

Table 1: Selected ransomwares

Unfortunately, tweets do not precisely match with infections, regarding both numbers and dynamics, because Twitter is first of all a massive communication tool. We however analyze tweets propagation timelines as a portfolio claims timelines. In the same time, we try to limit this biais by excluding clusters of tweets not directly linked to ransomware propagation and impacts. For instance, regarding WannaCry tweets, we identified 4 main clusters as highlighted in Figure 1. The first one corresponds with the first wave of Wannacry propagation. However, the second one is likely due to the FBI arrestation of Marcus Hutchins, the man that discovered a parade to WannaCry impacts, initially designed to prevent hackers from being themselves targetted. The third one because of its judgment and the last one because of the release of a TV show about Marcus Hutchins Story. Finally, for WannaCry, we focus on tweets posted from the 12 May 2017 to the 2 August 2017. For other ransomwares, less famous, we exclude tweets published outside the period of activity of ransomwares, based on the conclusion of experts.

We therefore obtain 7 subsets of tweets timelines, detailed in Table 2, that we further consider as ransomwares related claims spreading timelines. The dynamics of occurrence are illustrated in Figure 2. As expected, we clearly observe clusters in all the dynamics,

that support the choice of Hawkes process for modelling the occurrence of events. Also, the choice of a constant baseline intensity does not seem incoherent because no obvious trend nor seasonality is idendified.

## 4.2 Calibration of Hawkes processes kernels

In this section we calibrate for each ransomware timeline a Hawkes process and firstly introduce the basis of log-likelihood estimation of a point process. The log-likelihood of observing events at times $(\tau_i)_{1\leq i\leq \mathfrak{I}(T)}$ in $[0;T]$ from a point processs with a parametric intensity $\lambda$ indexed by $\theta$ and the resulting compensator $\Lambda_\theta$ can be expressed as follow.

$$\mathcal{L}_\theta\left(\mathfrak{I}(t),T\right)=\prod_{i=1}^{\mathfrak{I}(T)}\left(\lim_{t\overset{<}{\to}\tau_i}\lambda_\theta(t)\right)e^{-\Lambda_\theta(T)}.$$

A demonstration of this well known property is given in Section 6.1 for the completeness of the paper. Informally, the resulting estimate is a tradeoff between the maximization of the left limit of the intensity at observed events $(\tau_i)_{1\leq i\leq \mathfrak{I}(T)}$ , and its minimization during laps between events.

Regarding ransomware propagation, we expect clusters because of the underlying spreading process of such cyber attacks. Empirically, we observe such auto-excitation patterns for tweets timelines. For instance, for the Wannacry timeline, we analyze the delays of laps between events. We first gather all the laps in order to illustrate the global distribution of delays between events. Then, we focus on the laps that follow laps lower than a given threshold in order to analyze the delays between events conditionnally on the facts that the concentration of previous events is higher. Under the self excitation assumption, the concentration of the conditionnal distributions should increase with the concentration of previous events, that is when the threshold is getting closer to 0. Those trends are empirically observed for the Wannacry tweets timeline in Figure 3. This conforts the choice of using Hawkes processes.

We further assume the Hawkes kernel to be parametric and more specifically to be a truncated exponential kernel. This kernel is identifiable by a three dimensional parameter $\theta=(\alpha,\beta,\Delta)$ as follow.

$$h_\theta(0)=0,\ \forall\, t\in]0;\Delta],\ h_\theta(t)=\alpha e^{-\beta t}\text{ and }\forall\, t>\Delta,\ h_\theta(t)=0, \tag{4.1}$$

with $\alpha>0$, $\beta>0$, and $\frac{\alpha}{\beta}<1$. Therefore the auto-excitation phenomenon is bounded in time by $\Delta$. For the truncated exponential kernel, we can obtain a closed form for the

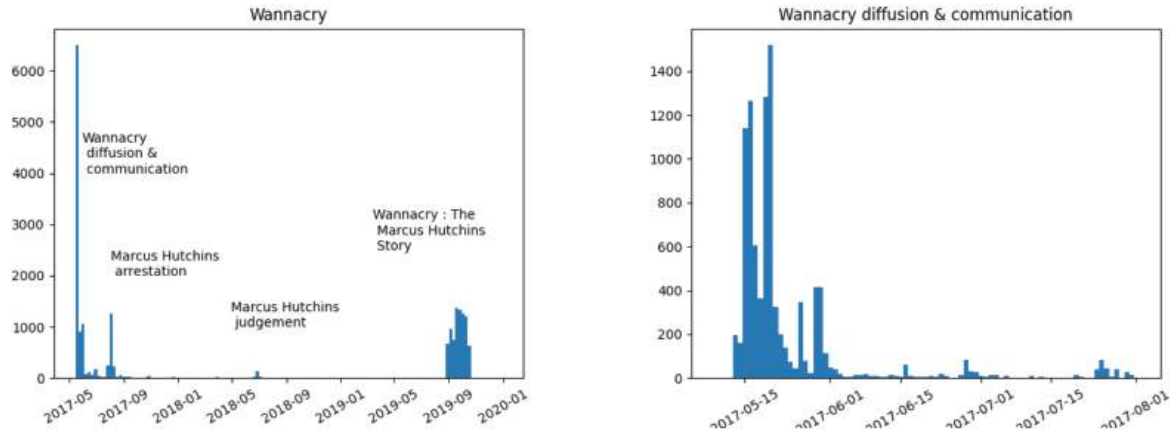
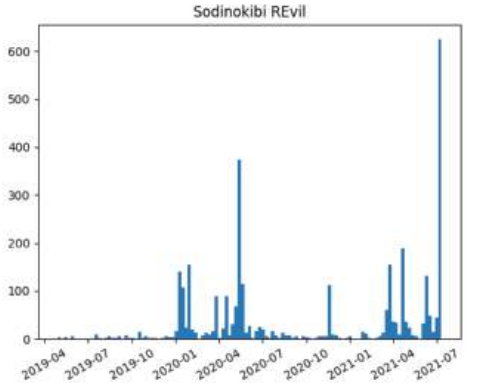


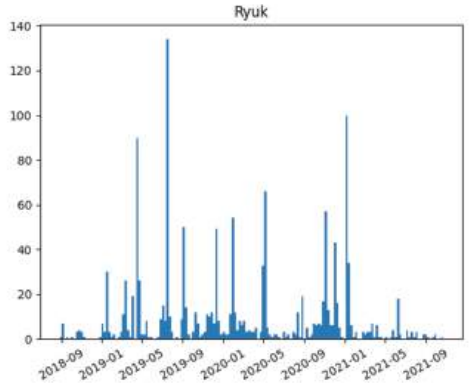

Figure 1: WannaCry tweets timeline. At the left hand side, the weekly number of tweets is represented from April 2017 to 2020. At the right hand side, the dayly number of tweets is represented, from April 2017 to August 2017.

Table 2: Selected ransomwares with descriptive statistics about their observed activity period by experts, the period with tweets and the number of tweets.

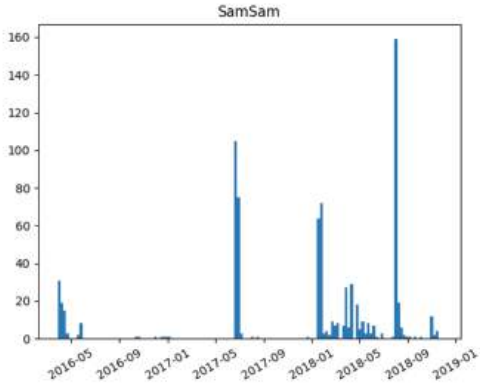
| Ransomware | Delay | Observation stop | Number |
|---|---|---|---|
| Wannacry | 0 | 2017-08-02 | 8964 |
| Sodinokibi REvil | 10 | 2021-07-12 | 3293 |
| Ryuk | $< 31$ | 2021-10-31 | 1371 |
| SamSam | 79 | 2018-11-30 | 853 |
| GrandCrab | $< 31$ | 2019-12-31 | 828 |
| Maze | 2 | 2020-12-31 | 810 |
| Dharma | 0 | 2021-10-31 | 285 |



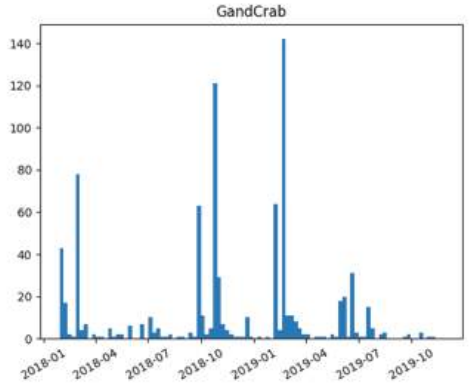

a) Sodinokibi REvil



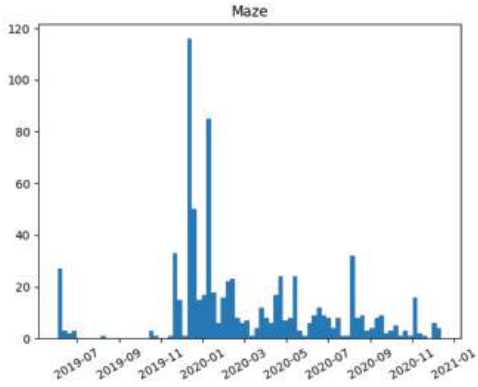

b) Ryuk



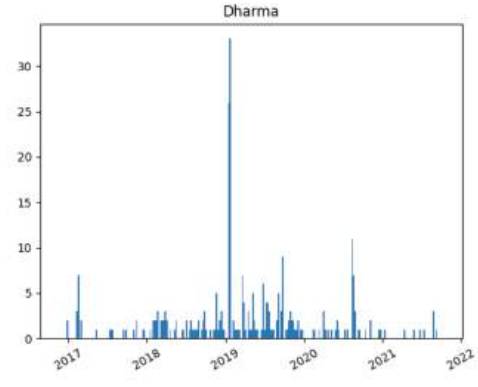

c) SamSam




d) GandCrab

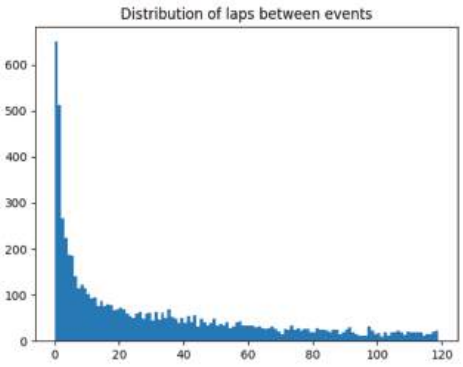



c) Maze

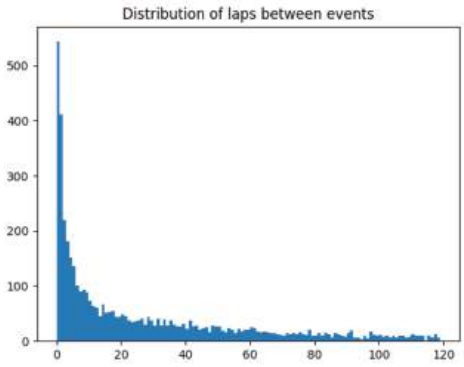



d) Dharma

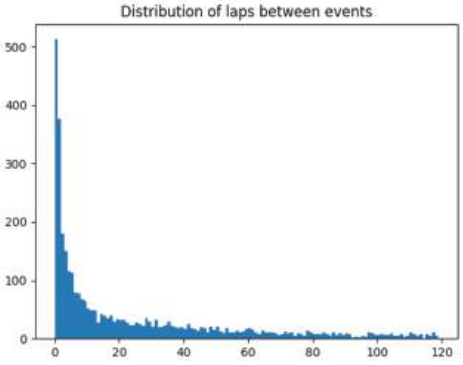
Figure 2: Evolution of the weekly number of tweets linked to selected ransomwares, excluding reply and retweets.



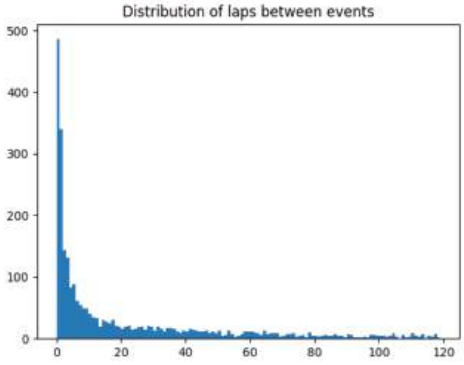

a) All laps



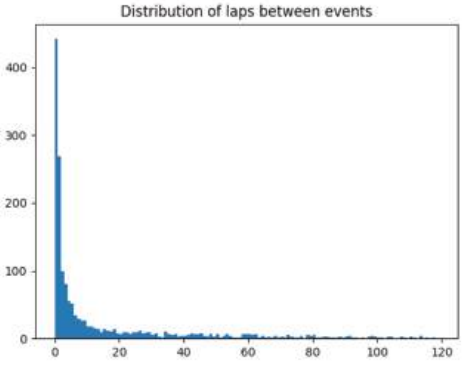

b) After a laps of 60' max



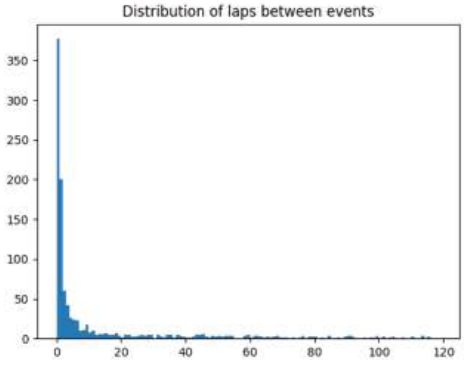

c) After a laps of 30' max




d) After a laps of 15' max




e) After a laps of 5' max




f) After a laps of 2' max

Figure 3: Evolution of the histograms of laps between Wannacry tweets with successive restrictions. In the plot a), all laps are considered. For other plots, we focus on laps lower than a given threshold, from 60 minutes to 2 minutes, before studying the histograms of laps following the previous selected laps.

intensity before the event $k > 1$, that is equal to $\mu$ for $k = 1$, and for the compensator at $T$ :

$$\text{for } h_\theta, \begin{cases} \lim\limits_{t \overset{<}{\to} \tau_k} \lambda^{(\mu,\theta)}(t) = \mu + \alpha \sum_{i=1}^{k-1} \mathbb{1}_{\tau_k - \tau_i \leq \Delta} e^{-\beta(\tau_k - \tau_i)}, \\ \\ \Lambda^{(\mu,\theta)}(T) = \mu T + \frac{\alpha}{\beta} \sum_{i=1}^{N^{(\mu,\theta)}(T)} \left(1 - e^{-\beta(\max(T-\tau_i, \Delta))}\right), \end{cases}$$

Next, we assume a deterministic bound for the support of the kernel approximately equal to 1 hour and a half, exactly $\Delta = 5000$. Resulting estimates of $(\alpha, \beta)$ are given in Table 3.

| Ransomware | $\mu$ | $\alpha$ | $\beta$ | $\|\|h_\theta\|\|_1$ |
|---|---|---|---|---|
| Wannacry | $7.248 \times 10^{-5}$ | $9.697 \times 10^{-4}$ | $1.022 \times 10^{-3}$ | 0.942 |
| Sodinokibi REvil | $1.469 \times 10^{-5}$ | $2.631 \times 10^{-4}$ | $2.906 \times 10^{-4}$ | 0.694 |
| Ryuk | $7.483 \times 10^{-6}$ | $2.432 \times 10^{-4}$ | $4.969 \times 10^{-4}$ | 0.449 |
| SamSam | $3.693 \times 10^{-6}$ | $3.427 \times 10^{-4}$ | $4.976 \times 10^{-4}$ | 0.631 |
| GrandCrab | $1.915 \times 10^{-6}$ | $1.476 \times 10^{-2}$ | $1.720 \times 10^{-2}$ | 0.858 |
| Maze | $9.372 \times 10^{-6}$ | $1.712 \times 10^{-4}$ | $3.306 \times 10^{-4}$ | 0.418 |
| Dharma | $1.360 \times 10^{-6}$ | $1.036 \times 10^{-4}$ | $3.394 \times 10^{-4}$ | 0.249 |

Table 3: Estimated parameters under the truncated exponential kernel model $h_\theta$ and the resulting value of $||h_\theta||_1$ .

The value of $||h_\theta||_1$ basically corresponds to the power of the auto-excitation phenomenon. Therefore, according to tweets timeline, we observe that Wannacry turns out to be the more contagious ransomware. This does not seems incoherent with regards to its worldwide spreading. However, an exhaustive comparison with cyber security indicator of ransomware spreading strenght may be appreciable.

## 4.3 Illustration of assistance needs scenarii

In this section, we apply the assistance needs bounds developped in this paper to the estimated Hawkes kernel. The obtained upper bounds for annual assistance needs quantiles are given in Table 4. We considered two scenarii regarding the duration of assistance need $\delta$. In the fisrt scenario we assume it to be equal to one hour and in the second scenario to be equal to one week. More precision on calculus are given in Section 6.3.

| Assistance duration | 1 hour | 1 hour | 1 week | 1 week |
|---|---|---|---|---|
| Ransomware | $q_{50\%}$ | $q_{99.5\%}$ | $q_{50\%}$ | $q_{99.5\%}$ |
| Wannacry | 23 028 | 94 099 | 132 232 | 534 588 |
| Sodinokibi REvil | 1 007 | 2 008 | 1 039 | 2 040 |
| Ryuk | 2 | 1 003 | 13 | 1 014 |
| SamSam | 2 | 1 003 | 9 | 1 010 |
| GrandCrab | 2 | 1 003 | 12 | 1 013 |
| Maze | 3 | 1 003 | 16 | 1 017 |
| Dharma | 1 | 1 002 | 2 | 1 003 |

Table 4: Upper bounds of quantiles 50% and 99.5% of the supremum of assistance needs $\sup_{0\leq t\leq T}\mathfrak{A}(t)$ over one year ($T = 1$ year), with two scenarii of assistance care durations, from one hour ($\delta = 3600$) to one week ($\delta = 3600 * 24 * 7$).

As expected, the Wannacry ransomware would need very high assistance capacity, especially in the second scenario. However for the others ransomwares, the extreme quantiles are quite similar on both scenarii. This stability may be appreciable for sizing assistance capacities, yet this observation is very empirical.

# 5 Conclusion

In this paper we illustrate the relevance of using point processes for understanding claims frequency. When clusters of claims are expected or observed, as for cyber risk and ransomware related claims, we investigate the use of Hawkes processes and their log-likelihood calibrations. For the tweets timelines considered in this paper, we empirically highlight the auto-excitation dependence between tweets and develop a framework for analyzing assistance needs considering tweets timelines as claims timelines.

We study the tail behaviour of the assistance needs stochastic process, relying on an analytical expression of its deviation from its expectation. We then study the tail behaviour of the deviation before obtaining a result on the assistance needs. We then apply those results for the calibrated Hawkes processes. This especially highlights the strength of the Wannacry tweets propagation and specifies the level of assistance capacity that can mitigate cluster of ransomware attacks. Yet our work is not based on insurance claims data, we try to demonstrate the relevance of using point processes for in-depth analysis of claims frequency. Our work can however be extended, regarding both the links

between kernel assumptions and mathematical bounds, but also regarding the assistance needs framework to manage the heterogeneity of claims.

# 6 Appendix

## 6.1 Point process likelihood

We aim to show that for a parametric point process and the log-likelihood linked with the observation of $(\tau_i)_{i\geq 1}$ can be expressed as follows,

$$\mathcal{L}_\theta\left(\mathfrak{I}(t),T\right)=\prod_{i=1}^{\mathfrak{I}(T)}\left(\lim_{t\overset{<}{\to}\tau_i}\lambda_\theta(t)\right)e^{-\Lambda_\theta(T)}.$$

For the demonstration we follow the steps of Rubin (1972).

$$\begin{aligned}\mathcal{L}_\theta\left(\mathfrak{I}(t),T\right) &= \mathbb{P}\left(\mathfrak{I}(\tau_1^-)=0\right)\mathbb{P}\left(\mathfrak{I}(\tau_1)=1\right)\\ &\times \prod_{i=1}^{\mathfrak{I}}\mathbb{P}\left(\mathfrak{I}(\tau_i^-)=\mathfrak{I}(\tau_{i-1})\right)\mathbb{P}\left(\mathfrak{I}(\tau_i)=\mathfrak{I}(\tau_{i-1})+1\right)\\ &\times \mathbb{P}(\mathfrak{I}(T)=\mathfrak{I}(\tau_{\mathfrak{I}(T)}))\end{aligned}$$

Since $\mathbb{P}(N(t^-)=N(s)|\mathcal{F}_s)=e^{-\int_s^t\lambda_\theta(u)du}$, the following calculus leads to the result.

$$\mathcal{L}_\theta\left(\mathfrak{I}(t),T\right)=\prod_{i=1}^{\mathfrak{I}(T)}\lambda_\theta(\tau_i)e^{-\int_0^{\tau_1}\lambda_\theta(s)ds}e^{-\sum_{i=2}^{\mathfrak{I}(T)}\int_{\tau_{i-1}}^{\tau_i}\lambda_\theta(s)ds}e^{-\int_{\tau_{\mathfrak{I}(T)}}^{\mathfrak{I}(T)}\lambda_\theta(s)ds}.$$

## 6.2 Proof of proposition 2.1 of Jaisson and Rosenbaum (2015)

This section follows the proof of Jaisson and Rosenbaum (2015) for the completeness of the paper.

Let us consider a Hawkes process $N^{(\mu,h)}$ with an intensity $\lambda^{(\mu,h)}_{N^{(\mu,h)}}$ defined by $(\mu,h)$. Let us write $\Lambda^{(\mu,h)}_{N^{(\mu,h)}}$ its compensator and $M_{N^{(\mu,h)}}=N^{(\mu,h)}-\Lambda^{(\mu,h)}_{N^{(\mu,h)}}$. Let us show that we can express the intensity as follows,

$$\lambda^{(\mu,h)}_{N^{(\mu,h)}}(s)=\mu\left(1+\int_0^s H(s-u)du\right)+\int_0^s H(s-z)dM_{N^{(\mu,h)}}(z).$$

Let us first give a result regarding $h$ and $H$ that will be useful in the following demonstration. We have that $H-h=H*h$. Indeed $H-h=\sum_{k\geq 2}(h^{*k-1}*h)=\sum_{k\geq 1}(h^{*k}*h)=H*h$ because of the linearity of the convolution product.

Then, let us recall the definition of $\lambda_{N^{(\mu,h)}}$ :

$$\lambda_{N^{(\mu,h)}}(s)=\mu+\int_0^s h(s-u)dN^{(\mu,h)}(u).$$

Then, using that $dN^{(\mu,h)}(u)=dM_{N^{(\mu,h)}}(u)+\lambda^{(\mu,h)}_{N^{(\mu,h)}}(u)du$, we get :

$$\lambda_{N^{(\mu,h)}}(s) = \mu + \int_0^s h(s-u)dM_{N^{(\mu,h)}}(u) + \int_0^s h(s-u)\lambda^{(\mu,h)}_{N^{(\mu,h)}}(u)du.$$

At this step, we rely on the following lemma, whose proof can for instance be found in Bacry and Muzy (2016) :

If $f(s) = b(s) + \int_0^s h(s-u)f(u)du$ with $b$ a locally bounded function, then :

$$f(s) = b(s) + \int_0^s H(s-u)b(u)du.$$

Its application leads to :

$$\lambda_{N^{(\mu,h)}}(s) = \mu + \int_0^s h(s-u)dM_{N^{(\mu,h)}}(u) + \int_0^s H(s-u)\left(\mu + \int_0^u h(u-z)dM_{N^{(\mu,h)}}(z)\right)du.$$

Because of the Fubini theorem, we get :

$$\int_0^s H(s-u)\int_0^u h(u-z)dM_{N^{(\mu,h)}}(z)du = \int_0^s \int_z^s H(s-u)h(u-z)dudM_{N^{(\mu,h)}}(z).$$

Then, letting $v = u - z$ :

$$\begin{array}{rcl} \int_0^s H(s-u)\int_0^u h(u-z)dM_{N^{(\mu,h)}}(z)du & = & \int_0^s \int_0^{s-z} H(s-z-v)h(v)dvdM_{N^{(\mu,h)}}(z) \\ & = & \int_0^s H * h(s-z)dM_{N^{(\mu,h)}}(z) \end{array}$$

Using $H - h = H * h$, we get :

$$\int_0^s H(s-u)\int_0^u h(u-z)dM_{N^{(\mu,h)}}(z)du = \int_0^s H(s-z)dM_{N^{(\mu,h)}}(z) - \int_0^s h(s-z)dM_{N^{(\mu,h)}}(z)$$

And finally get the result as follows :

$$\lambda^{(\mu,h)}_{N^{(\mu,h)}}(s) = \mu + \int_0^s h(s-u)dM_{N^{(\mu,h)}}(u) + \int_0^s H(s-u)\mu du + \int_0^s H(s-z)dM_{N^{(\mu,h)}}(z) - \int_0^s h(s-z)dM_{N^{(\mu,h)}}(z)$$

$$\lambda^{(\mu,h)}_{N^{(\mu,h)}}(s) = \mu\left(1 + \int_0^s H(s-u)du\right) + \int_0^s H(s-z)dM_{N^{(\mu,h)}}(z).$$

### 6.3 Settings for applications of Section 4

**Study of $\mathbb{E}[\mathfrak{A}(t)]$ :**

$$\forall t > 0,\ \mathbb{E}[\mathfrak{A}(t)] \leq \mu\delta \left(1 + \frac{1}{1 - \frac{\alpha}{\beta}\left(1 - e^{-\beta\Delta}\right)}\right).$$

**Study of $\mathfrak{H}$ :**

The analytic form of $\mathfrak{H}$ is hard to track when $\Delta < +\infty$. Therefore, we approximate the bound using results obtained for $\Delta = +\infty$, that is with non-truncated exponential kernel. We begin to show by recursivity that the following result for any $t > 0$ :

$$\text{Let } h_{(\alpha,\beta)}(t) = \alpha e^{-\beta t}. \text{ Then } \forall k \in \mathbb{N}^*, h^{*k}_{(\alpha,\beta)}(t) = \alpha^k \frac{t^{k-1}}{(k-1)!} e^{-\beta t}.$$

**Initialization :** $h^{*1}_{(\alpha,\beta)}(t) = h(t) = \alpha e^{-\beta t}$.

**Recursivity :** Let us suppose that for a $k \in \mathbb{N}^*$, $h^{*k}_{(\alpha,\beta)}(t) = \alpha^k \frac{t^{k-1}}{(k-1)!} e^{-\beta t}$.

Then, we have :

$$h^{*(k+1)}_{(\alpha,\beta)}(t) = \int_0^t h_{(\alpha,\beta)}(t-s) h^{*k}_{(\alpha,\beta)}(s) ds = \int_0^t \alpha e^{-\beta(t-s)} \alpha^{k-1} \frac{s^{k-2}}{(k-2)!} e^{-\beta s} ds$$

$$h^{*(k+1)}_{(\alpha,\beta)}(t) = \alpha^k e^{-\beta t} \int_0^t \frac{s^{k-2}}{(k-2)!} ds = \alpha^k e^{-\beta t} \left[\frac{s^{k-1}}{(k-1)!}\right]_0^t = \alpha^k \frac{t^{k-1}}{(k-1)!} e^{-\beta t}.$$

As a result,

$$H_{(\alpha,\beta)}(t) = \sum_{k\geq 1} h^{*k}_{(\alpha,\beta)}(t) = \sum_{k\geq 1} \alpha^k \frac{t^{k-1}}{(k-1)!} e^{-\beta t} = \alpha e^{-\beta t} \sum_{k\geq 1} \frac{(\alpha t)^{k-1}}{(k-1)!} = \alpha e^{(\alpha-\beta)t}.$$

When $\alpha < \beta$, which is the case under Assumption 1, $H_{(\alpha,\beta)}$ is decreasing, leading to

$$\mathfrak{H} = 1 + \int_0^\delta H_{(\alpha,\beta)}(v) dv = 1 + \frac{\alpha}{\alpha - \beta}\left(e^{(\alpha-\beta)\delta} - 1\right).$$

# References


d. e. d. r. a. a. i. C.-F. Agence nationale de la sécurité des systèmes d'information (ANSSI), Centre gouvernemental de veille. *État de la menace rançongiciel, à l'encontre des entreprises et des institutions*. 2020. URL `https://www.cert.ssi.gouv.fr/cti/CERTFR-2020-CTI-001/`.

E. Bacry and J.-F. Muzy. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016. 10.1109/TIT.2016.2533397.

A. Bonnet, C. Dion, F. Gindraud, and S. Lemler. Neuronal Network Inference and Membrane Potential Model using Multivariate Hawkes Processes. working paper or preprint, July 2021. URL https://hal.archives-ouvertes.fr/hal-03309709.

V. Chavez-Demoulin and J. McGill. High-frequency financial data modeling using hawkes processes. *Journal of Banking Finance*, 36(12):3415–3426, 2012. ISSN 0378-4266. https://doi.org/10.1016/j.jbankfin.2012.08.011. URL https://www.sciencedirect.com/science/article/pii/S0378426612002336. Systemic risk, Basel III, global financial stability and regulation.

D. V.-J. D. J. Daley. *An Introduction to the Theory of Point Processes : Volume II*. Springer, 2008. 10.1007/978-0-387-49835-5. URL https://doi.org/10.1007/978-0-387-49835-5.

R. L. Guével. Exponential inequalities for the supremum of some counting processes and their square martingales. *Comptes Rendus. Mathématique*, 359(8):969–982, 2021. 10.5802/crmath.206.

A. G. Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018. 10.1080/14697688.2017.1403131. URL https://doi.org/10.1080/14697688.2017.1403131.

T. Jaisson and M. Rosenbaum. Limit theorems for nearly unstable hawkes processes. *The Annals of Applied Probability*, 25(2):600–631, 2015. ISSN 10505164. URL http://www.jstor.org/stable/24519929.

B. D. Le, G. Wang, M. Nasim, and A. Babar. Gathering cyber threat intelligence from twitter using novelty classification. 2019. URL https://arxiv.org/abs/1907.01755.

P. Reynaud-Bouret and E. Roy. Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, 13(5):883 – 896, 2007. 10.36045/bbms/1170347811. URL https://doi.org/10.36045/bbms/1170347811.

I. Rubin. Regular point processesand their detection. *IEEE Transactions on information theory*, 18(5):547–557, 1972. URL https://ieeexplore.ieee.org/abstract/document/1054897.

F. T. Stéphane Lhuissier. Measuring cyber risk. 2021. URL https://blocnotesdeleco

.banque-france.fr/en/blog-entry/measure-evolution-cyber-risk.

M. F. Vogler Daniel. How users tweet about a cyber attack: An explorative study using machine learning and social network analysis. *Journal of Digital Media  Policy*, 11 (2):195–214, 2020. URL https://doi.org/10.1386/jdmp_00016_1.