Mémoire présenté le : 13 avril 2021

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Clément GERON

Titre : Méthode de projection ligne à ligne de la sinistralité pour la garantie Dommages Ouvrage

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

Membres présents du jury de l'Institut des Actuaires :

N. GARRIGUE

A. YOU

F. PERNOUD

Membre présent du jury de l'ISFA :

E. MASIELLO

Entreprise :
Nom : COVEA - MMA
Signature :

Directeur de mémoire en entreprise :
Nom : Pierre GOLHEN
Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : Actuariat non-vie, assurance construction, garanties décennales, Dommages Ouvrage, processus de Hawkes, copules, projection de la sinistralité, recours.

L'assurance Dommages Ouvrage est une des principales garanties que propose la branche de l'assurance construction. Elle permet à la personne pour laquelle sont effectués les travaux de se couvrir contre les dommages consécutifs à l'opération de construction, pendant une période de dix ans après la livraison de l'ouvrage. Elle repose sur un système à « double détente » spécifique à l'assurance construction, qui permet au maître d'ouvrage d'être indemnisé rapidement, avant que l'assureur n'engage des recours envers l'assureur de responsabilité civile décennale de l'intervenant responsable du sinistre.

La nature décennale de la garantie Dommages Ouvrage rend délicate l'approche de la projection de la sinistralité. Les modèles de projection communément utilisés dans l'actuariat tels que la méthode de Chain Ladder peuvent se révéler imprécis. De plus, les recours prennent une place primordiale dans le fonctionnement de la garantie. Ceci, ajouté au fait que de nombreux acteurs de l'assurance construction se sont désengagés dans les dernières années, implique que les montants de ces derniers doivent être considérés avec autant d'importance que les montants de règlement des sinistres.

Ce mémoire a pour ambition de présenter une méthode de projection de la sinistralité pour la garantie Dommages Ouvrage, mise en place afin d'estimer la rentabilité du portefeuille. Celle-ci s'inspire d'une méthode de projection ligne à ligne de la garantie responsabilité civile décennale proposée par Martin (2019). Le modèle avancé repose sur la simulation de scénarios en combinant l'utilisation de méthodes de *data science* classiques comme les arbres CART avec des outils plus complexes comme les processus de Hawkes ou l'explication de la dépendance entre plusieurs variables par copules. Cette approche projette individuellement chaque contrat du portefeuille en tenant compte de ses caractéristiques de souscription, mais aussi de sa sinistralité antérieure observée. Elle prend également en compte l'inflation en considérant une dimension temporelle, tant sur le moment de survenance des sinistres que sur leur durée de liquidation. L'inflation est en effet un élément capital pour une garantie qui engage un assureur sur plus d'une dizaine d'années après la souscription du contrat.

Abstract

Keywords : Non-life actuarial sciences, construction insurance, decennials guarantees, structural damages, Hawkes processes, copulas, claims forecasting, recourse.

Structural damages insurance is one of the main guarantees offered by the construction insurance branch. It allows the person for whom the work is carried out to cover himself against damages resulting from the construction operation, for a period of ten years after the building delivery. It is based on a "double trigger" system specific to the construction insurance, which allows the project owner to be compensated quickly, before the insurer initiates recourse to the decennial civil liability insurer of the worker responsible for the claim.

The decennial nature of the structural damage guarantee makes the projection claims approach difficult. The projection models commonly used in actuarial sciences such as the Chain Ladder method are imprecise. Moreover, recourse plays a key role in the guarantee operation. This, added to the fact that several construction insurers have withdrawn within the last years, implies that the recourse amounts must be considered with as much importance as the claims amounts payment.

This thesis aims to present a method for claims forecasting for the structural damage guarantee, implemented in order to estimate the portfolio profitability. This is inspired by a line-by-line projection method of the decennial civil liability guarantee proposed by Martin (2019). The offered model is based on scenarios simulations by combining the use of classical data science methods such as CART trees with more complex tools such as Hawkes processes or the explanation of the dependence between several variables by copulas. This approach projects individually each portfolio contract by taking into account its subscription characteristics, but also its observed previous claims experience. It also treats the inflation by considering a temporal dimension, for the occurrence of claims and for their liquidation period. Indeed, inflation is a capital element for a guarantee that commits an insurer for more than ten years after the contract subscription.

Remerciements

La réalisation de ce mémoire a été possible grâce à plusieurs personnes à qui je souhaite exprimer ma gratitude.

Je tiens tout d'abord à remercier chaleureusement Pierre GOLHEN, mon tuteur d'entreprise, qui a toujours su se montrer disponible pour répondre à mes interrogations. Il a su m'apporter avec pédagogie toute son expérience dans le domaine de l'assurance construction ainsi que de nombreux conseils sur les sujets techniques présentés dans ce mémoire.

Je souhaite aussi adresser toute ma reconnaissance à Erwan GALES, responsable de l'équipe « Actuariat Construction et Marchés Spécialisés », qui m'a accordé sa confiance pour cette année d'alternance. Il a également été d'un grand soutien sur les différents sujets traités, et a été déterminant dans le lancement de ma carrière professionnelle.

Je remercie toute l'équipe « Actuariat Construction et Marchés Spécialisés », et plus généralement l'ensemble de la direction Entreprise, de sa bienveillance qui m'a permis de me sentir rapidement intégré dans l'entreprise.

Je désire également remercier l'équipe pédagogique de l'Institut du Risque et de l'Assurance du Mans, en particulier Alexandre BROUSTE, de m'avoir apporté son aide sur certains sujets complexes.

Enfin, j'adresse mes remerciements à mon tuteur pédagogique Nicolas LEBOISNE, professeur et directeur de l'ISFA, pour ses conseils tout au long de la rédaction de ce mémoire. Je remercie plus généralement l'ensemble de l'équipe pédagogique de l'ISFA pour les enseignements de qualité dispensés pendant ma formation.

Table des matières

Chapitre 1

Introduction

La construction d'un ouvrage immobilier fait souvent appel à une multitude d'interlocuteurs. Du maître d'ouvrage au maître d'œuvre en passant par plusieurs types de réalisateurs, ces intervenants sont soumis à des règles strictes sur l'assurance de la construction. D'un côté, les individus ou entreprises prenant part aux travaux doivent couvrir leurs responsabilités sur les potentiels dommages causés à la construction. De l'autre, le maître d'ouvrage doit également se couvrir contre ces mêmes dommages dans l'objectif de percevoir une indemnisation dans un temps réduit. Du fait de la spécificité des garanties décennales, les techniques actuarielles classiques ne sont pas forcément adaptées à cette branche particulière de l'assurance.

Ces types d'assurances propres au domaine de la construction constituent une partie non négligeable du marché de l'assurance non-vie. Pour Covéa, SGAM leader français de l'assurance de biens et responsabilités réunissant les marques MMA, MAAF, et GMF, cette branche se place également à un rang important dans son activité en assurance non-vie. C'est au cœur de la direction entreprise de la marque MMA que cette étude a pu être réalisée.

Plusieurs travaux ont récemment été menés au sein de l'entreprise sur la projection des garanties de construction. Une nouvelle méthode a notamment été mise en place dans le cadre de la projection de la garantie responsabilité civile décennale, et a fait l'objet d'un mémoire soumis à l'Institut des Actuaires. Le modèle proposé permet d'obtenir une projection ligne à ligne de la sinistralité prenant en compte la déformation du portefeuille. Une transposition de cette méthodologie de projection a été proposée pour la garantie Dommages Ouvrage. Cependant, un aspect spécifique de ce produit rend une partie du modèle initial inadaptée. En effet, les recours prennent une place considérable dans le fonctionnement de l'assurance Dommages Ouvrage, ce qui n'est pas le cas pour l'assurance de responsabilité civile décennale.

Certaines parties du modèle actuel étant tout de même très efficaces, nous proposerons dans le cadre de ce mémoire de nouveaux procédés à intégrer dans la méthode initiale dans le but d'améliorer sa performance. L'utilisation d'un processus temporel pour l'estimation de la fréquence apportera une alternative paramétrique nécessitant un apprentissage moins lourd. La création de groupes de fréquence homogènes ainsi que la prise en compte de l'impact de la sinistralité observée dans l'historique du contrat permettront de garder un lien avec la modélisation en place. L'étude

de la structure de corrélation entre plusieurs variables nous guidera pour mieux appréhender le développement des sinistres, en particulier pour les montants de recours exercés.

Dans le but de présenter les nouvelles techniques proposées, nous commencerons par introduire l'environnement dans lequel a été menée cette étude ainsi que les caractéristiques spécifiques à l'assurance construction. Ensuite, il sera nécessaire de décrire les données et de présenter le modèle en place avant la réalisation de cette étude, pour mieux comprendre les besoins d'amélioration. Puis, nous présenterons en détails les apports de cette étude grâce à deux chapitres dédiés respectivement à la nouvelle méthode de projection de la fréquence des sinistres inconnus et à la structure de dépendance permettant d'obtenir une bonne prise en compte des recours. Enfin, nous conclurons sur les résultats obtenus et évoquerons les axes restant à améliorer.

Chapitre 2

Cadre de l'étude

L'objectif de ce chapitre est de fixer le cadre de rédaction de ce mémoire. Les enjeux de l'assurance construction pour MMA y seront développés, la nécessité d'engager des études approfondies sur le sujet de ce mémoire sera justifiée, et la démarche suivie sera introduite.

2.1 La place de l'assurance construction chez MMA

MMA (anciennement « Mutuelles du Mans Assurances ») est une entreprise du groupe Covéa. Fort de près de 200 ans d'activité, MMA est un acteur de proximité de la protection des particuliers, des professionnels, des entreprises, des associations et des collectivités locales. Grâce à son réseau de plus de 1000 agents généraux et au concours de 5000 courtiers partenaires, MMA s'investit pour assurer 3,1 millions de clients.

Dans le cadre de sa stratégie, le chiffre d'affaires concernant l'assurance auprès des professionnels et entreprises croît depuis plusieurs années. Le développement de ce secteur au sein de la compagnie place désormais le marché des entreprises en première position en termes de primes acquises. En effet, le chiffre d'affaires relatif à ce secteur représente près de 45 % du chiffre d'affaires de l'assurance IARD pour MMA.

L'assurance construction est classée chez MMA dans le secteur de l'assurance pour les professionnels et entreprises. Elle représente 17,6 % du chiffre d'affaires lié à ce secteur. Au sein de cette branche, MMA commercialise des garanties responsabilité civile décennale (45,6 % des primes perçues), responsabilité civile professionnelle (24,4 %), Dommages Ouvrages (17,4 %), tous risques chantier (2,7 %). La proportion du chiffre d'affaires restante correspond aux autres garanties appartenant à la branche construction.

Ainsi, la place importante de l'assurance construction chez MMA oblige la compagnie à posséder une bonne connaissance des risques qu'elle assure dans ce milieu. Afin d'être en adéquation avec les objectifs de rentabilité de l'entreprise, il convient de proposer une tarification maîtrisée et d'avoir la capacité d'identifier les typologies de risques à privilégier.

2.2 Motivations

Le régime décennal fonctionne de la manière suivante : la prime perçue par l'assureur à l'ouverture du chantier doit lui permettre de régler les sinistres pendant toute la période décennale, soit dix années à compter de la réception de l'ouvrage. La gestion des primes relatives aux garanties décennales est effectuée en capitalisation, ce qui rend le résultat sensible aux évolutions des marchés financiers. Or, dans le contexte économique actuel qui place les assureurs dans un environnement de taux bas, l'équilibre financier devient plus difficile à atteindre pour cette branche d'assurance. Les assureurs ne peuvent plus compter sur les produits financiers pour compenser une sous-évaluation du risque d'assurance, et doivent alors se recentrer sur leur cœur de métier pour assurer leur rentabilité.

De plus, la sinistralité observée les dernières années est en hausse, avec une augmentation annuelle moyenne des indemnisations supérieure à 5 % au niveau du marché. Cette croissance du coût de la sinistralité peut être expliquée par plusieurs facteurs. Tout d'abord, le développement insuffisamment maîtrisé de certaines techniques de construction a entraîné de la sinistralité sérielle. La sinistralité est aussi impactée par l'évolution de la règlementation du secteur du BTP qui provoque une inflation sur les indemnisations. Enfin, la crise subie par le secteur du bâtiment entre 2008 et 2015 est également l'une des causes probables, en ayant provoqué l'impossibilité de faire intervenir le service après-vente de l'entreprise ou de récupérer les franchises inopposables en responsabilité civile décennale. Bien que cette période de crise soit terminée, l'impact sur les assurances décennales subsiste puisque des contrats souscrits durant ces années sont toujours sous garantie. En outre, une nouvelle crise touchant le secteur du bâtiment se dessine à la suite de la crise sanitaire de la COVID-19 et pourrait entrainer des conséquences similaires.

Le marché de l'assurance construction a également beaucoup évolué au cours des dernières années. La précédente décennie, en particulier, a été marquée par le fort développement d'assureurs européens intervenant en libre prestation de service (LPS). Parmi les assureurs proposant ces offres, certains se sont concentrés sur des marchés considérés risqués au regard de la nature des activités ou de la situation de l'entreprise. Ces compagnies sont souvent de taille réduite et sans connaissance spécifique du régime français d'assurance construction. Les tarifs pratiqués sont le plus souvent ultra-compétitifs au regard des tarifs proposés par les compagnies établies en France. Cependant, depuis fin 2016, les difficultés de certains de ces assureurs se multiplient, puisque plusieurs acteurs se trouvent en situation de liquidation judiciaire ou se sont retirés du marché. Face à ces défaillances, les assureurs proposant des garanties Dommages Ouvrages peuvent alors craindre une recrudescence des défauts pour les recours à exercer.

L'ensemble des points précédemment détaillés montre que la branche de la construction est difficile à appréhender. L'engagement de l'assureur portant sur de longues périodes, il est toutefois primordial de maîtriser les risques qu'il assure. La compagnie doit alors être capable d'évaluer rapidement le risque inhérent à chaque police d'assurance, sous peine de se placer dans une situation délicate à cause d'un portefeuille sous tarifé pendant plusieurs années.

2.3 Problématique

La branche de l'assurance construction est difficile à appréhender pour les assureurs, de par ses spécificités qui la différencient des autres branches d'assurance IARD.

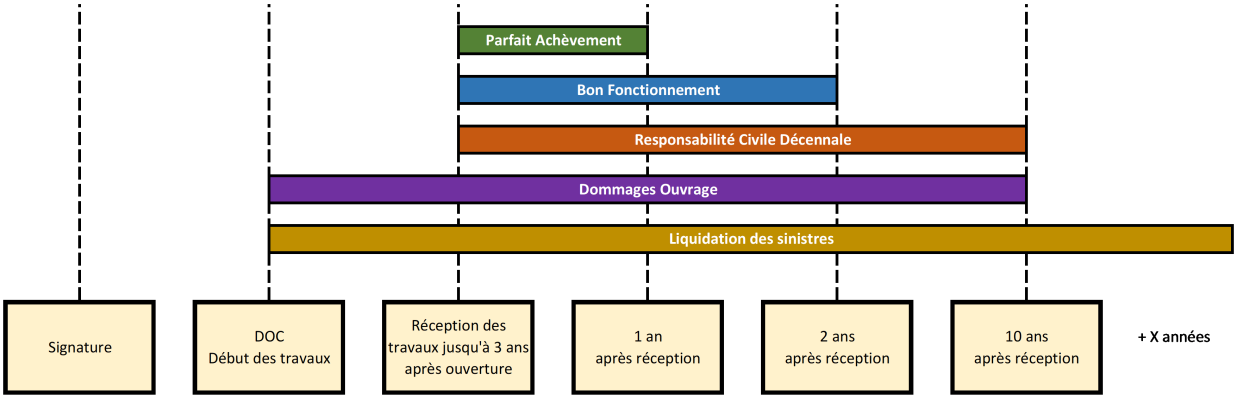

FIGURE 2.1 – Schéma représentatif des durées en assurance construction

Par définition, les garanties décennales, qui composent la majorité du portefeuille, couvrent les risques pendant une durée de dix ans après la réception du bien. Cette durée est très longue en comparaison de celle de la plupart des branches d'assurance non-vie. A cet effet, pour un contrat souscrit aujourd'hui, l'assureur ne connaitra le résultat final de ce dernier que 10 à 20 ans plus tard. Ainsi, il est primordial pour l'assureur d'avoir une connaissance précise et continue des risques qu'il possède en portefeuille sous peine de devoir réagir à retardement et de se retrouver dans une situation délicate.

Connaitre les risques du portefeuille revient alors à être capable de projeter le plus précisément possible le résultat ultime des contrats sous garantie, pour pouvoir adapter la souscription des polices futures. Or, les méthodes classiques de projection telles que Chain-Ladder ne sont pas adaptées à la projection au niveau contrat. En effet, ce type de méthode est pertinent dans le cas où l'assureur dispose d'une bonne connaissance de la sinistralité rapidement après la date d'effet du contrat. Ce n'est pas le cas en assurance construction, où la majorité des sinistres surviennent plusieurs années après la souscription du contrat. Par exemple, un contrat n'ayant pas subi de sinistre la première année ne pourra qu'être projeté à une sinistralité ultime nulle, la méthode Chain-Ladder s'appuyant sur des facteurs de développement.

Une autre des difficultés causées par le fait que l'assurance construction soit une branche longue est la sensibilité à l'inflation des indemnisations des sinistres pour un contrat. Effectivement, un sinistre survenant dans les premières années de la couverture décennale n'aura pas le même coût que ce même sinistre s'il survient pendant la dernière année de couverture, puisque l'économie aura évolué, tant pour le coût des matériaux que pour celui de la main d'œuvre. De la même façon, le temps nécessaire pour procéder au règlement d'un sinistre peut avoir un impact sur son coût.

Enfin, lorsque l'on se concentre sur l'assurance Dommages Ouvrage, on constate que les recours prennent une place très importante dans le résultat, du fait de la spécificité de la garantie. Une approche classique s'intéressant uniquement à la charge ultime nette de recours des sinistres peut manquer de précision. La variable relative aux montants de recours doit alors faire l'objet d'une étude à part entière et être considérée avec autant d'importance que les montants de règlement dans la projection.

2.4 Démarche et outils utilisés

2.4.1 Notations

Les notations utilisées dans ce mémoire, tant sur les aspects métier que théorique, se rapprochent au maximum de celles adoptées dans le mémoire de Martin (2019), afin que le lecteur puisse s'y référer. En particulier, on note :

DOC : Déclaration d'ouverture de chantier. Cette date correspond au début des travaux sur un chantier. C'est à partir de ce moment que les assurances construction, et en particulier la garantie Dommages Ouvrage, peuvent être mobilisées.

2.4.2 Outils informatiques utilisés

L'ensemble des travaux effectués dans le cadre de ce mémoire ont requis l'utilisation de plusieurs logiciels informatiques. Le logiciel SAS a été utilisé principalement pour l'extraction et la préparation des données. Le logiciel R a été utilisé pour les travaux de modélisation et de projection. Enfin, Excel a été utilisé pour la présentation de certains résultats.

2.4.3 Les produits sur mesure

Ce mémoire a été réalisé au sein d'un service actuariat spécialisé dans les produits d'assurance « sur mesure ». Même si la méthodologie proposée est transposable à l'ensemble des contrats des garanties décennales, l'apprentissage du modèle et la projection utilisent seulement ce type de contrats. Terminologie interne à MMA, un contrat est qualifié de « sur mesure » lorsque l'offre apportée sort du cadre standard de souscription. Dans ce cas, l'apporteur est dans l'obligation de solliciter MMA afin de construire un contrat adapté au risque. Dans le cas où l'offre entre dans le cadre standard de souscription, l'apporteur peut conclure l'affaire sans consulter MMA (on parle de souscription déléguée).

2.4.4 Introduction de la méthode

L'objectif des travaux réalisés dans ce mémoire est de mettre en place une méthode de projection des contrats de la garantie Dommages Ouvrage. Ce modèle de projection doit permettre d'estimer le résultat de cette garantie pour l'ensemble du portefeuille, que ce soit par exercice de souscription, mais aussi individuellement pour chaque contrat. Il est alors nécessaire de séparer le problème en plusieurs parties. Comme énoncé précédemment, les sinistres relatifs aux garanties décennales peuvent survenir plus de dix ans après la souscription du contrat. La sinistralité ultime des contrats est très peu représentée dans les premières années de ceux-ci. Il est alors primordial de pouvoir prévoir la sinistralité future d'un contrat.

Pour estimer cette sinistralité, il sera d'abord construit un modèle de fréquence, qui projettera la sinistralité future de chaque contrat. Ce modèle de fréquence s'appuiera sur la construction d'un arbre CART pour créer des classes de risques homogènes, avant d'ajuster un processus temporel permettant d'adapter la projection de chaque contrat en fonction de sa sinistralité observée.

Pour faire suite à ce modèle de fréquence, un modèle de coût des sinistres sera proposé. Grâce à la théorie des copules, la méthode décrite permettra d'attribuer aux sinistres inconnus un montant de règlement, mais aussi un montant de recours et une « durée de vie » du sinistre. Ce modèle s'appuie également sur des arbres de classification et diverses techniques actuarielles couramment utilisées.

Pour une projection complète du portefeuille, il est également nécessaire de s'intéresser à l'évolution des sinistres connus et non clos. La méthode utilisée est alors directement empruntée aux travaux proposés antérieurement par Martin (2019) dans le cadre de son mémoire. Ce mémoire, sur lequel s'appuient l'ensemble des travaux du présent rapport, sera résumé dans le cinquième chapitre.

2.4.5 Initiative de recherche Covéa - IRA - Ecole Polytechnique

Les travaux effectués s'inscrivent dans le cadre de l'initiative de recherche formée par Covéa, l'Institut du Risque et de l'Assurance (IRA) du Mans et l'Ecole Polytechnique. Pour Covéa, celle-ci a pour intérêt d'apporter un soutien théorique à des problématiques métier, en proposant des pistes d'amélioration aux méthodes utilisées.

Chapitre 3

Les spécificités de l'assurance construction

Une opération de construction regroupe l'ensemble des travaux de réalisation d'un ou plusieurs ouvrages, et peut être séparé en deux étapes.

La première regroupe l'ensemble des procédures effectuées avant l'ouverture du chantier, que ce soit sur le plan administratif (permis de construire, autorisation d'urbanisme, etc.), financier (prêts), ou technique (études de géomètres, géologues ou ingénieurs). La seconde phase rassemble tous les travaux effectués entre la déclaration d'ouverture de chantier (DOC) et la réception de l'ouvrage (fin des travaux). Durant cette phase, les différents réalisateurs interviennent pour le terrassement du terrain, le gros-œuvre et le second œuvre.

La construction d'un ouvrage fait donc appel à de nombreux intervenants et une multitude de potentiels dommages peuvent être subis par l'édifice. Plusieurs produits d'assurance permettent alors de se couvrir contre la survenance de ces derniers selon la règlementation en vigueur et le rôle de chacun sur le chantier.

3.1 Les intervenants

Une opération de construction, quelle que soit sa taille, fait intervenir différents types d'interlocuteurs que l'on peut décomposer en plusieurs catégories.

3.1.1 Le maître d'ouvrage

Le maître d'ouvrage est la personne morale ou physique pour laquelle sont effectués les travaux. Cette personne peut être un particulier, une société privée, une collectivité locale, l'État, ou encore des maîtres d'ouvrage professionnels faisant construire en vue de la revente (promoteurs).

Son rôle est de définir le programme de construction. Il fixe pour cela ses exigences en matière de prix, de délais et de qualité. Avant le lancement des travaux, il appartient au maître d'ouvrage d'attester de la faisabilité du projet et de son financement. Il doit aussi choisir le maître d'œuvre et les différentes entreprises qui interviendront sur le chantier (souvent à l'issue d'appels d'offre). Une fois les travaux lancés, il s'assure de leur bon déroulement et doit procéder à la réception de l'ouvrage une fois l'opération de construction terminée. La réception est l'acte par lequel le maître d'ouvrage déclare accepter l'ouvrage avec ou sans réserve.

3.1.2 Le maître d'œuvre

Le maître d'œuvre est une personne morale ou physique qui est chargée de concevoir la construction, diriger et contrôler l'exécution des travaux. Ce peut être un architecte, un bureau d'étude, ou encore un économiste de la construction.

Lié au maître d'ouvrage par un contrat de louage d'ouvrage, son rôle est d'orchestrer la réalisation de l'ouvrage, en garantissant le respect des délais, des coûts et du cahier des charges de la construction. Il définit en amont le budget à prévoir, conçoit les plans et apporte son assistance au maître d'ouvrage pour le choix des entreprises. Il est alors chargé de planifier l'intervention des entreprises sur le chantier, et s'assure du respect des délais et du coût des artisans, tout en vérifiant que le chantier répond parfaitement aux normes en vigueur. Enfin, il peut assister le maître d'ouvrage au moment de la réception des travaux.

3.1.3 Les réalisateurs

Les réalisateurs (parfois appelés entrepreneurs) sont l'ensemble des entreprises, artisans ou travailleurs indépendants qui participent intellectuellement ou matériellement à l'édification de l'ouvrage. Ce peut être une entreprise générale du bâtiment, c'est-à-dire qui exerce plusieurs activités liées à la construction, ou une entreprise titulaire d'un lot (une seule activité).

Ces sociétés sont recrutées par le maître d'ouvrage et pilotées techniquement par le maître d'œuvre. Elles sont liées au maître d'ouvrage par un contrat de louage d'ouvrage, de la même manière que le maître d'œuvre, et ont donc une obligation de résultat sur les travaux dont elles sont chargées. Une multitude de corps de métiers peuvent intervenir sur le chantier au titre de réalisateurs, on peut par exemple citer les entreprises de gros œuvre, de plomberie, d'électricité, de carrelage, de peinture, etc.

3.1.4 Autres intervenants

3.1.4.1 Les sous-traitants

Les sous-traitants sont chargés par les réalisateurs d'effectuer une partie des travaux qui leur ont été confiés par le maître d'œuvre. Ils n'ont pas de lien direct avec le maître d'ouvrage. Un sous-

traitant n'est pas assujetti à la responsabilité civile décennale, c'est-à-dire qu'il n'est pas présumé responsable en cas de dommages sur la construction. Cependant, il n'est pas pour autant libéré de toute responsabilité en cas de sinistre, puisqu'il a une obligation de résultat envers le réalisateur qui a sous-traité.

3.1.4.2 Les fournisseurs

Les fournisseurs mettent à disposition des entreprises le matériel et les matériaux nécessaires à la réalisation du chantier. Comme les sous-traitants, ils sont généralement en lien direct avec les réalisateurs et ne sont pas présumés responsables des malfaçons. La responsabilité décennale du fournisseur de matériaux peut tout de même être impliquée s'il joue un rôle actif sur le chantier, à la suite d'une jurisprudence datant de 2018.

3.1.4.3 Les partenaires du maître d'ouvrage

Le maître d'ouvrage a la possibilité ou l'obligation de faire appel à des partenaires pour mener à bien l'ensemble des travaux.

Il peut faire appel à **un conducteur d'opération**, dans le cas où il ne dispose pas des capacités nécessaires pour piloter l'opération. Il peut alors se faire assister sur le plan administratif, financier et technique, par une personne publique ou privée. Les missions qui lui sont confiées sont incompatibles avec les missions appartenant au maître d'œuvre.

Le coordinateur Sécurité et Protection de la Santé (SPS) est désigné pour veiller à ce que les principes généraux de prévention soient respectés. Il doit assurer la santé et la sécurité de l'ensemble des travailleurs sur le chantier de construction. Il est présent de la conception à la réception de l'opération, pour procéder à l'évaluation des risques sur le chantier. Pendant la réalisation du chantier, il gère les risques engendrés par les coactivités simultanées ou successives des différentes entreprises, ainsi que les risques provoqués par l'environnement du chantier.

Le contrôleur technique est engagé par le maître d'ouvrage pour vérifier la qualité des constructions et leur fiabilité pour assurer la sécurité des personnes. C'est un professionnel de l'immobilier dont la mission consiste à prévenir les défaillances techniques d'un chantier susceptibles d'engendrer des sinistres. Pour cela, il effectue une analyse de risques sur la base des documents de conception et de réalisation portés à sa connaissance. Il émet alors un avis en conséquence et rédige ensuite des rapports sur ses missions de vérification. L'exercice de la profession de contrôleur technique est soumis à un agrément ministériel et à une obligation d'assurance décennale.

3.2 Les garanties

Une opération de construction engendre une multitude de risques, pouvant provoquer des dommages pendant la réalisation des travaux, mais qui peuvent également se révéler après la réception

de l'ouvrage. Les différents interlocuteurs ont alors la possibilité ou l'obligation de se couvrir contre la survenance de sinistres, grâce à divers types d'assurances.

3.2.1 L'assurance de Responsabilité Civile Décennale Obligatoire (RCDO)

La responsabilité civile décennale est une responsabilité civile de plein droit pesant sur les constructeurs ou professionnels du bâtiment vis-à-vis du maître d'ouvrage pendant dix ans. L'assurance de cette dernière prévoit alors la couverture, pendant dix ans après la réception du chantier, des malfaçons subies par l'ouvrage et provoquées par ces intervenants.

La loi oblige les maîtres d'œuvre, les réalisateurs, les fournisseurs et les contrôleurs techniques à souscrire une assurance de ce type lorsqu'ils interviennent dans la construction d'un ouvrage neuf où dans une rénovation. Ces intervenants doivent être capables de justifier au maître d'ouvrage qu'ils disposent d'une telle garantie lors d'une déclaration d'ouverture de chantier. En cas de vente d'un logement dans les dix ans suivant sa construction, la mention de l'existence ou non des assurances obligatoires doit être annexée au contrat de vente. Ceci afin de permettre à l'acquéreur d'agir en cas de sinistre.

Cette assurance se déclenche lorsque l'assuré est déclaré responsable de vices ou dommages de construction pouvant affecter la solidité de l'ouvrage ou le rendre impropre à l'usage auquel il est destiné. Ces dommages peuvent résulter de défauts de conformité ou d'un vice de sol. Ces défauts concernent le gros-œuvre (murs, toiture, etc.), le second œuvre (électricité, isolation, etc.) mais aussi certains équipements lorsque le dysfonctionnement rend le bien impropre à son usage (système de chauffage). Une assurance RCDO comporte de nombreuses clauses obligatoires définies par le code des assurances.

Ce type d'assurance est donc obligatoirement souscrit par la plupart des intervenants sur le chantier pour le compte du maître d'ouvrage, que ce soit pour la conception ou la réalisation. De son côté, le maître d'ouvrage doit souscrire une garantie Dommages Ouvrage.

3.2.2 L'assurance Dommages Ouvrage (DO)

L'assurance Dommages Ouvrage permet à l'assuré, en cas de sinistre, d'être remboursé rapidement de la totalité des travaux de réparation des dommages couverts par la responsabilité civile décennale. Par complémentarité avec l'assurance RCDO, c'est une assurance décennale, c'est-à-dire qu'elle couvre l'assuré pendant dix ans à compter de la réception du chantier.

C'est une assurance obligatoire pour toute personne qui engage une ou plusieurs entreprises pour effectuer des travaux de construction. Elle doit être souscrite en amont de l'ouverture des travaux.

Ce type d'assurance couvre le même type de dommages que la RCDO. L'assurance Dommages Ouvrage permet de procéder aux remboursements ou à l'exécution des réparations sans recherche préalable de responsabilité. En effet, la recherche de responsabilité peut être longue dans le cas de

l'attente d'une décision de justice. Il est aussi important pour le maître d'ouvrage de souscrire cette garantie dans l'optique d'une potentielle revente du bien, puisqu'il est personnellement responsable vis-à-vis du nouvel acquéreur de toutes les conséquences résultant d'un défaut d'assurance.

Du point de vue de l'assureur, après avoir versé l'indemnisation à l'assuré dans le cadre d'un dommage à l'ouvrage, celui-ci procède ensuite à une recherche de responsabilité et exerce des recours contre les assureurs en responsabilité décennale en cas de responsabilité du constructeur.

De la même manière que pour la garantie RCDO, le code des assurances définit les clauses obligatoires d'un contrat DO.

3.2.3 Les garanties facultatives

Complémentairement aux deux garanties décennales décrites précédemment, les différents acteurs du chantier peuvent également se protéger contre d'autres risques engendrés par l'opération de construction. Le champ du possible pour les assureurs est alors très large, puisque de nombreuses garanties facultatives peuvent être proposées, sous la condition de présence d'un aléa.

Ces garanties facultatives peuvent être adossées ou non aux garanties décennales obligatoires. Il est alors possible de soumettre une liste non exhaustive d'assurances facultatives proposées sur le marché :

- L'assurance effondrement ou risque d'effondrement, souscrite par les constructeurs, prévoit la couverture des frais de démolition, déblaiement et reconstruction de l'ouvrage après effondrement dans le cas où une menace grave et imminente d'effondrement pèse sur l'ouvrage en construction.
- L'assurance Tous Risques Chantier (ou TRC) a pour but de garantir les dommages matériels pouvant survenir en cours de chantier. Elle couvre les dommages affectant l'ouvrage en cours de réalisation, mais aussi potentiellement les installations annexes nécessaires aux travaux. Ces dommages peuvent être causés par un événement extérieur (vandalisme, vol) ou lié aux travaux (mauvaise conception, vice de matériaux, etc.).
- La garantie des dommages immatériels intervient en complément de la garantie Dommages Ouvrage. Elle permet de couvrir les dommages immatériels en prenant en charge des frais liés, par exemple, à des troubles de jouissance, à des pertes de loyer ou à des pertes d'exploitation. Elle est notamment utile si les dommages relevant de la garantie décennale (ou leurs travaux de réparation) entraînent l'impossibilité d'habiter ou d'utiliser le bien.

L'ensemble de ces garanties, qu'elles soient obligatoires ou facultatives, sont soumises à une règlementation que doivent respecter les assureurs, en matière juridique mais aussi comptable.

3.3 Règlementation

3.3.1 Loi Spinetta

La loi Spinetta du 4 janvier 1978, entrée en vigueur au 1er janvier 1979, fixe la réglementation relative à la responsabilité et à l'assurance dans le domaine de la construction. Le principal objectif de cette loi est de protéger le propriétaire du bien sinistré en procédant à une indemnisation rapide, tout en déterminant la responsabilité de chacun des intervenants dans le dommage. Le système d'assurance mis en place par cette loi est à double détente et repose sur la présomption de responsabilité des constructeurs.

3.3.1.1 Présomption de responsabilité

Dans un premier temps, l'article 1792 du code civil issu de la loi Spinetta stipule que *« Tout constructeur d'un ouvrage est responsable de plein droit, envers le maître ou l'acquéreur de l'ouvrage, des dommages, même résultant d'un vice du sol, qui compromettent la solidité de l'ouvrage ou qui, l'affectant dans l'un de ses éléments constitutifs ou l'un de ses éléments d'équipement, le rendent impropre à sa destination. Une telle responsabilité n'a point lieu si le constructeur prouve que les dommages proviennent d'une cause étrangère. »*

Cette présomption de responsabilité des constructeurs dispense alors le maître d'ouvrage de prouver la faute des constructeurs. Pour s'exonérer de sa responsabilité, le constructeur doit alors prouver que le sinistre a été causé par la victime, un tiers, ou par une force majeure.

3.3.1.2 Système à double détente

Dans un second temps, la loi Spinetta impose la souscription de deux types d'assurances lors d'une opération de construction :

- Une assurance Dommages Ouvrage, souscrite par le maître d'ouvrage ;
- Une assurance de responsabilité civile décennale, obligatoire pour chaque constructeur.

En effet, ces obligations d'assurance sont énoncées dans deux articles du code des assurances. Dans un premier temps, l'article L241-1 notifie que *« Toute personne physique ou morale, dont la responsabilité décennale peut être engagée sur le fondement de la présomption établie par les articles 1792 et suivants du code civil, doit être couverte par une assurance. A l'ouverture de tout chantier, elle doit justifier qu'elle a souscrit un contrat d'assurance la couvrant pour cette responsabilité. Tout candidat à l'obtention d'un marché public doit être en mesure de justifier qu'il a souscrit un contrat d'assurance le couvrant pour cette responsabilité. Tout contrat d'assurance souscrit en vertu du présent article est, nonobstant toute stipulation contraire, réputé comporter une clause assurant le maintien de la garantie pour la durée de la responsabilité décennale pesant sur la personne assujettie à l'obligation d'assurance. »* L'article L242-1 du code des assurances énonce ensuite que

« Toute personne physique ou morale qui, agissant en qualité de propriétaire de l'ouvrage, de vendeur ou de mandataire du propriétaire de l'ouvrage, fait réaliser des travaux de construction, doit souscrire avant l'ouverture du chantier, pour son compte ou pour celui des propriétaires successifs, une assurance garantissant, en dehors de toute recherche des responsabilités, le paiement de la totalité des travaux de réparation des dommages de la nature de ceux dont sont responsables les constructeurs au sens de l'article 1792-1, les fabricants et importateurs ou le contrôleur technique sur le fondement de l'article 1792 du code civil. »

Le système est alors qualifié de « à double détente » car l'assurance Dommages Ouvrage doit dans un premier temps indemniser (sous soixante jours maximum) l'assuré au titre des garanties prévues au contrat. Cette indemnisation est destinée à financer les travaux de réparation des dommages. Dans un second temps, il incombe à l'assureur Dommages Ouvrage d'exercer des recours contre les responsables des dommages subis par l'ouvrage, ou généralement envers leur assureur de responsabilité civile décennale.

3.3.1.3 Sanctions en cas de non-assurance

L'absence de souscription d'une assurance fait encourir des sanctions pénales aux professionnels concernés. En effet, L'article L243-3 du code des assurances stipule que *« Quiconque contrevient aux dispositions des articles L. 241-1 à L. 242-1 du présent code sera puni d'un emprisonnement de six mois et d'une amende de 75 000 euros ou de l'une de ces deux peines seulement. Les dispositions de l'alinéa précédent ne s'appliquent pas à la personne physique construisant un logement pour l'occuper elle-même ou le faire occuper par son conjoint, ses ascendants, ses descendants ou ceux de son conjoint. »*

3.3.2 Les garanties de parfait achèvement et de bon fonctionnement

En complément de la responsabilité civile décennale, les constructeurs sont responsables vis-à-vis du maître d'ouvrage du parfait achèvement et du bon fonctionnement de la construction.

L'article 1792-6 du Code Civil stipule que *« La garantie de parfait achèvement, à laquelle l'entrepreneur est tenu pendant un délai d'un an, à compter de la réception, s'étend à la réparation de tous les désordres signalés par le maître de l'ouvrage, soit au moyen de réserves mentionnées au procès-verbal de réception, soit par voie de notification écrite pour ceux révélés postérieurement à la réception. »*

L'article 1792-3 indique que les équipements de l'ouvrage n'entrant pas dans le cadre de la présomption de responsabilité détaillée précédemment font l'objet d'une garantie de bon fonctionnement d'une durée minimale de deux ans à compter de la réception du chantier.

3.3.3 Convention de Règlement de l'Assurance Construction (CRAC)

Consécutivement à l'entrée en vigueur de la loi Spinetta en 1978, un agrément a été signé par la plupart des assureurs intervenant sur le marché de l'assurance construction dans le but d'améliorer le fonctionnement du dispositif d'indemnisation. Mise en place en 1983, la CRAC a pour objectif de faciliter l'instruction des sinistres et d'en accélérer le paiement au bénéfice de la victime.

Cette convention facilite l'application de la loi Spinetta pour les assureurs en simplifiant la gestion des sinistres. Lors de la survenance d'un sinistre, un unique expert est mandaté par l'assureur Dommages Ouvrage pour le compte de tous les assureurs concernés par le chantier. Cet expert rédigera un rapport dans lequel le coût du sinistre et les responsabilités des divers intervenants sont évalués. Sur la base des conclusions de l'expert, les différents assureurs prennent position sur la mise en jeu de leurs garanties. À la suite de cela, les assureurs des constructeurs responsables disposent de trois mois pour procéder au remboursement, en prenant en compte un barème de responsabilité préétabli par la convention. Le recours exigé prend en compte, en plus du cout du sinistre, 50 % des honoraires et des frais d'expertise, sous déduction d'un ticket modérateur indexé de 1600 euros. Dans le cas où l'indemnité réglée n'atteint pas le montant du ticket modérateur, l'assureur de Dommages Ouvrage ne présente pas de recours et garde à sa charge la totalité des frais et honoraires.

Cette convention entraîne donc une réduction des frais de gestion et d'expertise, et met en place des règles pour faciliter les recours de l'assureur Dommages Ouvrage envers les assureurs de responsabilité.

3.4 Provisionnement

Bien que ce mémoire n'ait pas pour but principal le calcul des provisions, il semble indispensable d'évoquer le provisionnement exigé par la règlementation comptable en assurance construction.

3.4.1 Les provisions pour sinistres à payer (PSAP)

La branche de l'assurance construction, tout comme les autres branches d'assurance non-vie, requiert un provisionnement pour les sinistres à payer. L'article R343-7 du code des assurances définit cette provision comme la *« valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise ».*

C'est une provision calculée exercice par exercice. Selon l'article 143-10 du règlement de l'Autorité des Normes Comptables (ANC), l'évaluation de celle-ci doit être *« effectuée dossier par dossier, le coût d'un dossier comprenant toutes les charges externes individualisables ; elle est augmentée d'une estimation du coût des sinistres survenus mais non déclarés ».*

Pour résumer, cette provision comprend :

- Le coût total, estimé dossier par dossier, des sinistres qui ont été déclarés jusqu'à la date de l'inventaire, diminué des règlements déjà effectués et des frais déjà payés ;
- Une estimation prudente des sinistres survenus mais non encore déclarés (IBNyR - Incured But Not yet Reported).

De plus, ce même article rajoute que cette provision *« doit toujours être calculée pour son montant brut, sans tenir compte des recours à exercer ; les recours à recevoir font l'objet d'une évaluation distincte »*. Cette remarque est importante pour la branche de l'assurance construction, sujette à l'exercice de recours, notamment pour la garantie Dommages Ouvrage.

3.4.2 Les provisions pour sinistres non encore manifestés (PSNEM)

En supplément des provisions pour sinistres à payer, l'ANC impose un provisionnement spécifique aux garanties décennales d'assurance construction : Les provisions pour sinistres non encore manifestés. En effet, selon l'article 143-13 du règlement, l'assureur doit prévoir une provision correspondant à l'estimation *« du coût des sinistres non encore manifestés et qui devraient se manifester d'ici à l'expiration de la période de prescription décennale ».*

En d'autres termes, la provision pour sinistres non encore manifestés correspond au montant estimé des sinistres à survenir pendant la période de garantie restante au-delà de l'exercice de l'inventaire. La nécessité d'une telle provision s'explique par le fait que la manifestation des sinistres pour un contrat s'étale sur une période supérieure à dix ans, contrairement aux autres branches de l'assurance non-vie. Il s'agit alors de constituer une provision destinée à couvrir les sinistres futurs. L'ANC impose alors une méthode d'estimation des PSNEM dans l'article 143-14 de son règlement :

« Les entreprises calculent, pour chaque exercice d'ouverture de chantier, séparément pour les garanties décennales de responsabilité civile et pour les garanties décennales de dommage aux ouvrages, l'ancienneté n des chantiers ainsi que les montants A_n et B_n, définis comme suit :

- *n = différence de millésime entre l'exercice sous inventaire et l'exercice d'ouverture des chantiers ;*
- *A_n = coût total, estimé dossier par dossier, des sinistres afférents aux garanties décennales d'assurance construction délivrées pour des chantiers d'ancienneté n et qui se sont manifestés jusqu'à la date de l'inventaire, diminué des recours encaissés ou à encaisser ;*
- *B_n = montant des primes émises et des primes restant à émettre, nettes des primes à annuler et des frais d'acquisition, afférent à ces mêmes garanties. L'estimation des sinistres non encore manifestés, effectuée séparément pour les garanties décennales de responsabilité civile et pour les garanties décennales de dommage aux ouvrages, est égale au plus élevé des deux montants MS_n et MP_n suivants :*
- $MS_n = a_n \times A_n$
- $MP_n = b_n \times B_n$

a_n et b_n prenant les valeurs suivantes :

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
a_n	0	0	3,4	2	1,4	1	0,7	0,5	0,35	0,25	0,20	0,15	0,10	0,05
b_n	1	1	0,95	0,85	0,75	0,65	0,55	0,45	0,35	0,25	0,20	0,15	0,10	0,05

TABLE 3.1 – Coefficients utilisés pour le calcul des PSNEM

3.5 Chiffres clés du marché

Les chiffres présentés ci-après proviennent des données clés de l'assurance de biens et de responsabilité publiés par la fédération française de l'assurance (FFA).

En 2018, le marché de l'assurance français a perçu 219,4 milliards d'euros de cotisations. Sur l'ensemble de ces primes, 74,4 % sont consacrées aux assurances de personnes contre 25,6 % pour les assurances de biens et responsabilité. L'assurance construction représente 1 % des primes perçues, ce qui représente 3,9 % des cotisations des assurances de biens et responsabilité. La part de l'assurance construction dans les prestations versées est sensiblement la même. Sur la totalité des primes versées dans le cadre de l'assurance construction, 28 % sont consacrées à la garantie Dommages Ouvrage.

Même si une tendance légèrement à la baisse des cotisations consacrées à l'assurance construction est constatée sur les dernières années, l'évolution est plutôt stable, que ce soit pour la garantie responsabilité civile décennale ou Dommages Ouvrage.

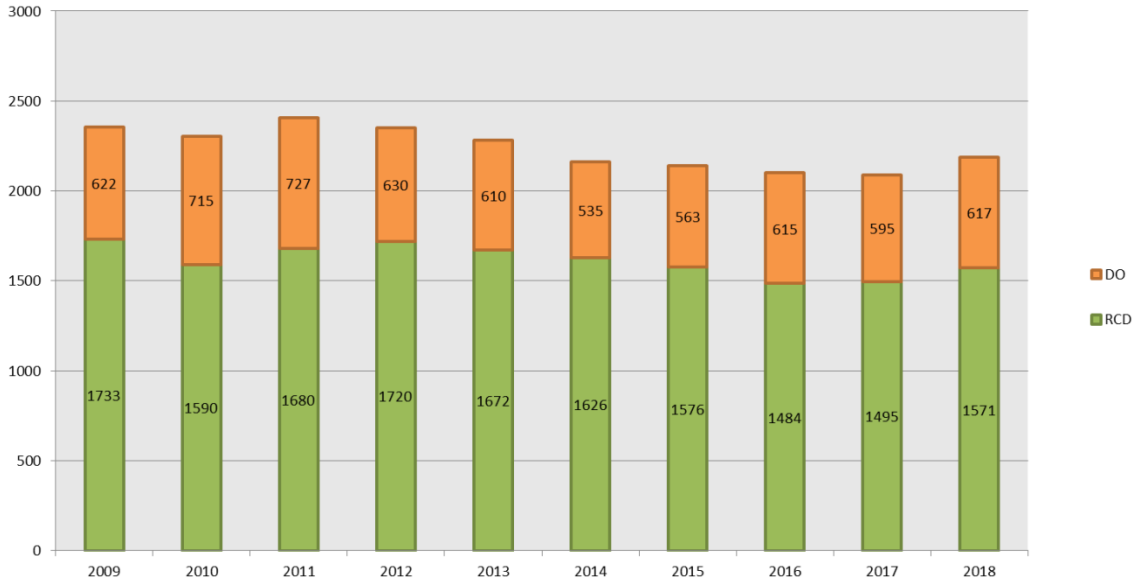

FIGURE 3.1 – Evolution des cotisations de l'assurance construction

Au contraire, la tendance des prestations sur les mêmes années est à la hausse, ce qui fait qu'une partie plus importante de la prime est consacrée au règlement des sinistres. Cette augmentation du

rapport sinistres à primes motive le sujet de ce mémoire en questionnant l'évolution de la rentabilité du secteur.

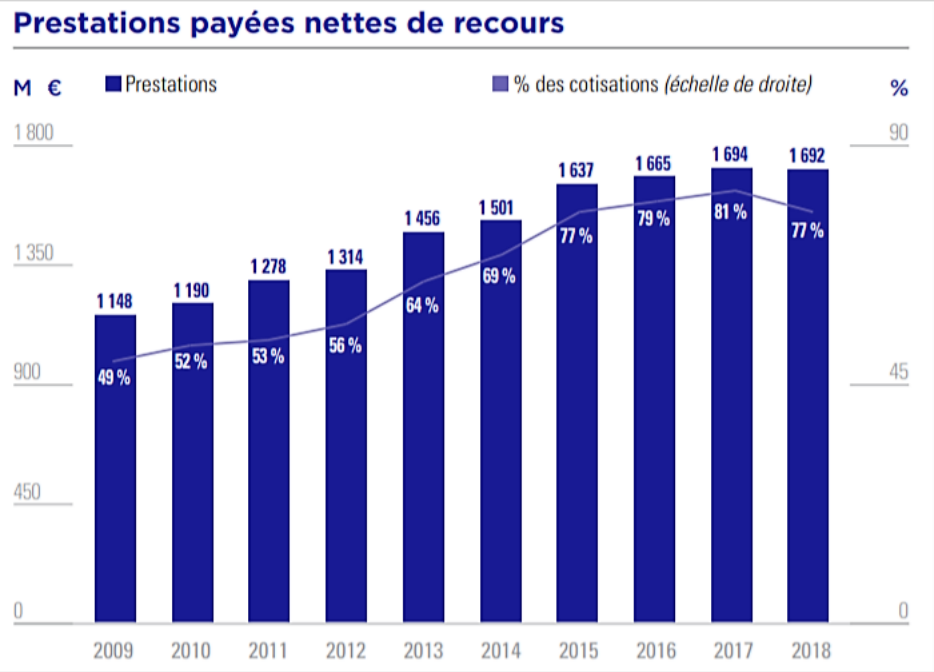

FIGURE 3.2 – Evolution des prestations de l'assurance construction

Chapitre 4

Travaux préliminaires sur les données

Avant de s'intéresser aux étapes de modélisation, une phase de préparation et d'exploration des données est nécessaire pour s'assurer de la cohérence des démarches entreprises.

4.1 Préparation des données

La démarche de préparation des données est une étape primordiale dans la construction d'un modèle de projection. En effet, un modèle qui s'appuierait sur des données incohérentes conduirait à des résultats et conclusions sans valeurs. De ce fait, il est nécessaire de s'intéresser à celles-ci afin de vérifier leur pertinence. Par suite, des hypothèses sont émises afin de rendre les données propres à la modélisation. Cette section s'intéresse alors à ces hypothèses.

4.1.1 Actualisation des données

L'évolution de l'économie au cours des années implique que les données monétaires ne doivent pas être considérées de la même manière selon l'ancienneté du versement. En effet, l'inflation constatée dans les dernières années doit être prise en compte pour pouvoir travailler uniquement sur des montants correspondant à des évènements qui surviendraient aujourd'hui. Seule l'étape d'actualisation des montants relatifs à la sinistralité sera présentée, puisque les primes n'interviennent pas dans le modèle de projection. Les données actualisées seront alors utilisées dans la phase d'apprentissage de la modélisation présentée dans la suite de ce mémoire.

Les coefficients d'actualisation doivent être déterminés pour représenter au plus juste l'évolution des prix dans le secteur de la construction. Pour cela, l'utilisation d'un indice de place est privilégiée à la construction d'un indice interne. L'INSEE référence chaque mois des indices qui représentent l'évolution des coûts pour les différentes activités de la construction : les index bâtiment (BT). Ils sont appréhendés à l'aide de six postes de coût : le coût du travail, les matériaux, le matériel, l'énergie, le transport et les frais divers. Il existe un indice pour une multitude d'activités (voire de

sous-activités). On peut par exemple citer le BT03 pour la maçonnerie, le BT47 pour l'électricité et le BT53 pour l'étanchéité. Cependant, le type d'activité sinistrée n'est pas disponible au sein des bases de données dont nous disposons. Par conséquent, c'est l'indice BT01 qui semble le plus approprié. Ce dernier correspond à l'ensemble des activités du bâtiment tous corps d'état confondus, et permet de donner une vision globale de l'évolution des prix dans le secteur de la construction.

Même si cet indice est publié mensuellement, c'est une actualisation annuelle qui est effectuée puisque l'historique des indemnisations manque de précision pour procéder à une actualisation mensuelle. L'indice considéré pour chaque année écoulée est donc le BT01 du mois de décembre, et le dernier connu pour l'année en cours. Une méthode de projection de cet indice a été mise en place par Martin (2019) et sera utilisée pour pouvoir prendre en compte l'inflation des sinistres qui surviendront dans les prochaines années.

4.1.2 Constitution des bases

La constitution des bases nécessaires à la construction du modèle suit les mêmes étapes que celles décrites dans la partie 3.3 du mémoire de Martin (2019). Celles-ci sont mêmes simplifiées étant donné que les contrats de la garantie Dommages Ouvrage ne sont liés qu'à une seule DOC.

Tout d'abord, une base relative aux sinistres est créée. Une fonction d'actualisation s'appuyant sur l'indice BT01 permet de mettre en *as-if* les montants de règlement et de recours en fonction de leur année de versement. Par exemple, pour exprimer un montant réglé en 2014 en euros d'aujourd'hui, on écrit :

$$\text{MontantRéglé}_{2014\rightarrow \text{ Aujourd'hui}} \;=\; \text{MontantRéglé}_{2014} \times \frac{BT01_{\text{Courant}}}{BT01_{2014}}$$

où :

- $\text{MontantRéglé}_{2014\rightarrow \text{ Aujourd'hui}}$ est le montant réglé en 2014 en euros d'aujourd'hui ;
- $\text{MontantRéglé}_{2014}$ est le montant réglé en 2014 en euros de 2014 ;
- $BT01_{\text{Courant}}$ est la dernière valeur de l'indice $BT01$ publiée pour l'année en cours ;
- $BT01_{2014}$ est la valeur de l'indice $BT01$ en décembre 2014.

En plus de l'actualisation des montants, une variable indiquant si un sinistre a été clos à 0 est calculée. La base est ensuite transformée d'une vision comptable à une vision connaissance sinistre. Autrement dit, on connait pour chaque sinistre son positionnement par rapport à la DOC et son développement dans le temps à partir de l'instant où il est connu. Cela permet *in fine* d'obtenir des montants en euros constants. Ces différents montants actualisés permettront de mettre en place le modèle de coût.

Cette base de sinistres est alors jointe à la base répertoriant l'ensemble des contrats, dans le but de construire le modèle de fréquence. Pour chaque contrat, il est calculé le nombre de sinistres observés ainsi que la durée en jours entre la date de début de la garantie et la date de connaissance pour chaque sinistre.

Enfin, une base d'apprentissage consacrée aux coefficients de liquidation des sinistres est considérée. Cette base, contrairement aux deux précédentes qui sont utilisées pour la modélisation des sinistres inconnus, servira à la projection des sinistres connus mais non clos, et sera personnalisée pour chaque sinistre. Pour un sinistre non clos donné ouvert depuis n jours, la base d'apprentissage contiendra l'ensemble des sinistres clos ayant été ouverts pendant au moins n jours. Les variables récupérées sont le coût ultime des sinistres, le coût prévisionnel brut de provision de recours à l'instant le plus proche de la durée n, la proportion de PSAP et la proportion de réglé diminué des recours dans ce coût.

4.2 Description des données utilisées

Le modèle présenté dans ce mémoire cherche à expliquer plusieurs variables à l'aide des caractéristiques des contrats. Il est alors important, avant toute modélisation, de bien comprendre ce que représentent les variables d'intérêt et les variables explicatives.

Les variables à expliquer de ce modèle sont :

- La fréquence de sinistralité d'un contrat, qui correspond au nombre de sinistres qu'un contrat va engendrer du début à la fin de sa période de garantie ;
- Le montant de règlement ultime d'un sinistre, qui correspond à l'indemnisation versée à l'assuré sans prendre en compte les recours exercés ;
- La présence ou non de recours et leur montant ultime exercé le cas échéant ;
- La « durée de vie » d'un sinistre, qui correspond à la durée en jours entre l'instant de connaissance du sinistre et l'instant de sa clôture.

Pour expliquer ces variables, les modèles proposés peuvent entre autres s'appuyer sur des caractéristiques de souscription du contrat. Les variables explicatives inhérentes au contrat envisagées sont :

- La région du chantier ;
- Le coût du chantier (variable quantitative) ;
- Le type de construction, qui correspond à l'usage que va avoir l'ouvrage après la construction (bâtiment agricole, pavillon individuel, bureaux, etc.) ;
- La qualité du souscripteur, qui correspond au statut de la personne qui souscrit la garantie Dommages Ouvrage. Ce peut être par exemple le propriétaire de l'ouvrage, un promoteur immobilier, l'État, etc. ;
- La présence ou non d'un accord cadre pour la souscription du contrat. Dans notre contexte, un accord cadre est un accord conclu entre l'assureur et un acteur de la construction qui permet d'établir les règles relatives à la souscription des contrats pendant une période donnée. Ces accords fixent les quantités de contrats envisagés et leurs tarifs, généralement avantageux pour le souscripteur.

4.3 Analyse préliminaire des dépendances entre variables explicatives et variables à expliquer

Une étape d'analyse préliminaire des données permet de connaitre les variables qui pourront être intégrées dans les différents modèles. En effet, pour qu'une variable puisse en expliquer une autre, une dépendance entre les valeurs observées de celles-ci doit être constatée. Pour chaque variable d'intérêt que l'on souhaite modéliser en intégrant ces variables inhérentes au contrat, des coefficients représentant l'intensité du lien avec chaque variable à expliquer sont calculés.

- Si la variable d'intérêt est quantitative et la variable explicative est qualitative, on calcule le rapport de corrélation, résultat de l'analyse de variance (ANOVA), et décrit en annexe ;
- Si la variable d'intérêt et la variable explicative sont quantitatives, on calcule le coefficient de corrélation linéaire, décrit dans le chapitre 7 de ce mémoire ;
- Si la variable d'intérêt et la variable explicative sont qualitatives, on calcule le V de Cramer, décrit en annexe.

Ces indicateurs de corrélation entre les covariables et la fréquence de sinistralité d'un contrat sont :

	Coefficient de corrélation linéaire (Cor) ou rapport de corrélation (RCor)
Coût de chantier	0,31 (Cor)
Type de construction	0,113 (Rcor)
Région	0,009 (Rcor)
Qualité du souscripteur	0,104 (Rcor)
Accord cadre	$< 10^{-5}$ (Rcor)

TABLE 4.1 – Dépendance des variables explicatives avec la fréquence de sinistralité

Ces derniers sont significatifs excepté celui pour la variable relative à la présence d'accord cadre. Les quatre premières variables pourront alors être intégrées dans le modèle de projection de la fréquence.

Les mêmes calculs, effectués cette fois avec la variables des montants de règlement, donnent les résultats suivants :

	Coefficient de corrélation linéaire (Cor) ou rapport de corrélation (RCor)
Coût de chantier	0,027 (Cor)
Type de construction	0,020 (Rcor)
Région	0,003 (Rcor)
Qualité du souscripteur	0,014 (Rcor)
Accord cadre	0,005 (Rcor)

TABLE 4.2 – Dépendance des variables explicatives avec les montants de règlement des sinistres

Les indicateurs, bien que de valeur nettement plus faible que pour la fréquence, indiquent une corrélation significative pour chaque variable. Cela signifie que la fréquence est beaucoup plus sensible aux caractéristiques du contrat que le coût des sinistres. Elles pourront tout de même être intégrées dans le modèle de coût, même si la segmentation sera moins significative.

4.4 Impact de l'antériorité de sinistralité

Outre l'ensemble des attributs du contrat disponibles au moment de la souscription, le modèle s'appuie sur la sinistralité observée depuis le début de période de garantie pour projeter la sinistralité future. En effet, l'observation de sinistres augmente la probabilité d'en observer d'autres d'ici la fin de la période décennale. On peut illustrer cela en représentant le taux de contrats sinistrés l'année $DOC + i$ $(i = 1, ..., 13)$ conditionné à l'observation d'au moins un sinistre l'année précédente $(DOC + i - 1)$ contre ce même taux conditionnellement à l'observation d'aucun sinistre.

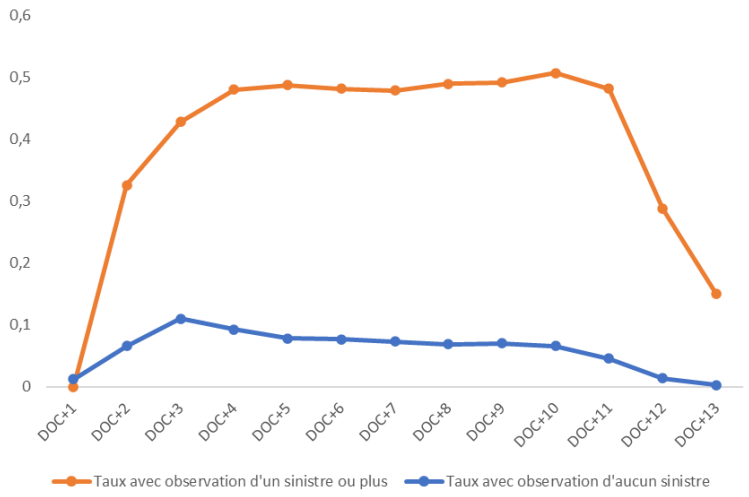

FIGURE 4.1 – Taux de présence de sinistralité conditionnellement à la sinistralité de l'année passée

La significativité de la courbe construite est plus faible pour les premières et dernières années car la sinistralité est moindre. Cependant, pour l'ensemble des années relative à la DOC, le constat est clair : le taux de contrats sinistrés est bien plus important lorsqu'au moins un sinistre a été observé l'année précédente.

L'ensemble de ces travaux élémentaires permettra par la suite d'aborder sereinement l'étape de modélisation. Cette étape de modélisation s'appuiera sur des travaux effectués pour la projection de la garantie RCDO, qui fait l'objet du chapitre suivant.

Chapitre 5

Méthodologie en place

Longtemps utilisée dans l'entreprise grâce à sa simplicité de mise en place et de compréhension, la méthode de projection Chain-Ladder a montré des limites de précision pour les garanties décennales. En effet, cette méthode se montre peu robuste lorsque la part de la sinistralité ultime connue est faible, ce qui est le cas des garanties décennales puisque les sinistres peuvent s'étaler sur une dizaine d'années. Des travaux ont récemment été menés afin d'élaborer un modèle de projection plus adapté à ce type de garanties. Ces derniers ont fait l'objet d'un mémoire présenté devant l'Institut des Actuaires (Martin, 2019). Ce chapitre a pour but de donner des éléments de compréhension au lecteur en synthétisant le procédé de projection précédemment instauré. Il reprend notamment certaines parties ou graphiques utilisés dans le mémoire en question.

5.1 Contexte de mise en place de la méthode

Les récentes faillites de plusieurs acteurs du marché de l'assurance construction montrent que cette dernière est une branche de l'assurance difficile à maitriser. En effet, la spécificité des garanties décennales rend obscure l'appréhension des risques, d'autant plus que le portefeuille assuré est en constante évolution. Dans ce contexte, les méthodes usuelles de l'assurance non-vie atteignent rapidement leurs limites. Il a alors été nécessaire de produire un modèle plus adapté (mais aussi plus complexe) pour estimer plus justement la sinistralité future du portefeuille. Cette étude a initialement été mise en place pour la garantie Responsabilité Civile Décennale Obligatoire (RCDO), puis a été transposée à la garantie Dommages Ouvrage (DO).

Pour prendre en compte la déformation du portefeuille, le modèle projette la sinistralité contrat par contrat, que ce soit pour les sinistres connus non clos ou pour les sinistres non encore manifestés. Pour son apprentissage, il s'appuie sur l'information des contrats de toutes les DOC (et non pas seulement des DOC « éteintes »). Ainsi, cela lui permet d'être dynamique et de pouvoir être réévalué dès que de nouvelles données sont disponibles. Ce modèle mélange des approches paramétriques et non paramétriques en prenant en compte les informations obtenues à la souscription du contrat mais aussi l'antériorité de la sinistralité inhérente à celui-ci.

5.2 Modélisation de la fréquence des sinistres inconnus

La méthode de modélisation de la fréquence fait tout d'abord appel à des modèles non paramétriques (arbres CART) pour classer les risques grâce à certaines variables exogènes. L'approche consiste en la création de groupes de risques homogènes en matière de fréquence de sinistralité. Cette dernière peut être divisée en deux indicateurs, qui sont la probabilité de survenance d'au moins un sinistre et le nombre de sinistres le cas échéant. Cela permet de ne pas perturber la distribution du nombre de sinistres par une trop grande représentation du nombre de zéros. On parle de modèle à inflation de zéro. Elle prend ensuite en compte dans chaque classe de manière paramétrique l'impact de la sinistralité passée du contrat.

La démarche s'effectue indépendamment pour chaque année de connaissance relative à la DOC. Ce choix est justifié par la variabilité de la sinistralité au cours de la vie du contrat. Pour illustrer ce propos, si l'on considère qu'un contrat peut engendrer des sinistres pendant n années après la DOC, alors $n + 1$ modèles indépendants (il y a aussi un modèle pour l'année de la DOC) seront ajustés.

5.2.1 Segmentation de la probabilité de survenance d'au moins un sinistre

Pour commencer, le procédé passe par la création de groupes de risques homogènes quant à la probabilité de survenance d'au moins un sinistre au cours de l'année modélisée.

Ces groupes sont construits à partir de variables exogènes aux contrats telles que le chiffre d'affaires de l'entreprise, son activité principale, l'ancienneté de l'entreprise, son effectif, etc. Il est important de remarquer que toutes les covariables utilisées sont invariables à tout instant du déroulement du contrat relativement à la DOC. Cela permet de fixer a priori les groupes d'appartenance de chaque contrat pour chaque année de vie de celui-ci.

La segmentation est alors réalisée en utilisant des arbres CART. Le fonctionnement des arbres CART peut être consulté en annexe. Ces arbres modélisent la probabilité qu'au moins un sinistre survienne selon certaines caractéristiques du contrat ainsi que son année relativement à la DOC. L'échantillon d'apprentissage évolue en fonction de l'année du contrat considérée par rapport à la DOC. En effet, On ne peut pas considérer les contrats aux DOC récentes pour apprendre de la sinistralité d'un contrat dix ans après la DOC. Il serait par exemple absurde d'apprendre de la DOC 2016 lorsque l'on s'intéresse à l'année D+7 puisque l'on ne connaîtra la sinistralité correspondante qu'en 2023 (ce qui explique la différence d'effectif d'apprentissage entre l'année D+0 et D+8 sur l'illustration).

La construction de ces arbres CART modélisant la survenance d'au moins un sinistre est la première partie de la modélisation de la fréquence. Elle est alors complétée par la modélisation du nombre de sinistres conditionné à la survenance d'au moins un sinistre.

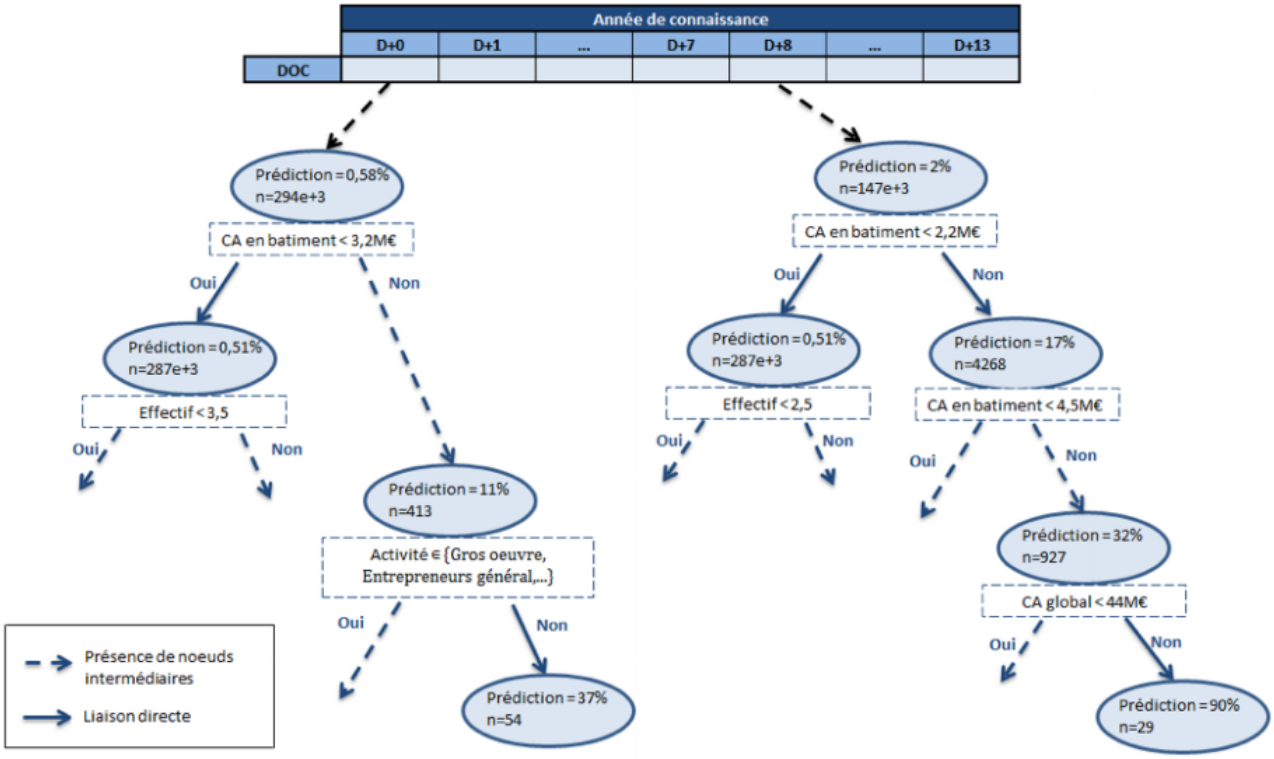

FIGURE 5.1 – Segmentation des contrats par rapport à la probabilité de survenance d'au moins un sinistre par arbre CART

5.2.2 Segmentation du nombre de sinistres conditionnellement au fait qu'au moins un sinistre soit survenu

La méthode employée est exactement la même que pour la partie précédente, à l'exception près que la variable à expliquer devient le nombre de sinistres conditionnellement au fait qu'au moins un sinistre soit survenu.

On peut remarquer que ces arbres sont bien moins développés que ceux expliquant la survenance d'au moins un sinistre (l'arbre construit pour l'année D+8 se compose même seulement de sa racine). Cela traduit le fait que les variables explicatives segmentent peu le nombre de sinistres, mais aussi que le volume de données est moins important (seuls les contrats avec au moins un sinistre sont considérés).

Comme indiqué dans la partie précédente, les groupes sont formés avec des variables invariantes au cours de la vie du contrat. La segmentation ne prend donc pas en compte la sinistralité observée les années précédentes. Or, on peut facilement se convaincre du fait que la sinistralité observée peut permettre d'expliquer plus précisément la sinistralité future, puisque pour un même contrat RCDO, ce sont sensiblement les mêmes employés qui sont intervenus et les mêmes matériaux qui ont été utilisés sur les différents chantiers. La même remarque peut être effectuée sur les contrats de la garantie DO, au détail près qu'un contrat est lié à un seul chantier. L'antériorité de sinistralité

sera donc prise en compte seulement sur ce même chantier.

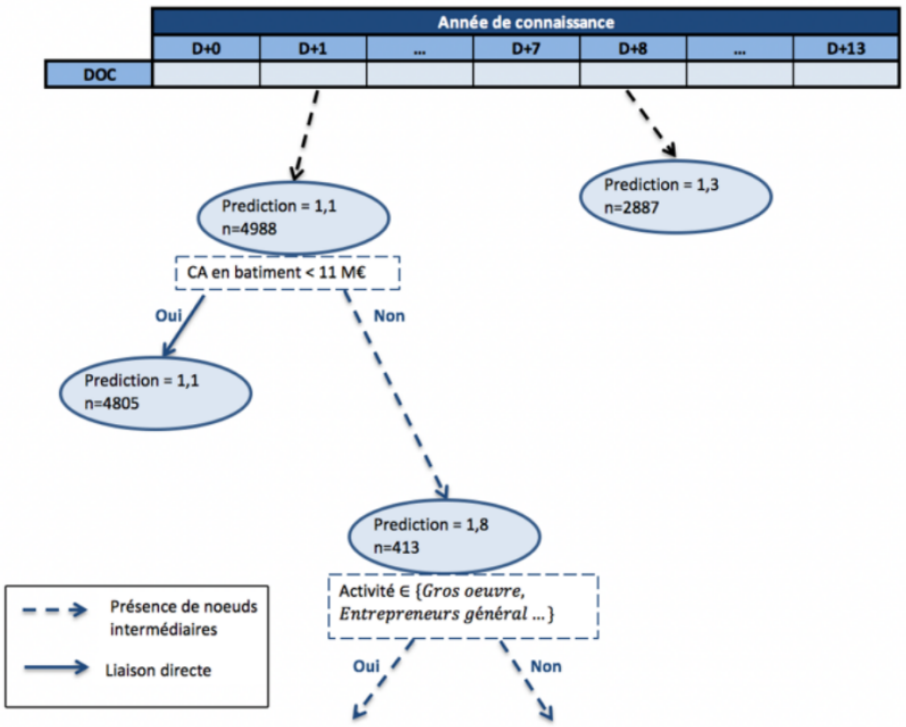

FIGURE 5.2 – Segmentation des contrats par rapport au nombre de sinistres conditionnellement au fait qu'au moins un sinistre soit survenu par arbre CART

5.2.3 Intégration de l'antériorité de sinistralité

Une fois les groupes de risques homogènes créés (selon les deux variables à expliquer précédemment évoquées), seule l'antériorité de sinistralité pourrait permettre de segmenter plus précisément le risque au sein d'un même groupe. Par souci de simplification, seule la méthode adaptée à la garantie DO sera décrite.

La première étape de cette partie consiste à identifier les variables relatives à l'antériorité de sinistralité segmentant le risque. Cette analyse est réalisée de manière non paramétrique pour expliquer la sinistralité de l'année $D+i \quad (i=1,...,m)$, en utilisant des arbres CART avec comme variables explicatives :

- Le nombre de sinistres observés pendant l'année $D+j \quad (j=0,...,i-1)$;
- Le nombre de sinistres cumulés l'année $D+j \quad (j=0,...,i-1)$.

Cette démarche est alors répétée pour chaque année relative à la DOC et pour chaque groupe de risques homogène constitué, que ce soit pour expliquer la survenance d'au moins un sinistre ou le nombre de sinistres le cas échéant.

5.2.3.1 Intégration de l'antériorité de sinistralité pour le nombre de sinistres conditionnellement au fait qu'au moins un sinistre soit survenu

Lorsqu'il est possible de segmenter via l'antériorité de sinistralité, le modèle mis en place s'inspire des Modèles Linéaires Généralisés (MLG). La méthode décrite ci-dessous va s'appliquer pour chaque groupe de risques homogène créé précédemment. Le modèle s'inspire d'un modèle log-Poisson.

Pour rappel, dans un modèle log-Poisson classique, en notant Y la variable d'intérêt et X la matrice des covariables, on modélise :

$$\ln(\mathbb{E}[Y \mid X]) = \alpha + \beta^T X \Longrightarrow \lambda \triangleq \mathbb{E}[Y \mid X] = \exp\left(\alpha + \beta^T X\right)$$

Dans notre cas, la modélisation considérée est celle d'une loi de Poisson tronquée en 0, étant donné le conditionnement exigé. L'estimation des paramètres α et β s'effectue par maximisation de la fonction de vraisemblance. Autrement dit, cela revient à résoudre :

$$(\widehat{\alpha}, \widehat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmax}} \prod_{i=1}^{m} \mathbb{P}\left(N = n_i \mid N > 0; x_i'; \alpha; \beta\right)$$

En écartant les étapes calculatoires, la fonction de log-vraisemblance à maximiser s'écrit :

$$l\left(\alpha, \beta \mid N, X'\right) = \sum_{i=1}^{m} \left[n_i \left(\alpha + \beta^T x_i'\right) - e^{\alpha + \beta^T x_i'} - \ln\left(n_i!\right) - \ln\left(1 - e^{-\exp\left(\alpha + \beta^T x_i'\right)}\right)\right]$$

Il suffit de résoudre :

$$\begin{cases} \frac{\partial l(\alpha,\beta|N,X')}{\partial \alpha} = 0 \\ \frac{\partial l(\alpha,\beta|N,X')}{\partial \beta} = 0 \end{cases}$$

La solution est calculée numériquement. Cela permet alors, connaissant l'antériorité de sinistralité du contrat pour une année donnée, de déterminer le paramètre de la loi de Poisson tronquée ajustée au nombre de sinistres pour cette même année.

Pour les groupes de risques homogènes dans lesquels il est impossible de segmenter via l'antériorité de sinistralité, un simple ajustement du nombre de sinistres à une loi de comptage usuelle est effectué. Le modèle estime par maximum de vraisemblance les paramètres optimaux pour les lois de Poisson, binomiale et binomiale négative (tronquées en zéro). La meilleure loi est ensuite choisie grâce au critère de minimisation de la somme des écarts quadratiques entre les fonctions de répartition empirique et théorique.

5.2.3.2 Intégration de l'antériorité de sinistralité pour la probabilité qu'au moins un sinistre survienne

L'intégration de l'antériorité de sinistralité pour estimer la probabilité de survenance d'au moins un sinistre fonctionne de la même manière qu'expliquée dans la partie précédente. Lorsque la segmentation via l'antériorité de sinistralité est possible, le modèle s'inspire cette fois-ci de la régression logistique puisque le problème est binaire (« Survenance » ou « Non survenance »).

Pour rappel, en notant $Y \in \{0,1\}$ la variable d'intérêt et X la matrice des covariables, la régression logistique utilisant la fonction de lien Logit modélise :

$$\ln\left(\frac{\mathbb{P}(Y=1 \mid X)}{1-\mathbb{P}(Y=1 \mid X)}\right) = \alpha + \beta^T X \Longrightarrow \mathbb{P}(Y=1 \mid X) = \frac{e^{\alpha+\beta^T X}}{1+e^{\alpha+\beta^T X}}$$

Dans notre cas, on modélise P la variable aléatoire qui vaut 1 si au moins un sinistre survient et 0 sinon. L'estimation des paramètres α et β s'effectue par maximisation de la fonction de vraisemblance. Autrement dit, cela revient à résoudre :

$$(\widehat{\alpha},\widehat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmax}} \prod_{i=1}^{m} \mathbb{P}\left(P=1 \mid x';\alpha;\beta\right)^{\max(0;\min(n_i;1))} \mathbb{P}\left(P=0 \mid x';\alpha;\beta\right)^{1-\max(0;\min(n_i;1))}$$

On écarte de nouveau les étapes calculatoires pour obtenir la log-vraisemblance à maximiser :

$$l\left(\alpha,\beta \mid N,X'\right) = \sum_{i=1}^{m} \left[\max\left(0;\min\left(n_i;1\right)\right)\left(\alpha+\beta^T x_i'\right) - \left(1+e^{\alpha+\beta^T x_i'}\right)\right]$$

Il suffit de résoudre :

$$\begin{cases} \frac{\partial l(\alpha,\beta|N,X')}{\partial \alpha} = 0 \\ \frac{\partial l(\alpha,\beta|N,X')}{\partial \beta} = 0 \end{cases}$$

La solution est également calculée numériquement. Cela permet de calculer la probabilité de survenance d'au moins un sinistre d'un contrat pour une année donnée en prenant en compte l'antériorité de sinistralité.

Dans les groupes de risques homogènes où il est impossible de segmenter via l'antériorité de sinistralité, la modélisation est effectuée en considérant une loi de Bernoulli ayant comme paramètre la moyenne empirique du groupe en question.

5.3 Modélisation du coût des sinistres et de la corrélation avec leur durée de vie

L'intérêt de l'étude de la corrélation entre ces deux grandeurs se trouve dans le fait que l'inflation (relativement élevée) tient un rôle important dans l'analyse du coût des sinistres. Il est alors intéressant de prendre en compte la dimension temporelle des sinistres pour mieux appréhender leur coût. Il est important de notifier que dans le modèle construit, la variable d'intérêt représente la charge ultime nette de recours du sinistre, qui correspond au montant payé par l'assureur diminué des recours encaissés par ce dernier une fois le sinistre clos. Cette agrégation de deux variables en une seule est l'un des points sur lequel des modifications de la méthode seront apportées et expliquées dans ce mémoire.

Le modèle s'appuie sur l'ajustement des distributions de chaque variable et l'étude de la corrélation par copules. Des démarches similaires seront effectuées dans le septième chapitre, et remettront totalement en cause cette partie. Le nouveau modèle reprend à l'identique la méthode d'ajustement de lois, et utilise également des copules. Ces éléments seront donc uniquement détaillés dans le chapitre dédié du présent mémoire, afin d'éviter une certaine redondance.

5.4 Simulations des sinistres inconnus

La méthode de projection proposée utilise la simulation d'un grand nombre de scénarios. L'intégralité de la démarche de simulation découlant des modélisations réalisées dans ce chapitre peut alors être représentée par les schémas suivants, directement tirés du mémoire de Martin (2019) :

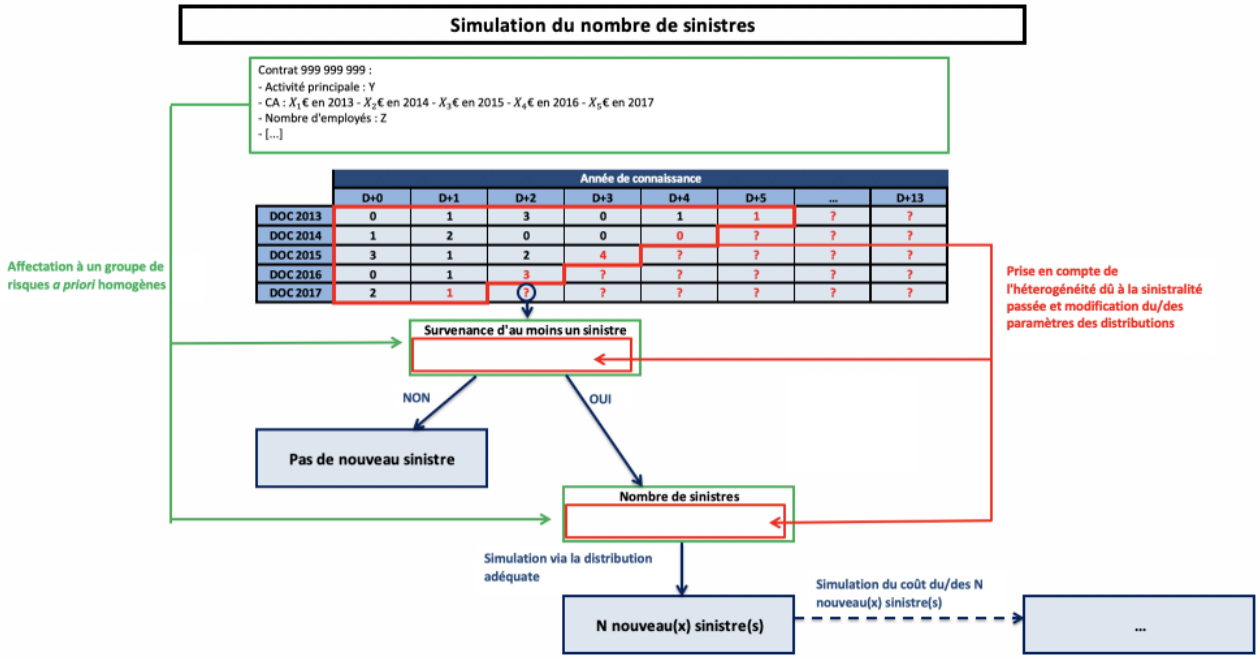

FIGURE 5.3 – Schéma récapitulatif de la démarche de simulation relative à l'estimation de la fréquence pour la méthode en place

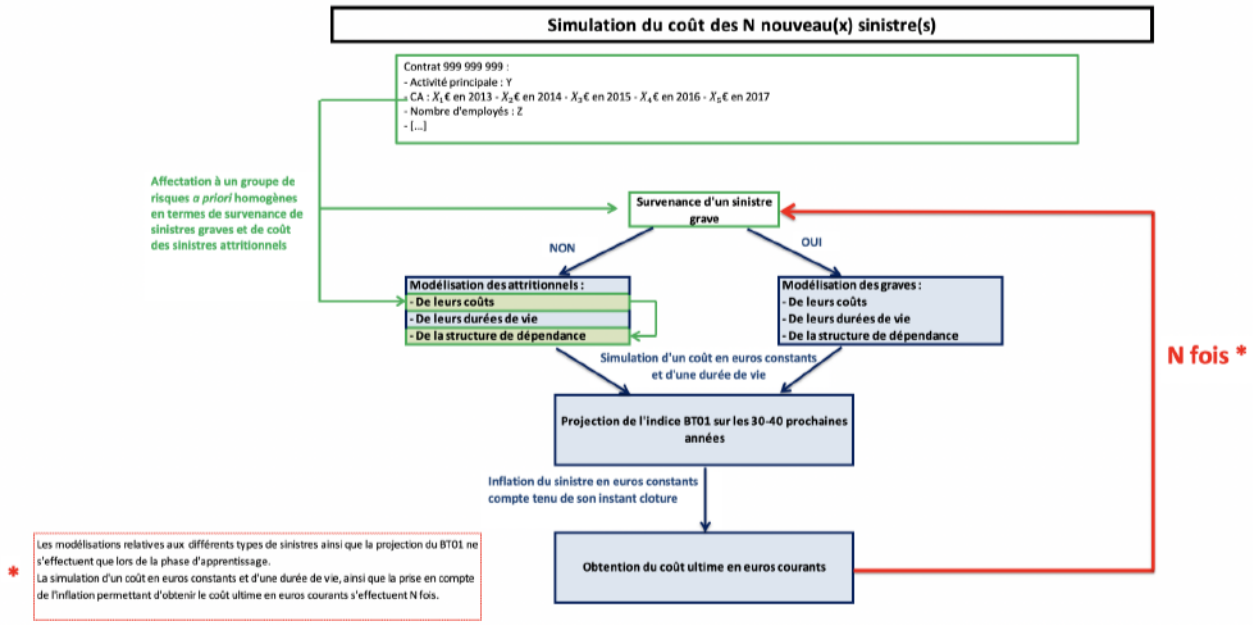

FIGURE 5.4 – Schéma récapitulatif de la démarche de simulation relative à l'estimation du coût et de la durée de vie des sinistres pour la méthode en place

5.5 Projection des sinistres connus et non clos

La projection de la sinistralité inconnue constitue une partie importante de la projection du portefeuille pour la garantie construction. Néanmoins, la capacité à estimer les trajectoires que vont emprunter les sinistres connus au moment de la projection tient également une place non négligeable. L'estimation du coût d'un sinistre réalisée à dire d'expert par les gestionnaires évolue au cours du temps en fonction des nouvelles informations sur le sinistre. L'intérêt se porte donc sur la prédiction du coût ultime du sinistre. Pour cela, le modèle en place procède en deux temps.

5.5.1 Estimation de la probabilité qu'un sinistre se clôture à zéro

Premièrement, la probabilité que le sinistre se clôture à zéro est estimée à l'aide d'un arbre CART. Ce dernier est construit grâce à une base d'apprentissage dépendant de la durée de vie actuelle du sinistre. A titre d'exemple, si un sinistre est ouvert depuis 100 jours, la base d'apprentissage pour ce sinistre contient l'ensemble des sinistres clos étant restés ouverts au moins 100 jours. L'arbre construit s'appuie alors sur les données suivantes pour estimer la probabilité qu'un sinistre se clôture à zéro :

- Le coût ultime du sinistre à sa clôture ;
- Le coût prévisionnel brut de provision de recours à l'instant le plus proche de la durée d'ouverture considérée ;
- La proportion de PSAP dans la variable précédente ;
- La proportion de réglé diminué des recours reçus dans cette même variable.

Grâce à ces caractéristiques, il est alors possible d'attribuer à chaque sinistre à projeter un groupe d'appartenance et donc une probabilité de se clôturer à zéro. Cette probabilité sera alors utilisée lors des simulations comme le paramètre d'une loi de Bernoulli.

5.5.2 Estimation du coefficient de liquidation conditionnellement au fait que le sinistre ne se clôture pas à zéro

Dans un second temps, c'est le coefficient de liquidation (conditionnellement au fait que le sinistre ne se clôture pas à zéro) qui est estimé. De la même manière que pour estimer la probabilité que le sinistre se clôture à zéro, un arbre CART prédisant le coefficient de liquidation est construit à l'aide des covariables citées supra pour chacun des sinistres. La base d'apprentissage est restreinte aux sinistres ne se clôturant pas à zéro, afin de prendre en compte le conditionnement. Une fois ces groupes de sinistres homogènes constitués, une distribution usuelle est ajustée pour prendre en compte l'hétérogénéité à l'intérieur des groupes lors de la phase de simulation. A l'instar de l'ajustement des lois pour la projection des sinistres inconnus, plusieurs lois usuelles sont modélisées pour la variable du coefficient de liquidation. Le choix de la meilleure se fait alors selon le critère de minimisation de la somme des écarts quadratiques entre les fonctions de répartition empirique et théorique.

5.6 Limites de la méthode pour la garantie DO

La méthodologie en place satisfait les critères nécessaires pour obtenir une bonne projection pour les garanties construction, à savoir le traitement du développement des sinistres connus mais aussi la prédiction des sinistres futurs. Cependant, même si ce modèle se montre efficace pour projeter la garantie RCDO, des pistes d'amélioration sur celui-ci peuvent être évoquées, ainsi que des solutions alternatives pour mieux s'adapter à la projection de la garantie DO.

Dans un premier temps, on peut s'apercevoir que le modèle dans son ensemble s'appuie sur une segmentation très importante des risques, en particulier pour le modèle de fréquence. Cette forte segmentation au niveau des contrats engendre, avec la prise en compte paramétrique de la sinistralité antérieure, un modèle avec un très grand nombre de paramètres. En effet, une loi est ajustée pour chaque groupe de risques homogène créé à partir d'un arbre CART pour chaque année de développement du contrat, et ce pour les deux variables à expliquer (probabilité de survenance d'au moins un sinistre et nombre de sinistres le cas échéant). Si ce grand nombre de paramètres permet un bon ajustement aux données d'apprentissage, il questionne la qualité du modèle en matière de prédiction (application sur de nouvelles données).

Dans un second temps, nous avons vu que le modèle de coût des sinistres cherche à expliquer la charge ultime de l'assureur, à savoir le montant réglé au moment de la clôture du sinistre diminué du montant de recours encaissé. Dans le cas de la garantie RCDO, cela fonctionne plutôt bien puisque les recours ne prennent pas une place très importante du fait de la nature de ce produit d'assurance, et ont donc peu d'impact sur la charge finale de l'assureur. En revanche, les recours se révèlent primordiaux dans le fonctionnement de la garantie DO, et influencent de manière non négligeable la charge ultime de l'assureur. Une perte d'information sur le comportement de la sinistralité est alors constatée lorsque l'étude s'effectue sur une variable qui est en réalité la réunion de deux variables distinctes. La démarche relative à la modélisation des coûts peut alors faire l'objet d'une tentative d'amélioration, en intégrant trois variables dans la structure de dépendance au lieu de deux actuellement.

Chapitre 6

Modélisation de la fréquence par processus de Hawkes

Modéliser la fréquence des sinistres constitue une étape primordiale dans la projection de la sinistralité pour la garantie Dommages Ouvrage. En effet, pour les DOC récentes, l'incertitude liée au résultat correspondant à cette garantie dépend essentiellement des sinistres non encore déclarés/survenus. La période de garantie étant très longue, il peut être complexe d'estimer la sinistralité que va observer chaque contrat. Le modèle actuel dispose d'une méthode à priori performante dans la prédiction de la fréquence de sinistralité future. Cependant, ce modèle est complexe et la segmentation est très importante, ce qui peut faire émerger un risque de sur-apprentissage. L'objectif est alors de proposer un modèle alternatif faisant intervenir moins de paramètres pour plus de robustesse. Une segmentation des contrats selon leurs caractéristiques intrinsèques reste tout de même primordiale, de même que la prise en compte de la sinistralité connue. A cet effet, on introduit un type particulier de processus temporels : les processus de Hawkes.

6.1 Présentation des processus de Hawkes

Avant d'introduire concrètement la définition et les propriétés d'un processus de Hawkes, il est nécessaire de rappeler progressivement les types de processus qui le composent.

6.1.1 Rappels sur les processus temporels

Définition 6.1.1. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé complet. Soit $(t_i)_{i\in\mathbb{N}^*}$ une suite strictement croissante de variables aléatoires positives ou nulles. $(t_i)_{i\in\mathbb{N}^*}$ est alors appelé un* ***processus ponctuel*** *sur $\mathbb{R}^+$.*

Dans notre situation, chaque variable aléatoire t_i représente l'instant de survenance d'un sinistre pour un même contrat. On peut représenter un processus de ce type avec le schéma suivant :

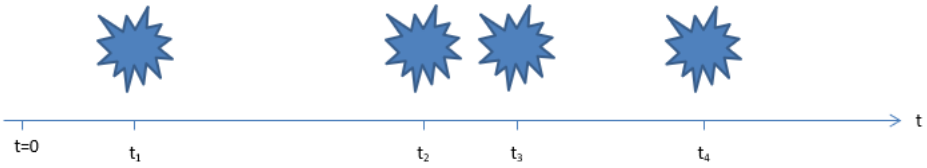

FIGURE 6.1 – Représentation d'un processus ponctuel

Définition 6.1.2. *Soit $(t_i)_{i\in\mathbb{N}^*}$ un processus ponctuel sur $\mathbb{R}^+$. On appelle* ***processus de comptage*** *le processus N continu à droite tel que :*

$$N(t) = \sum_{i\in\mathbb{N}^*} \mathbb{1}_{[t_i,+\infty[}(t) \quad t \geq 0$$

.

Le processus de comptage considéré sera le nombre total de sinistres observés à un instant t depuis la date d'entrée en garantie du contrat. Ce type de processus est alors étroitement lié à un processus ponctuel. Ce lien peut être représenté graphiquement en parallèle avec le schéma du processus ponctuel.

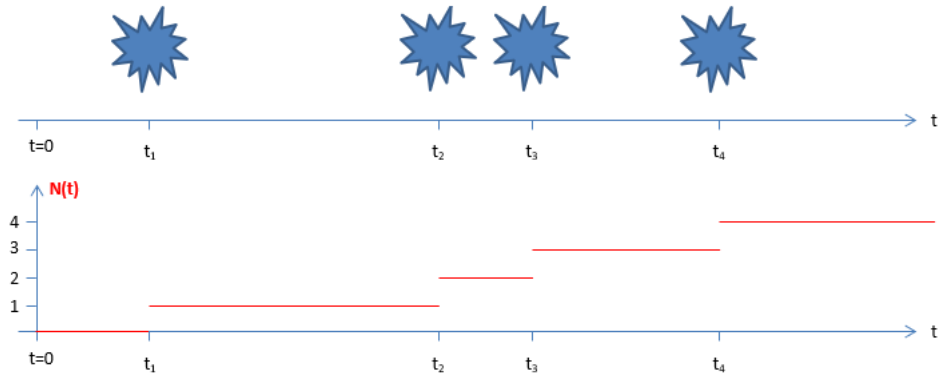

FIGURE 6.2 – Représentation d'un processus de comptage

Définition 6.1.3. *Soit N un processus de comptage adapté à la filtration $(\mathcal{F}_t)_{t\geq 0}$, le* ***processus d'intensité*** *est défini par sa fonction d'intensité λ telle que :*

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0} \mathbb{E}\left[\frac{N(t+h)-N(t)}{h} \mid \mathcal{F}_t\right] \\ &= \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left[N(t+h)-N(t) > 0 \mid \mathcal{F}_t\right], t \geq 0.\end{aligned}$$

L'intensité $\lambda(t)$ du processus représente la probabilité qu'il y ait un saut entre le temps t et t+h lorsque h tend vers 0, connaissant le nombre d'évènements déjà survenus avant t. Autrement dit, cela représente le risque qu'un évènement survienne au temps t.

Remarque : L'intensité n'est pas une probabilité, elle peut prendre des valeurs supérieures à 1.

Exemple :

Le processus de Poisson, fréquemment utilisé dans les modèles de fréquence, permet d'illustrer simplement les rappels énoncés supra.

Définition 6.1.4. *Un processus de comptage* $\{N(t), t \geq 0\}$ *est un processus de Poisson d'intensité* $\lambda \geq 0$ *si :*

- $N(0) = 0$ *;*
- *Le processus est à accroissements indépendants ;*
- $\forall 0 \leq s < t$, *la variable aléatoire* $N(t) - N(s)$ *suit une loi de Poisson de paramètre* $\lambda(t-s)$.

Sa simplicité rend aisée son interprétation ainsi que sa mise en place. Cependant, ce type de processus manque de flexibilité. En effet, il est d'intensité constante et est totalement déterminé par un unique paramètre. De plus, ses accroissements sont stationnaires et indépendants, ce qui signifie que le processus n'est pas influencé par les évènements passés. Pour remédier à ces faiblesses, on introduit un type de processus plus élaboré : les processus linéaires auto-excitants.

6.1.2 Les processus linéaires auto-excitants

Définition 6.1.5. *Un* ***processus linéaire auto-excitant*** N *est un processus de comptage de fonction d'intensité* λ *de la forme :*

$$\begin{aligned}\lambda(t) &= \lambda_0(t) + \int_{-\infty}^{t} \nu(t-s) dN_s \qquad où \begin{cases} \lambda_0 : \mathbb{R}^+ \longrightarrow \mathbb{R}^+ \\ \nu : \mathbb{R}^+ \longrightarrow \mathbb{R}^+ \end{cases} \\ &= \lambda_0(t) + \sum_{t_i < t} \nu\left(t - t_i\right)\end{aligned}$$

λ_0 est une fonction déterministe qui représente l'intensité de base.
ν est une fonction qui représente l'influence positive des évènements passés $t_i \leq t$ sur la valeur de l'intensité en t.

6.1.3 Les processus de Hawkes

On se limitera dans notre étude aux processus de Hawkes unidimensionnels, définis comme suit :

Définition 6.1.6. *Un* ***processus de Hawkes*** *(1971) est un processus linéaire auto-excitant, où la fonction* ν *est définie grâce à la formule suivante :*

$$\nu(t) = \sum_{j=1}^{P} \alpha_j e^{-\beta_j t} \mathbb{1}_{t \in \mathbb{R}_+}$$

Ainsi, la fonction d'intensité devient :

$$\begin{aligned}\lambda(t) &= \lambda_0(t) + \int_0^t \sum_{j=1}^{P} \alpha_j e^{-\beta_j(t-s)} dN_s \\ &= \lambda_0(t) + \sum_{t_i<t} \sum_{j=1}^{P} \alpha_j e^{-\beta_j(t-t_i)}\end{aligned}$$

L'intensité du processus est, contrairement à un processus de Poisson, variable dans le temps. Elle est constituée d'une intensité de base λ_0 (qui peut aussi dépendre du temps) et d'une intensité positive additionnelle représentant l'impact des évènements survenus depuis l'origine du processus.

6.1.4 Les processus de Hawkes simples

Définition 6.1.7. *On appelle* ***processus de Hawkes simple*** *un processus de Hawkes vérifiant $P = 1$ et $\lambda_0(t) = \lambda_0$ constante. La formule de de la fonction d'intensité λ d'un processus de Hawkes simple est donnée par :*

$$\begin{aligned}\lambda(t) &= \lambda_0 + \int_0^t \alpha e^{-\beta(t-s)} dN_s \\ &= \lambda_0 + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}\end{aligned}$$

La condition $P = 1$ est traduite dans notre cas par le fait que seuls les sinistres antérieurs correspondant au contrat étudié ont un impact sur la sinistralité future, puisqu'ils sont relatifs à un seul chantier (et donc une seule DOC). On peut par exemple supposer que ce ne serait pas le cas pour la trajectoire d'un contrat en garantie RDCO, dans laquelle la sinistralité des contrats des DOC précédentes pour une même entreprise permet d'expliquer la sinistralité des contrats de la DOC étudiée, puisque ce sont les mêmes travailleurs et les mêmes matériaux qui ont été utilisés sur les différents chantiers.

La condition λ_0 constante permet la stationnarité du processus en conditionnant les paramètres α et β, et simplifie amplement l'ajustement des paramètres. En effet, dans le cadre de cette étude, l'intensité est supposée stationnaire, c'est à dire que l'on considère $\mathbb{E}[\lambda(t)] = \mu$ avec μ constante. Par suite :

$$\begin{aligned}\mu = \mathbb{E}[\lambda(t)] &= \mathbb{E}\left[\lambda_0 + \int_{-\infty}^{t} \nu(t-s) dN_s\right] \\ &= \lambda_0 + \mathbb{E}\left[\int_{-\infty}^{t} \nu(t-s)\lambda(s) ds\right] \\ &= \lambda_0 + \int_{-\infty}^{t} \nu(t-s)\mu ds \\ &= \lambda_0 + \mu \int_0^{+\infty} \nu(v) dv\end{aligned}$$

On peut alors écrire μ comme :

$$\begin{aligned}\mu &= \frac{\lambda_0}{1 - \int_0^\infty \nu(v)dv} \\ &= \frac{\lambda_0}{1 - \frac{\alpha}{\beta}}\end{aligned}$$

La fonction d'intensité étant une fonction positive, on doit avoir :

$$\begin{aligned}&\mu > 0 \\ \Leftrightarrow &\frac{\lambda_0}{1 - \frac{\alpha}{\beta}} > 0 \\ \Leftrightarrow &1 - \frac{\alpha}{\beta} > 0 \\ \Leftrightarrow &\beta > \alpha\end{aligned}$$

6.2 Ajustement de processus de Hawkes pour la fréquence

Pour rappel, la garantie Dommages Ouvrage est une garantie décennale couvrant le propriétaire d'un ouvrage des malfaçons après la livraison du chantier. L'auto-excitation du processus décrite dans la partie précédente parait cohérente. En effet, la survenance ou non de sinistres donne de l'information sur la qualité des travaux sur le chantier, et donc sur la probabilité de survenance de sinistres sur ce même ouvrage dans le futur.

Pour ajuster un processus de Hawkes à des observations, on doit disposer de trajectoires que l'on suppose provenir du même processus, et donc d'une classe de risques homogène. Or, l'ensemble des contrats du portefeuille ne peut pas être considéré comme étant du même type de risque. La première phase consiste donc en la création de groupes de risques homogènes.

6.2.1 Segmentation des contrats en groupes de fréquence

Pour segmenter les contrats en classes de risques homogènes, on fait de nouveau appel aux arbres CART. Cette classification permet d'expliquer la fréquence de sinistres des contrats de la garantie Dommages Ouvrage, grâce aux covariables :

- Coût du chantier ;
- Type de construction ;
- Qualité du souscripteur ;
- Région du chantier ;
- Présence ou non d'un accord cadre.

Une condition d'effectif minimal de 5000 contrats par feuille est imposée, afin de disposer d'assez d'observations pour obtenir un estimateur robuste des paramètres des processus de Hawkes.

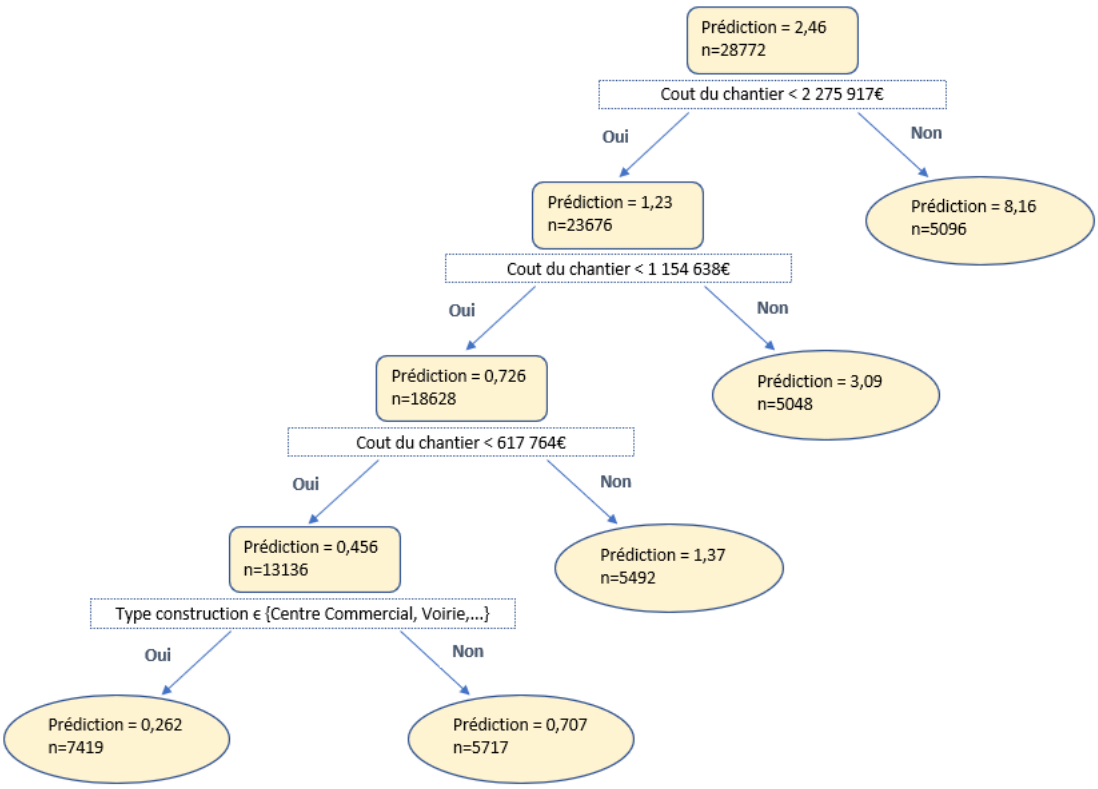

FIGURE 6.3 – Classification des contrats selon la fréquence de sinistralité

6.2.2 Estimation des paramètres des processus de Hawkes

Une fois les classes de risques homogènes construites pour expliquer la fréquence, il reste à ajuster les paramètres λ_0, α et β du processus de Hawkes pour chaque classe. Cette estimation est effectuée par maximum de vraisemblance. Le temps est exprimé en jours, et le processus est supposé avoir une maturité de 13 ans, soit environ 4750 jours. Cette durée est expliquée par le fait que la garantie Dommages Ouvrage débute dès l'ouverture du chantier et pendant dix ans après que ce dernier est terminé (le chantier pouvant durer jusqu'à 3 ans).

La log-vraisemblance d'un processus de comptage N d'intensité λ est définie par :

$$\ln \mathcal{L}\left((N_t)_{t\in[0,T]}\right) = \int_0^T (1-\lambda(s))ds + \int_0^T \ln \lambda(s) dN(s)$$

Dans le cas d'un processus de Hawkes, elle est exprimée par :

$$\ln \mathcal{L}\left(\{t_i\}_{i=1,\dots,n}\right) = t_n - \int_0^{t_n} \lambda(s)ds + \sum_{i=1}^{n} \ln \lambda\left(t_i\right)$$
$$= t_n - \int_0^{t_n} \lambda_0(s)ds - \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\alpha_j}{\beta_j}\left(1 - e^{-\beta_j(t_n - t_i)}\right) + \sum_{i=1}^{n} \ln\left(\lambda\left(t_i\right)\right)$$

Lorsque nos hypothèses sont respectées (processus de Hawkes simple), la log-vraisemblance devient :

$$\ln\left(L\left\{t_i\right\}_{i=1,\dots,n}\right) = t_n - t_n\lambda_0 - \sum_{i=1}^{n} \frac{\alpha}{\beta}\left(1 - e^{-\beta(t_n - t_i)}\right) + \sum_{i=1}^{n} \ln\left(\lambda_0 + \sum_{t_k < t_i} \alpha e^{-\beta(t_i - t_k)}\right)$$

Par conséquent, pour obtenir les paramètres optimaux, il suffit de résoudre :

$$(\widehat{\lambda_0}, \widehat{\alpha}, \widehat{\beta}) = \underset{\lambda 0, \alpha, \beta}{\operatorname{argmax}} \left(t_n - t_n\lambda_0 - \sum_{i=1}^{n} \frac{\alpha}{\beta}\left(1 - e^{-\beta(t_n - t_i)}\right) + \sum_{i=1}^{n} \ln\left(\lambda_0 + \sum_{t_k < t_i} \alpha e^{-\beta(t_i - t_k)}\right)\right)$$

Cette estimation de paramètres est effectuée pour chaque classe de contrats construite précédemment.

	λ_0^*	α^*	β^*
Classe 1	0,000014	0,00032	0,00032
Classe 2	0,000050	0,00039	0,00039
Classe 3	0,000097	0,00095	0,00131
Classe 4	0,000189	0,00155	0,00208
Classe 5	0,000345	0,00263	0,00319

TABLE 6.1 – Paramètres estimés des processus de Hawkes de chaque classe

6.3 Validation du modèle et pistes d'amélioration

6.3.1 Validation

Pour analyser la qualité du modèle construit, il convient de le comparer avec des méthodes alternatives, en particulier avec le modèle en place. Pour refléter la qualité de chaque modèle, on prendra en compte le critère de l'erreur quadratique moyenne (ou MSE : *Mean Square Error*) entre les valeurs prédites et les valeurs observées. Cette valeur présente l'avantage d'être comparable à la variance des fréquences observées, qui représente sensiblement l'erreur quadratique moyenne dans

le cas où l'estimateur se résume à l'espérance (sans segmentation). Cette valeur est donc donnée par la formule :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Trois modèles seront alors comparés :

- La modélisation de la fréquence par processus de Hawkes ;
- Le modèle de fréquence en place (segmentation par arbre CART pour chaque DOC, et prise en compte de l'antériorité de sinistralité par une méthode apparentée aux MLG, pour expliquer d'un côté la probabilité de survenance d'au moins un sinistre, et de l'autre le nombre dans le cas où au moins un sinistre est observé) ;
- Un modèle simplifié, estimant la fréquence par MLG pour chaque année relative à la DOC.

Les modèles linéaires généralisés sont des modèles de référence dans les approches fréquence $\times$ coût en assurance de biens et de responsabilité. Ils donnent une référence de qualité de modèle dans un cas simple. Dans notre cas, le modèle cherche à expliquer la fréquence pour chaque année relative à la DOC à l'aide des covariables suivantes :

- Le coût du chantier ;
- Le type de construction ;
- La qualité du souscripteur ;
- La présence ou non d'un accord cadre ;
- Le nombre de sinistres observés l'année précédente ;
- Le nombre de sinistres cumulés jusqu'à l'année précédente.

Un développement théorique des modèles linéaires généralisés est effectué en annexe.

Pour cette phase de validation, chaque modèle effectue son apprentissage sur les données relatives à la sinistralité disponibles au 31 décembre 2018. La prédiction est ensuite effectuée sur l'année 2019, et est alors comparée à la sinistralité réellement observée pour chaque contrat. Pour chaque contrat, la valeur prédite considérée est la moyenne de la fréquence simulée sur 1000 scénarios. La variance observée est de 1,48.

	MSE
Processus de Hawkes	0,80
Modèle actuel	0,92
Modèle simplifié MLG	0,91

TABLE 6.2 – MSE des différents modèles de projection de la fréquence

Le MSE constaté pour la prédiction de l'année 2019 est le meilleur pour le modèle construit utilisant les processus de Hawkes. Il est toutefois important de prendre du recul par rapport à cette validation, puisque s'il semble le meilleur sur la projection de la sinistralité à un an, cela pourrait être différent sur une projection de plusieurs années.

6.3.2 Améliorations possibles du modèle

La principale limite du modèle construit pour l'ajustement des trajectoires de sinistralité par des processus de Hawkes repose sur une hypothèse prise initialement. Le processus de Hawkes ajusté est un processus de Hawkes simple, c'est-à-dire que l'intensité de base λ_0 est supposée constante. En d'autres termes, la probabilité de survenance de sinistres est a priori présumée comme étant la même à n'importe quel moment de développement du contrat. Un processus pris au temps $t = 0$ prédira donc en moyenne autant de sinistres la première année du contrat que la septième ou la treizième. Cependant, ce n'est pas le cas en pratique, où la sinistralité se concentre entre la quatrième et la onzième année des contrats.

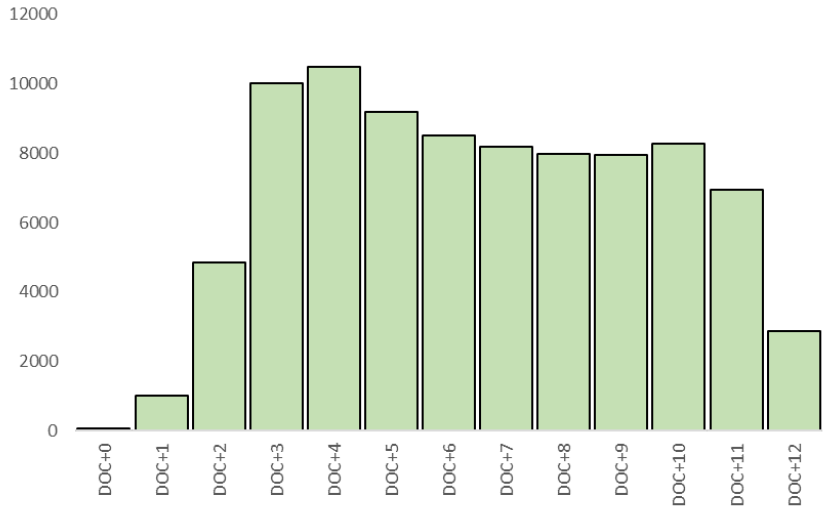

FIGURE 6.4 – Sinistralité par année relative à la DOC

Cette différence s'explique par l'hétérogénéité des contrats. En effet, certains chantiers se terminent plus rapidement que d'autres, dans ce cas la garantie décennale se déclenche plus tôt et se terminera donc avant les 13 ans supposés du processus. Dans le cas contraire où le chantier dure longtemps, le processus aura bien une maturité de 13 ans, mais les trois premières années seront en moyenne moins sinistrées. Une solution serait alors de considérer l'intensité de base du processus comme variable au cours du temps. L'estimation des paramètres se révèlerait alors bien plus complexe.

Un autre défaut de cette procédure est qu'elle prend peu en compte la déformation du portefeuille, contrairement au modèle actuellement en place. En effet, l'apprentissage sur la fréquence effectué pour la classification en groupes de contrats homogènes ainsi que l'estimation des para-

mètres des processus de Hawkes se fait exclusivement grâce aux contrats dont la période de garantie est terminée, et ne prend donc pas en compte les données relatives aux contrats actuellement en cours. Une évolution du risque au sein du portefeuille ne serait alors pas immédiatement détectée et le modèle de fréquence pourrait donc s'avérer déficient.

Chapitre 7

Modélisation de la dépendance règlement-recours-durée de vie à l'aide de copules

Les recours étant une variable primordiale dans le fonctionnement de la garantie DO, il semble important de considérer cette variable dans le modèle pour rendre ce dernier plus précis. Compte tenu de ce constat, on cherche alors à intégrer cette nouvelle variable dans le modèle en place, en gardant l'idée de dépendance par copule entre les trois variables.

L'analyse préliminaire individuelle des trois variables d'intérêt renvoie immédiatement vers une particularité : le nombre de sinistres ayant un montant de recours ultime nul est très important. Ceci peut être expliqué par la présence de nombreux sinistres ayant donné lieu à un montant d'indemnisation nul. Seuls les frais d'expertise composent alors le montant de règlement, et les recours ne sont pas exercés sur ces paiements, conformément à la CRAC. Pour remédier à cette spécificité, le nouveau modèle sera divisé en deux. Une fois le montant de règlement modélisé, on cherchera dans un premier temps à estimer la probabilité qu'un sinistre se clôture avec un montant de recours égal à zéro. Dans un second temps, deux modèles distincts seront mis en place. Le premier expliquant la structure de corrélation entre le montant de règlement et la durée de vie du sinistre conditionnellement à l'absence de recours, et le second expliquant la structure de corrélation entre les trois variables conditionnellement à un recours strictement positif.

Avant de développer la procédure mise en place, il est nécessaire d'évoquer les aspects essentiels de la théorie des copules.

7.1 Généralités sur les copules

7.1.1 Les copules multi-variées

Cette partie a pour but d'énoncer les principes généraux concernant les copules, quelle que soit leur dimension. On parle alors de copules multi-variées.

En termes pratique, on peut définir une copule comme une fonction permettant de lier la fonction de répartition jointe de plusieurs variables aléatoires avec les fonctions de répartition marginales de celles-ci. Autrement dit, si on note $X_1, ..., X_d$ les variables aléatoires et C la copule qui lie ces variables aléatoires, on a :

$$\mathbb{P}\left(X_1 \leq x_1, ..., X_d \leq x_d\right) = C\left(\mathbb{P}\left(X_1 \leq x_1\right), ..., \mathbb{P}\left(X_d \leq x_d\right)\right)$$

En termes plus techniques, il est nécessaire de rappeler la définition et les principales propriétés d'une fonction de répartition multi-variée, avant de définir formellement ce que sont les copules.

Définition 7.1.1. *Soit $X = (X_1, ..., X_d)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^d$.*

- *La fonction F définie par*

$$F(x) = \mathbb{P}\left(\bigcap_{i=1}^{d} X_i \leq x_i\right) \qquad x = (x_1, ..., x_d) \in \mathbb{R}^d$$

  est appelée ***la fonction de répartition multivariée*** *de X.*

- *Les lois des X_i prises séparément et définies par les fonctions de répartition*

$$F_i(x_i) = \mathbb{P}\left(X_i \leq x_i\right) \quad i = 1, ..., d$$

  sont appelées les ***marginales*** *(ou* ***lois marginales****) de X.*

Propriété 7.1.1. *Soit F une fonction de répartition multivariée, elle vérifie alors les propriétés suivantes :*

- *F est continue à droite ;*
- $\lim\limits_{x_i \to -\infty} F(x_1, ..., x_d) = 0 \quad \forall i \in (1, ..., d)$;
- $\lim\limits_{x_i \to +\infty \, \forall i} F(x_1, ..., x_d) = 1$.

Définition 7.1.2. *Une* ***copule*** *est une fonction de répartition multivariée dont les marginales sont uniformes sur $[0, 1]$.*

Théorème 7.1.1. *Soit $d \geq 2$, C une fonction de $[0,1]^d$ à valeurs dans $[0,1]$ et $u = (u_1, ..., u_d) \in [0,1]^d$. La fonction C est une copule si et seulement si elle vérifie les conditions suivantes :*

- *Soit $k \in \{1, ..., d\}$. Si $u_k = 0$ alors $C(u) = 0$.*
- *Soit $k \in \{1, ..., d\}$. Si $u_i = 1 \; \forall i \neq k$, alors $C(u) = u_k$.*
- *C est supermodulaire. C'est à dire que pour tous $u^1 = (u_1^1, ..., u_d^1)$ et $u^2 = (u_1^2, ..., u_d^2) \in [0,1]^d$ tels que $u_i^1 \leq u_i^2 \quad \forall i \in \{1, ..., d\}$, l'inégalité suivante est vérifiée :*

$$\sum_{k_1=1}^{2} ... \sum_{k_d=1}^{2} (-1)^{k_1+...+k_d} C\left(u_1^{k_1}, ..., u_d^{k_d}\right) \geq 0$$

Le théorème suivant, énoncé par Sklar (1959), est primordial dans la théorie des copules, il permet d'assurer l'existence d'une fonction liant plusieurs variables aléatoires, quelles qu'elles soient, permettant d'expliquer la structure de dépendance.

Théorème 7.1.2.

- *Soient C une copule de dimension d et $(F_i)_{1 \leq i \leq d}$ d fonctions de répartition. Alors $F(x) = C(F_1(x_1), ..., F_d(x_d))$ est une fonction de répartition multivariée, ayant les fonctions F_i pour marginales.*
- *Réciproquement, soit F une fonction de répartition multivariée de marginales F_i, $i \in \{1, ..., d\}$. Il existe une copule C telle que pour tout $x \in \mathbb{R}^d$:*

$$F(x) = C(F_1(x_1), ..., F_d(x_d))$$

De plus, si les F_i sont continues, C est unique.

Dans le cas de marges continues, on peut extraire l'unique copule associée à F par la formule suivante :

$$C(u_1, ..., u_d) = F\left(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)\right) \quad \forall u = (u_1, ..., u_d) \in [0,1]^d$$

où F_i^{-1} est la fonction inverse de F_i, $i = 1, ..., d$.

Une propriété intéressante d'une copule est son invariance par rapport à des transformations strictement croissantes des variables.

Propriété 7.1.2.
Soit $(X_1, ..., X_d)$ un vecteur aléatoire de fonction de répartition F et de marginales $F_1, ..., F_d$. Soient $(t_j)_{1 \leq j \leq d}$ des fonctions strictement croissantes.

Alors, la copule $\widetilde{C}$ associée à la loi $\widetilde{H}$ du vecteur aléatoire $(t_1(X_1), ..., t_d(X_d))$ est la même que la copule C associée à H.

Ce résultat permet de constater que les lois marginales n'affectent pas la structure de dépendance. En d'autres termes, on peut normaliser les variables de telle sorte qu'elles soient distribuées uniformément sur l'intervalle $[0,1]$, tout en conservant la nature de la dépendance.

Définition 7.1.3. *On appelle* ***densité*** *d'une copule la fonction c définie sur* $[0,1]^d$ *définie par :*

$$c\left(u_1,...,u_d\right)=\frac{\partial^d C\left(u_1,...,u_d\right)}{\partial u_1...\partial u_d}$$

Rappel : Soit $X=(X_1,...,X_d)$ un vecteur aléatoire de fonction de répartition F et $x=(x_1,...,x_d)\in\mathbb{R}^d$. La loi conditionnelle de la variable X_i est donnée par la formule suivante :

$$\mathbb{P}\left(X_i\leq x_i|X_1=x_1,...,X_{i-1}=x_{i-1},X_{i+1}=x_{i+1},...,X_d=x_d\right)=\frac{\partial^{d-1}F\left(x_1,...,x_d\right)}{\partial x_1...\partial x_{i-1}\partial x_{i+1}...\partial x_d}$$

Il est intéressant de définir deux fonctions remarquables, permettant de borner toutes les copules.

Définition 7.1.4.
Soient $W:[0,1]^d\mapsto[0,1]$ *et* $M:[0,1]^d\mapsto[0,1]$ *les fonctions définies par les formules :*

$$W\left(u\right)=\max\left(\sum_{i=1}^{d}u_i-d+1,0\right)$$

$$M\left(u\right)=\min_{1\leq i\leq d}u_i.$$

Alors pour toute copule C *et pour tout* $u=(u_1,...,u_d)\in[0,1]^d$*, on a :*

$$W\left(u\right)\leq C\left(u\right)\leq M\left(u\right).$$

Les fonctions W *et* M *sont appelées les bornes de Fréchet-Hoeffding.*

Remarque : En dimension $d>2$, la fonction W n'est pas une copule.

Définition 7.1.5. *La* ***copule indépendante*** *est donnée par la formule :*

$$C_{\perp}\left(u_1,...,u_d\right)=\prod_{i=1}^{d}u_i$$

C'est la copule permettant de lier la fonction de répartition jointe aux lois marginales lorsque les variables sont indépendantes.

Définition 7.1.6.
Soient $X = (X_1, ..., X_d)$ un vecteur aléatoire et $x^1, ..., x^n \in \mathbb{R}^d$ des observations de ce vecteur. On note R_i^j le rang de l'observation x_i^j parmi les observations de la variable X_i. Soit $u = (u_1, ..., u_d) \in \mathbb{R}^d$.
La fonction $C_n : [0,1]^d \mapsto [0,1]$ donnée par la formule :

$$C_n(u_1, ..., u_d) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\left\{ \frac{R_1^j}{n} \leq u_1, ..., \frac{R_d^j}{n} \leq u_d \right\}}$$

*est appelée **copule empirique** associée aux observations.*

Il existe deux principales catégories de copules usuelles. D'une part les copules elliptiques comme la copule Gaussienne ou de Student et de l'autre les copules archimédiennes comme les copules de Gumbel ou Clayton. Cette dernière catégorie sera développée ultérieurement dans le but d'aborder les copules archimédiennes hiérarchiques. Les copules elliptiques, occupant une place secondaire dans le cadre de ce mémoire, seront présentées en annexe.

7.1.2 Particularités du cas bi-varié

L'étude des corrélations entre deux variables permet d'utiliser des outils supplémentaires propres à cette dimension.

7.1.2.1 Les mesures de corrélation

Le cas bi-varié permet de calculer des mesures de corrélation entre les deux variables, et donc de résumer la structure de dépendance en valeurs numériques. On note X et Y deux variables aléatoires.

Le coefficient de corrélation linéaire

Le coefficient de corrélation linéaire (ou coefficient de Pearson) est la mesure de dépendance la plus fréquemment utilisée. Il est défini par la formule suivante :

$$r_{lin}(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Il permet, comme son nom l'indique, de mesurer la corrélation linéaire entre deux variables. Cependant, il ne capte pas les effets de corrélation non linéaires. Par exemple, si X suit une loi $\mathcal{U}(-1, 1)$, $r_{lin}(X, X) = 1$ tandis que $r_{lin}(X, X^2) = 0$, alors que la variable X^2 est entièrement déterminée par X.

Cet indicateur a aussi la particularité de nécessiter une variance finie, ce qui n'est pas toujours le cas pour les variables étudiées. Dans le cas où les variances entre les variables comparées ne sont pas du même ordre de grandeur ($Var(X)$ très grande et $Var(Y)$ très petite par exemple), le coefficient de corrélation linéaire peut être très proche de 0 alors que la corrélation est maximale.

En connaissance de ces défauts, il est alors opportun d'introduire d'autres mesures de corrélation.

Le τ de Kendall

Le τ de Kendall permet de mesurer la corrélation entre deux variables en observant la concordance des observations par comparaison de leurs rangs. De la même manière que le coefficient de corrélation linéaire, le τ de Kendall est compris entre -1 (corrélation négative) et 1 (corrélation positive).

Soient $\widetilde{X}$ et $\widetilde{Y}$ des variables aléatoires de mêmes lois que X et Y respectivement. Cette mesure est donnée par la formule suivante :

$$\tau = \mathbb{P}\left(\left(X - \widetilde{X}\right)\left(Y - \widetilde{Y}\right) > 0\right) - \mathbb{P}\left(\left(X - \widetilde{X}\right)\left(Y - \widetilde{Y}\right) < 0\right)$$

Le terme $\mathbb{P}\left(\left(X - \widetilde{X}\right)\left(Y - \widetilde{Y}\right) > 0\right)$ est appelé la probabilité de concordance.
Le terme $\mathbb{P}\left(\left(X - \widetilde{X}\right)\left(Y - \widetilde{Y}\right) < 0\right)$ est appelé la probabilité de discordance.

Soient n observations du vecteur aléatoire (X, Y). La version empirique du τ de Kendall est donnée par :

$$\widehat{\tau} = \frac{\text{(nombre de paires concordantes)} - \text{(nombre de paires discordantes)}}{\frac{1}{2}n(n-1)}$$

Exemple :

Pour illustrer ce calcul, on considère $n = 3$ et les observations suivantes pour les variables X et Y

X	Y
10	3
15	2
100	20

Nous allons alors considérer chaque paire d'observations et observer si X est dans le même ordre que Y (paire concordante) ou si X et Y sont en ordre inversé (paire discordante).

Nous observons donc une paire discordante (X et Y ne sont pas dans le même ordre) :

X	Y
10	3
15	2

Et deux paires concordantes :

X	Y
10	3
100	20

X	Y
15	2
100	20

Cela donne donc :

$$\widehat{\tau} = \frac{2-1}{\frac{1}{2} \times 3 \times 2} = \frac{1}{3}$$

Dans le cas de deux variables X et Y liées par une copule C, le taux de Kendall est donné par la formule suivante :

$$\widehat{\tau} = 4 \int_0^1 \int_0^1 C(u,v)\, dC(u,v) - 1$$

Une formule explicite découle de cette expression pour la plupart des copules usuelles.

Le ρ de Spearman

Une autre mesure de corrélation alternative est le ρ de Spearman. De la même manière que le τ de Kendall, le ρ de Spearman s'intéresse à la relation entre les rangs des deux variables, et est donné par la formule :

$$\rho_S(X,Y) = r_{lin}(F_X(x), F_Y(y))$$

En fait, le ρ de Spearman correspond au coefficient de corrélation linéaire appliqué aux fonctions de répartition, qui correspondent aux rangs des variables dans son expression empirique. Ainsi, cette mesure ne dépend plus des lois marginales.

A l'instar du τ de Kendall, le ρ de Spearman admet une expression sous forme intégrale dans le cas de deux variables X et Y liées par une copule C :

$$\rho_S = 12 \int_0^1 \int_0^1 C(u,v) du dv - 3$$

Les coefficients de dépendance forte des extrêmes

Les coefficients de dépendance forte des extrêmes, comme leur nom l'indique, ne s'intéressent pas à la liaison entre deux variables dans sa globalité, mais à leurs comportements joints lorsqu'elles prennent des valeurs extrêmes.

Soient $U \sim \mathcal{U}(0,1)$ et $V \sim \mathcal{U}(0,1)$, les coefficients de dépendance forte des extrêmes sont définis de la manière suivante :

$$\lambda_L = \lim_{u \to 0} P(U \leq u \mid V \leq u) = \lim_{u \to 0} \frac{C(u,u)}{u}$$

$$\lambda_U = \lim_{u \to 1} \mathbb{P}(U > u \mid V > u) = \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1-u}$$

En posant $U = F_X(X)$ et $V = F_Y(Y)$, on dit alors que X et Y présentent de la dépendance forte des extrêmes à droite (respectivement à gauche) si $\lambda_U > 0$ (respectivement $\lambda_L > 0$).

Malgré la possibilité de décrire la structure de corrélation entre deux variables aléatoires par un seul nombre dans le cas bi-varié, cela reste un résumé trop violent. La structure de dépendance décrite par les copules se révèle bien plus complète.

7.1.2.2 Les représentations graphiques des copules

L'étude d'une structure de corrélation par copule dans le cas bivarié présente une simplicité de représentation qui ne se retrouve pas dans les dimensions supérieures. Il est en effet possible de représenter la corrélation par plusieurs graphiques. Pour l'ensemble de ce paragraphe, il est représenté deux variables aléatoire X et Y suivant une loi $\mathcal{N}(0,1)$ et liées par une copule Gaussienne de paramètre $0,7$. Les graphiques sont basés sur 10000 observations de ces variables.

Les nuages de points

Le premier type de représentations possible pour illustrer la corrélation entre deux variables sont les nuages de points. Le Scatter Plot représente simplement les observations (X,Y) de la variable aléatoire. Le Rank-rank Plot représente les rangs normalisés associés aux observations des variables aléatoires. C'est en fait le nuage de points $(F_X(X), F_Y(Y)) = (U,V)$ représentant les fonctions de répartitions marginales empiriques appliquées aux observations. Ce dernier permet de faire abstraction de l'effet d'échelle des variables aléatoires ainsi que de leur loi marginale, les observations étant toutes représentées sur $[0,1] \times [0,1]$. Il peut alors être utile pour déterminer le type de copules liant les deux variables.

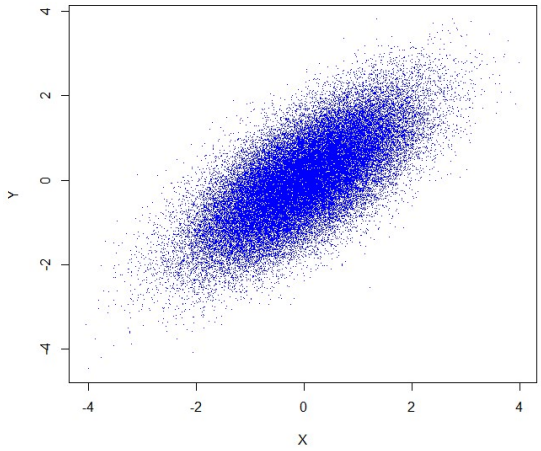

FIGURE 7.1 – Représentation d'un scatter plot

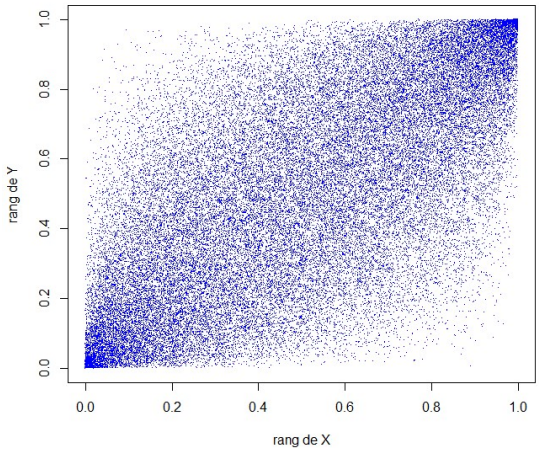

FIGURE 7.2 – Représentation d'un rank-rank plot

La densité

La densité d'une copule peut être représentée dans le cas bi-varié. Le rank-rank plot donne une première idée de la forme de la densité empirique, et peut être complété par une heatmap ou un histogramme 3D.

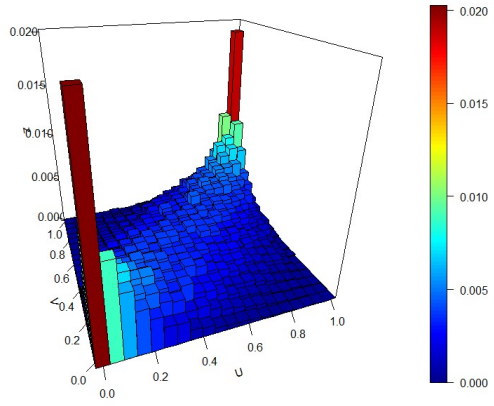

FIGURE 7.3 – Représentation d'un histogramme 3D

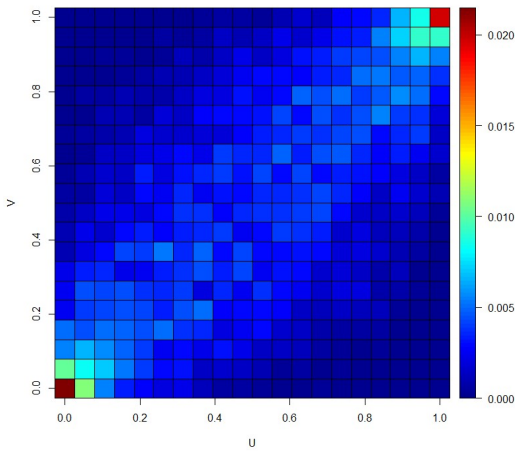

FIGURE 7.4 – Représentation d'une heatmap

La hauteur et/ou la couleur zones sont représentatives du nombre d'observations se trouvant dans chaque zone du graphique. Ces deux graphiques peuvent confirmer la structure de dépendance entre deux variables.

La fonction de répartition bi-variée des rangs

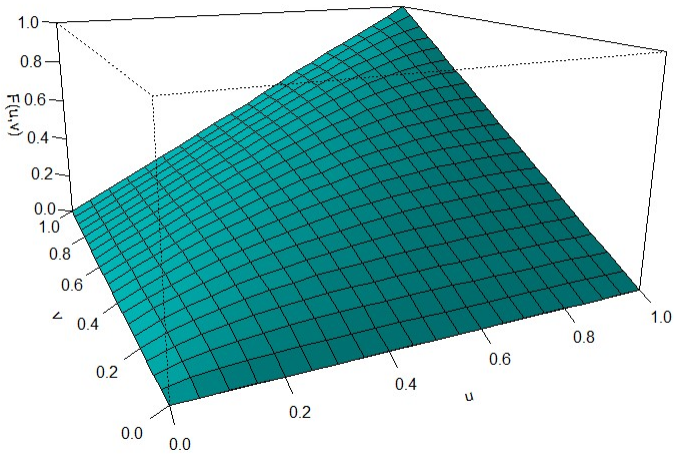

FIGURE 7.5 – Représentation d'une fonction de répartition bi-variée des rangs

La fonction de répartition jointe (la fonction copule empirique) peut également être représentée sur un graphique 3D, bien que ce graphique ne permette pas d'identifier aisément la structure de corrélation.

Le chi-plot

Le chi-plot, proposé par Fisher et Switzer (1985) s'appuie exclusivement sur les couples de rangs, et permet de détecter la dépendance entre deux variables aléatoires continues. Il tire son nom du test d'indépendance du χ_2 dont il s'inspire.

On considère n observations du vecteur aléatoire (X, Y). Etant donnée une paire (X_i, Y_i), $i \in \{1, ..., n\}$, on pose :

$$H_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i, Y_j \leq Y_i\}$$
$$F_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i\}$$
$$G_i = \frac{1}{n-1} \# \{j \neq i, Y_j \leq Y_i\}$$

Sous l'hypothèse d'indépendance entre les deux variables, on s'attend à ce que :

$$H_i = F_i G_i \quad \forall i = 1, ..., n$$

C'est cette propriété qui est exploitée pour construire un chi-plot.

Pour construire ce graphique, les paires (λ_i, χ_i) sont représentées, avec :

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i\,(1-F_i)\,G_i\,(1-G_i)}}$$

et

$$\lambda_i = 4\,\text{sign}\left(\left(F_i - \frac{1}{2}\right)\left(G_i - \frac{1}{2}\right)\right)\max\left(\left(F_i - \frac{1}{2}\right)^2, \left(G_i - \frac{1}{2}\right)^2\right) \qquad 1 \leq i \leq n$$

$\lambda_i \in [-1, 1]$ mesure la distance entre les couples (X_i, Y_i) et le centre du nuage de points. Afin de supprimer les valeurs aberrantes, les paires vérifiant l'inégalité suivante sont retirées.

$$|\lambda_i| \geq 4\left(\tfrac{1}{n-1} - \tfrac{1}{2}\right)^2$$

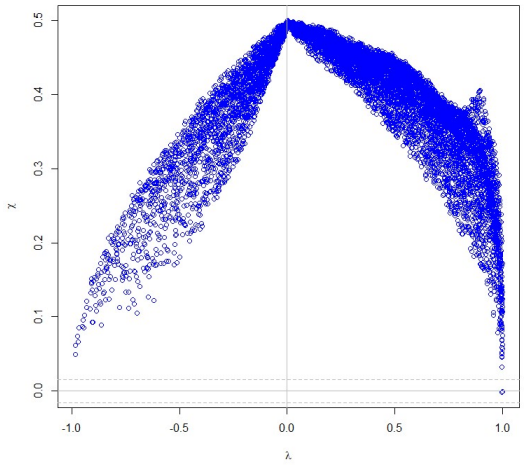

FIGURE 7.6 – Représentation d'un chi-plot avec des variables dépendantes

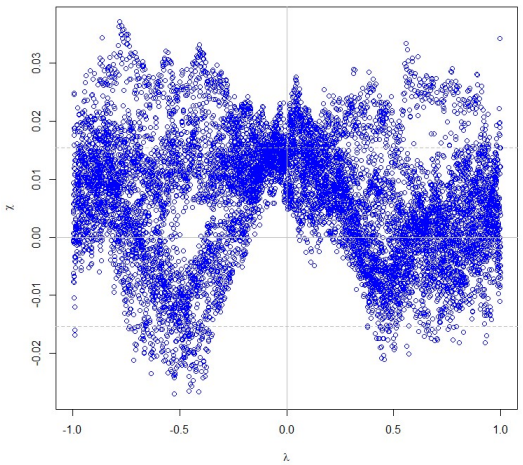

FIGURE 7.7 – Représentation d'un chi-plot avec des variables indépendantes

Lorsque la majorité des points se trouvent hors de l'intervalle de confiance, cela indique une dépendance entre les deux variables. A l'inverse, une majorité de points à l'intérieur de l'intervalle de confiance révèle l'indépendance des variables.

Le Kendall plot

Genest et Boies (2003) sont à l'origine d'une autre représentation graphique permettant de représenter la dépendance. Le Kendall plot (ou K-plot) est, à l'instar du chi-plot, basé sur les rangs des variables aléatoires. Il s'inspire du diagramme quantile-quantile (QQplot) et de la transformation intégrale de probabilité (TIP).

Soit F la fonction de répartition jointe du vecteur (X, Y). La loi $Z = F(X, Y) = C(U, V)$ est appelée la loi de Kendall. Sa fonction de répartition est donnée par la formule :

$$K(w) = \mathbb{P}(F(X, Y) \leq w) \quad w \in [0, 1]$$

On note K_0 la fonction de répartition de la variable Z dans le cas où les variables (X, Y) (et donc (U, V)) sont indépendantes. Cette fonction est donnée par la formule :

$$K_0(w) = \mathbb{P}(UV \leq w) = w - w\ln(w) \quad w \in [0, 1]$$

La procédure utilisée pour tracer le Kendall plot est alors la suivante :

- Calculer $H_i = \frac{1}{n-1}\#\{j \neq i : X_j \leq X_i, Y_j \leq Y_i\} \quad \forall i \in \{1, ..., n\}$;
- Les ordonner : $H_{(1)}, ..., H_{(n)}$;
- Tracer les paires $(W_{i:n}, H_{(i)}) \quad i \in \{1, ..., n\}$, où $W_{i:n}$ est l'espérance de la i-ème statistique d'ordre dans un échantillon aléatoire de taille n de K_0.

Ainsi, le but est de comparer K_0 et la loi empirique de Z donnée par les H_i en observant leur QQplot. Si ces deux variables sont de même loi (les points sont alignés le long de la bissectrice), cela indique l'indépendance de X et Y. Au contraire, plus les points sont éloignés de la bissectrice, plus la dépendance entre les variables est forte.

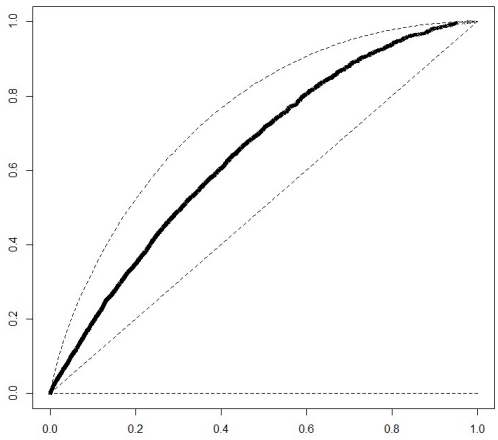

FIGURE 7.8 – Représentation d'un Kendall plot avec des variables dépendantes

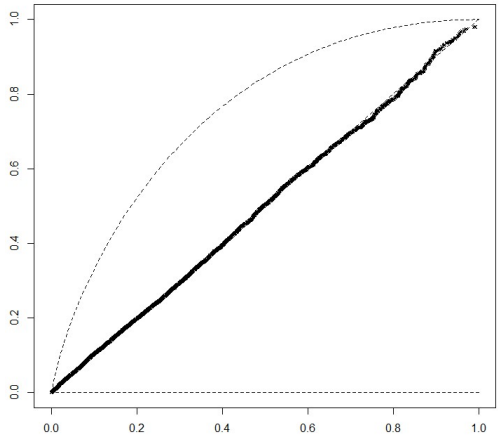

FIGURE 7.9 – Représentation d'un Kendall plot avec des variables indépendantes

Pour illustrer cette remarque, on observe le K-plot des variables dépendantes selon la copule gaussienne de paramètre $0,7$, qui admet une courbe concave au-dessus de la bissectrice. Par opposition, la courbe du K-plot dans le cas de l'indépendance est confondue avec la bissectrice.

7.1.3 Les méthodes d'ajustement

Lorsqu'un ajustement de la structure de dépendance entre plusieurs variables est nécessaire et que des observations sont à disposition, plusieurs méthodes peuvent être utilisées pour estimer les paramètres d'une copule proposée.

On note :

- $X_1, ..., X_d$ les variables aléatoires considérées, $F_1, ..., F_d$ leurs fonctions de répartition et $f_1, ..., f_d$ leurs densités respectives ;
- F la fonction de répartition jointe de $(X_1, ..., X_d)$;
- $x^j = (x_1^j, ..., x_d^j) \in \mathbb{R}^d$, $j \in \{1, ..., n\}$ les observations du vecteur $X = (X_1, ..., X_d)$;
- C_θ la copule de paramètre θ sous-jacente à la loi du vecteur $(X_1, ..., X_d)$, de densité associée $c(\cdot\,; \theta)$.

7.1.3.1 La méthode du maximum de vraisemblance

Pour utiliser la méthode du maximum de vraisemblance, on suppose que chaque loi marginale X_i appartient à une famille de lois indexée par un paramètre α_i. Les paramètres des lois et de la copule sont alors estimés conjointement en maximisant la fonction de vraisemblance :

$$\left(\widehat{\alpha_1}, ..., \widehat{\alpha_d}; \widehat{\theta}\right) = \underset{\alpha_1,...,\alpha_d,\theta}{\operatorname{argmax}} \prod_{j=1}^{n} \frac{\partial^d F(x_1^j, ..., x_d^j)}{\partial x_1 ... \partial x_d}$$

$$\Leftrightarrow \quad \left(\widehat{\alpha_1}, ..., \widehat{\alpha_d}; \widehat{\theta}\right) = \underset{\alpha_1,...,\alpha_d,\theta}{\operatorname{argmax}} \prod_{j=1}^{n} \frac{\partial^d C_\theta(F_1(x_1^j; \alpha_1), ..., F_d(x_d^j; \alpha_d))}{\partial x_1 ... \partial x_d}$$

$$\Leftrightarrow \quad \left(\widehat{\alpha_1}, ..., \widehat{\alpha_d}; \widehat{\theta}\right) = \underset{\alpha_1,...,\alpha_d,\theta}{\operatorname{argmax}} \prod_{j=1}^{n} c_\theta(F_1(x_1^j; \alpha_1), ..., F_d(x_d^j; \alpha_d)) f_1(x_1^j; \alpha_1) \times ... \times f_d(x_d^j; \alpha_d)$$

$$\Leftrightarrow \quad \left(\widehat{\alpha_1}, ..., \widehat{\alpha_d}; \widehat{\theta}\right) = \underset{\alpha_1,...,\alpha_d,\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \ln(c_\theta(F_1(x_1^j; \alpha_1), ..., F_d(x_d^j; \alpha_d))) + \sum_{j=1}^{d} \sum_{i=1}^{d} \ln(f_i(x_i^j; \alpha_i))$$

Remarque : Seule la première somme fait intervenir le paramètre θ de la copule.

L'optimisation d'une fonction de plusieurs variables est exigeante et nécessite donc un temps de calcul qui peut être très long, en particulier lorsque la dimension est grande. De plus, il y a un risque non négligeable d'identifier un maximum local et non le maximum global lors de l'optimisation numérique.

7.1.3.2 La méthode IFM

La méthode Inference Functions for Margins (IFM) a été proposée par Joe (1997) et fonctionne en deux temps. Les notations introduites dans la partie précédente sont reprises.

- On procède d'abord à l'estimation des paramètres $\alpha_1, ...\alpha_d$ des lois marginales, indépendamment de la copule, en maximisant leurs log-vraisemblance :

$$\widehat{\alpha_i} = \underset{\alpha_i}{\operatorname{argmax}} \sum_{j=1}^{n} \ln(f_i(x_i^j; \alpha_i)) \quad \forall i \in \{1, ..., d\}$$

- Ensuite, on estime le paramètre θ de la copule par maximum de vraisemblance, en incorporant dans la formule les estimateurs des paramètres des lois marginales précédemment déterminés :

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \ln(c_\theta(F_1(x_1^j; \widehat{\alpha_1}), ..., F_d(x_d^j; \widehat{\alpha_d})))$$

Cette méthode a l'avantage d'être très facile à implémenter. Cependant, un mauvais choix de lois marginales risque de détériorer l'estimation du paramètre de dépendance θ. Tout comme pour la méthode par maximum de vraisemblance, il est alors primordial d'être très confiant sur la sélection des distributions marginales. De ce fait, lorsque l'ajustement des marges est discutable, il est préférable d'estimer le paramètre de la copule de manière semi-paramétrique. Pour cela, on introduit la méthode CML.

7.1.3.3 La méthode CML

La méthode Canonical Maximum Likelihood (CML) a été proposée par Oakes (1994).

Lors de l'utilisation de cette méthode, on suppose que les marges sont inconnues et on les estime par leur fonction de répartition empirique :

$$F_{i,n}(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{x_i^j \leq x\}} \quad i = 1, ..., d$$

Le paramètre θ de la copule est ensuite déterminé en maximisant la log-vraisemblance :

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \ln(c_\theta(F_{1,n}(x_1^j), ..., F_{d,n}(x_d^j)))$$

L'emploi d'estimations non paramétriques de la méthode peut induire une perte d'information. En effet, la méthode IFM donne des résultats plus satisfaisants lorsque les lois marginales sont connues avec certitude. Toutefois, la méthode CML permet d'obtenir une estimation du ou des paramètres de la copule indépendamment des ajustements des marginales. Elle est aussi facile à implémenter que la méthode IFM, il n'y a qu'à travailler avec les rangs des observations.

7.1.3.4 La méthode des moments pour le cas bi-varié

Si une certaine mesure de corrélation peut s'écrire en fonction de l'unique paramètre de la copule considérée, la méthode des moments permet, par inversion de cette fonction appliquée à la

mesure de corrélation empirique, d'estimer ce paramètre. Les inversions du τ de Kendall ou du ρ de Spearman permettent par exemple de construire des estimateurs des moments pour certaines familles de copules. Cette méthode a l'avantage d'être très facile à implémenter, elle est cependant très peu robuste et ne sera donc pas utilisée par la suite.

Les principaux résultats de la théorie des copules ayant été énoncés, ils pourront alors être mis en pratique au sein de notre démarche aux moments d'ajuster des structures de corrélation. Avant cela, la procédure commence par l'ajustement de la loi marginale du montant de règlement.

7.2 Ajustement de la loi des montants de règlement

Afin de prendre en compte la variabilité des sinistres dans nos scénarios de simulation, il est nécessaire d'ajuster une distribution théorique sur les données. La première étape de la modélisation des sinistres inconnus est celle de leur montant de règlement. On cherche alors à ajuster la loi de cette variable, indépendamment d'un quelconque conditionnement. Pour cette modélisation et les suivantes, seuls les sinistres clos sont considérés.

L'étude de la distribution des montants de règlement fait appel à la théorie des valeurs extrêmes, décrite en annexe. L'ensemble des analyses des lois marginales qui seront décrites dans ce mémoire suivront la même procédure.

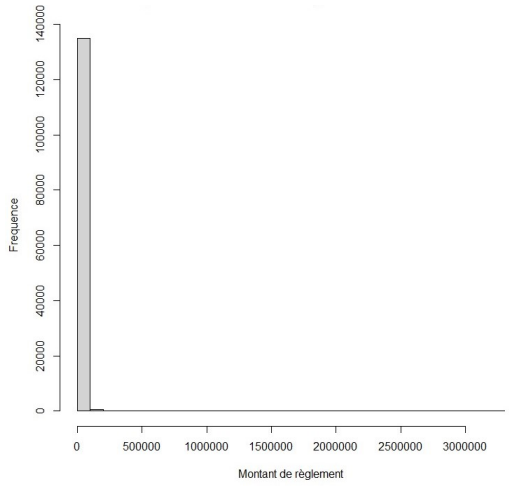

FIGURE 7.10 – Histogramme de la distribution empirique des montants de règlement

Une représentation de la densité empirique de la variable indique que la distribution est concentrée sur les faibles montants, malgré la présence de valeurs de règlement très élevées. L'approche classique est alors de chercher à scinder les sinistres en deux catégories : les sinistres larges et les sinistres attritionnels. Pour cela, après une brève phase exploratoire des données, un seuil doit être déterminé à partir duquel on supposera que la variable suit une GPD. Ensuite les paramètres des distributions seront estimés.

7.2.1 Identification du domaine d'attraction de la loi

La visualisation du diagramme quantile-quantile, dont le but est de comparer la distribution empirique à une distribution de loi exponentielle, peut permettre d'identifier le domaine d'attraction de la loi.

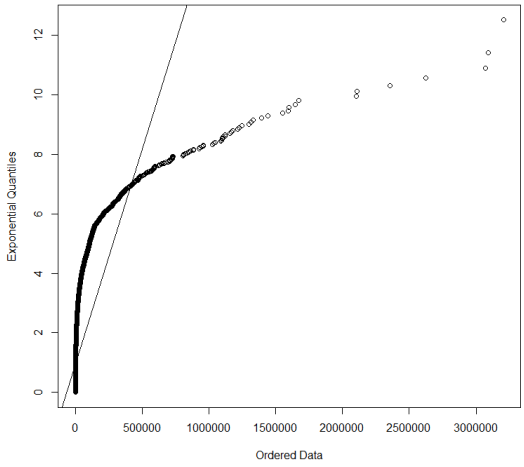

FIGURE 7.11 – QQplot entre la distribution des montants de règlement et la loi exponentielle

On constate que le nuage de points tracé s'approche d'une courbe concave. Autrement dit, la queue de distribution empirique est plus lourde que la distribution exponentielle. La distribution empirique semble donc appartenir au domaine d'attraction de Fréchet.

7.2.2 Identification du seuil

D'après la théorie des valeurs extrêmes développée en annexe, le seuil des sinistres larges doit être fixé à la valeur à partir de laquelle l'espérance résiduelle par rapport à cette valeur est linéaire. On note alors la fonction des excès moyens :

$$e(u) = \mathbb{E}[X - u \mid X > u]$$

On peut aussi considérer la version empirique de cette fonction et l'appliquer sur nos données. Soit $\Delta_n(u) = \{i \mid X_i > u,\ i = 1, ..., n\}$. La fonction des excès moyens empirique est définie par :

$$e_n(u) = \frac{\sum_{i \in \Delta_n(u)} (X_i - u)}{card(\Delta_n(u))}$$

Le graphique des excès moyens (*mean excess plot*) est alors construit en traçant les points

$$\left\{ X_{(i)}, e_n(X_{(i)}), \quad i = 1, ..., n \right\}.$$

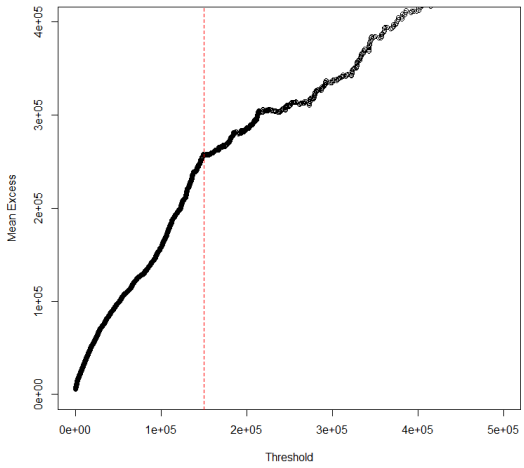

FIGURE 7.12 – Mean excess plot des montants de règlement

On peut alors repérer une « cassure » sur le graphique. En lisant la courbe de droite à gauche, la contrainte de linéarité semble violée aux alentours de 150 000. Cette analyse est limitée à une simple interprétation graphique et est donc subjective. Grâce à la fixation de ce seuil, les sinistres peuvent alors être partagés en deux catégories à ajuster.

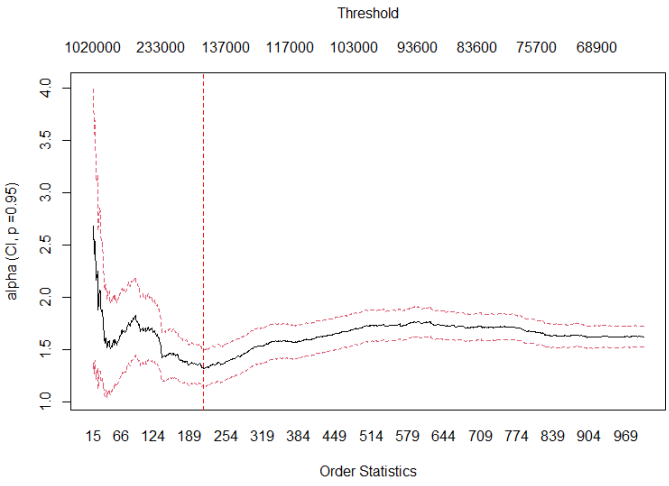

FIGURE 7.13 – Hill plot des montants de règlement

Le tracé du Hill-plot conforte le choix de la valeur seuil, puisque l'estimateur semble se stabiliser autour du seuil de 150 000. De nouveau, cette interprétation graphique reste subjective.

7.2.3 Distribution des montants de règlement supérieurs au seuil

Du fait de la logique inhérente à la construction du seuil, la loi suivie par les montants de règlement des sinistres dépassant ce même seuil (sinistres larges) est une GPD. Les paramètres peuvent alors être estimés par maximum de vraisemblance.

On note :

- $(Y_i)_{1\leq i\leq n}$ le coût résiduel des sinistres au-delà du seuil ;
- $(Y_{(i)})_{1\leq i\leq n}$ les statistiques d'ordre de $(Y_i)_{1\leq i\leq n}$.

On cherche alors $\hat{\xi}$ et $\widehat{\beta}$ tels que :

$$(\hat{\xi},\widehat{\beta}) = \underset{\widehat{\xi},\widehat{\beta}}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\beta}\left(1+\xi\left(\frac{Y_{(i)}}{\beta}\right)\right)^{-\frac{1}{\xi}-1}$$

$$\Longleftrightarrow (\hat{\xi},\widehat{\beta}) = \underset{\widehat{\xi},\widehat{\beta}}{\operatorname{argmax}} \left(-n\times\ln(\beta) - \frac{1}{\xi}\sum_{i=1}^{n}\ln\left(1+\xi\left(\frac{Y_{(i)}}{\beta}\right)\right) - \sum_{i=1}^{n}\ln\left(1+\xi\left(\frac{Y_{(i)}}{\beta}\right)\right)\right)$$

Il n'existe pas de formule explicite pour l'estimation des paramètres. La solution est alors de passer par un algorithme d'optimisation numérique. Le paramètre de forme optimal obtenu est le suivant :

	Valeur optimale
ξ	0,4284

TABLE 7.1 – Paramètre de forme optimal de la GPD pour les montants de règlement

Les analyses graphiques comparant les distributions empirique et théorique indiquent un bon ajustement de la loi aux données. De plus, selon le test de Kolmogorov-Smirnov, on ne peut pas rejeter l'hypothèse selon laquelle les distributions sont identiques. Les paramètres estimés seront alors utilisés dans la suite.

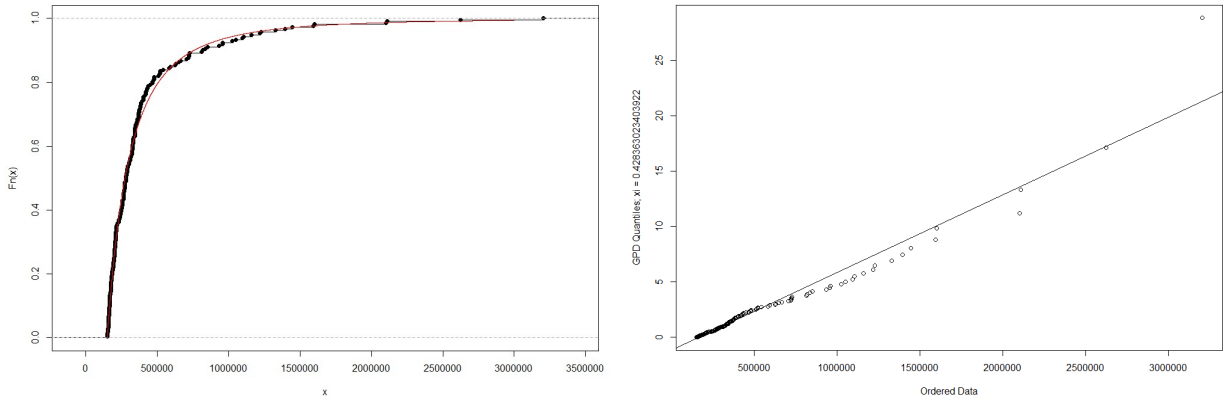

FIGURE 7.14 – Comparaison des distributions empirique et théorique pour les montants de règlement supérieurs au seuil

7.2.4 Distribution des montants de règlement inférieurs au seuil

Outre la variabilité des sinistres que l'on souhaite prendre en compte dans les scénarios de simulation, le montant de règlement des sinistres possède aussi une dépendance avec les caractéristiques du contrat correspondant. Il est supposé que cette dépendance n'est pas présente pour les sinistres larges, puisqu'ils sont par nature exceptionnels quel que soit le type de chantier. Avant l'ajustement d'une distribution pour les montants de règlement en dessous du seuil, une classification par arbre CART de la sévérité des règlements selon les attributs du contrat est mise en place.

7.2.4.1 Classification des contrats par rapport à la sévérité des sinistres attritionnels

On considère les variables suivantes inhérentes aux contrats :

- Le coût du chantier
- Le type de construction
- La qualité du souscripteur
- La région du chantier
- La présence ou non d'un accord cadre

Ces variables doivent permettre de classifier les contrats pour expliquer le montant de règlement des sinistres attritionnels. L'arbre CART optimal est alors construit, sous condition d'un nombre minimal de 15 000 sinistres par feuille, dans le but de disposer d'un effectif suffisant pour ajuster une loi théorique à la distribution empirique de chaque classe. Sous ces conditions, trois classes de contrats sont alors créées.

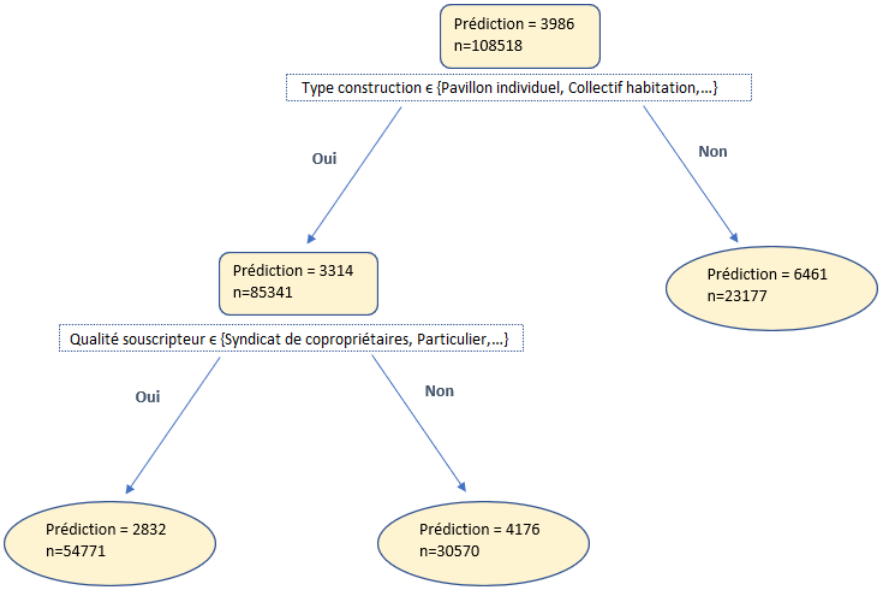

FIGURE 7.15 – Arbre de classification des contrats par rapport à la sévérité du montant de règlement

7.2.4.2 Ajustement d'une distribution pour chaque classe créée

Pour chaque classe de contrats, il est alors nécessaire d'ajuster aux montants de règlement des sinistres une loi tronquée au seuil déterminé auparavant. Ainsi, les paramètres de plusieurs lois usuelles tronquées sont estimés dans le but de sélectionner la distribution théorique représentant le mieux la variable. A cet effet, on introduit les notations suivantes :

- Λ : l'ensemble des distributions considérées ;
- $\mathcal{L}_E, \mathcal{L}_{LN}, \mathcal{L}_W, \mathcal{L}_G, \mathcal{L}_B$: les cinq distributions entrant en considération (Exponentielle, Log-normale, Weibull, Gamma et Burr) ;
- $\Theta_E, \Theta_{LN}, \Theta_W, \Theta_G, \Theta_B$: les domaines de définition respectifs de chaque loi ;
- $\theta_E, \theta_{LN}, \theta_W, \theta_G, \theta_B$: les paramètres respectifs des distributions ;
- S le seuil déterminé.

Pour l'utilisation d'un estimateur de maximum de vraisemblance sur une loi tronquée, la fonction de vraisemblance doit être adaptée en conséquence. On cherche alors à résoudre :

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f_X\left(x_i \mid X \leq S; \theta\right)$$

$$\iff \theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{f_X\left(x_i; \theta\right)}{\mathbb{P}(X \leq S; \theta)}$$

$$\Longleftrightarrow \theta^* = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln\left(f_X\left(x_i;\theta\right)\right) - \ln(\mathbb{P}(X \leq S;\theta))$$

De nouveau, les paramètres optimaux des lois n'admettent pas de formules explicites. Une optimisation numérique est alors effectuée. La démarche est ainsi déroulée pour chaque distribution appartenant à Λ. La meilleure loi parmi les cinq est ensuite déterminée selon le critère de minimisation de la moyenne des écarts quadratiques entre les fonctions de répartition empirique et théorique. Mathématiquement, cela correspond à déterminer $\mathcal{L}^{**}$ par :

$$\mathcal{L}^{**} = \underset{\mathcal{L}\in\Lambda^*}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(F_n\left(x_i\right) - F_{\mathcal{L}}\left(x_i\right)\right)^2$$

avec :

- $F_n(c) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(x_i \leq c\right)$: la fonction de répartition empirique ;
- $F_{\mathcal{L}}(.)$: la fonction de répartition de la loi théorique considérée, tronquée au seuil.

Pour illustrer ce choix de distribution, il est intéressant de lister ces valeurs pour chaque loi considérée et pour chaque classe.

	Exponentielle	Log-normale	Gamma	Weibull	Burr
Classe 1	0,01722	0,00365	0,01245	0,00870	0,00113
Classe 2	0,02753	0,00384	0,01352	0,00839	0,00091
Classe 3	0,03824	0,00440	0,01430	0,00844	0,00073

TABLE 7.2 – Moyenne des écarts quadratiques avec les distributions théoriques pour les montants de règlement en dessous du seuil

La valeur minimum parmi les cinq lois correspond à la loi de Burr pour chaque groupe de risques homogène. C'est alors cette distribution théorique qui ajuste le mieux les données et qui est utilisée par la suite pour chaque classe. Les paramètres de celle-ci sont :

	m	s	f
Classe 1	499,2	4,88	0,19
Classe 2	497,0	4,86	0,15
Classe 3	533,5	5,35	0,12

TABLE 7.3 – Paramètres des lois de Burr ajustées pour chaque classe

De la même manière que pour les sinistres larges, une validation graphique peut permettre de consolider l'ajustement effectué. On construit donc plusieurs représentations graphiques comparant les distributions théoriques et empiriques, à savoir la densité, la fonction de répartition et le QQplot.

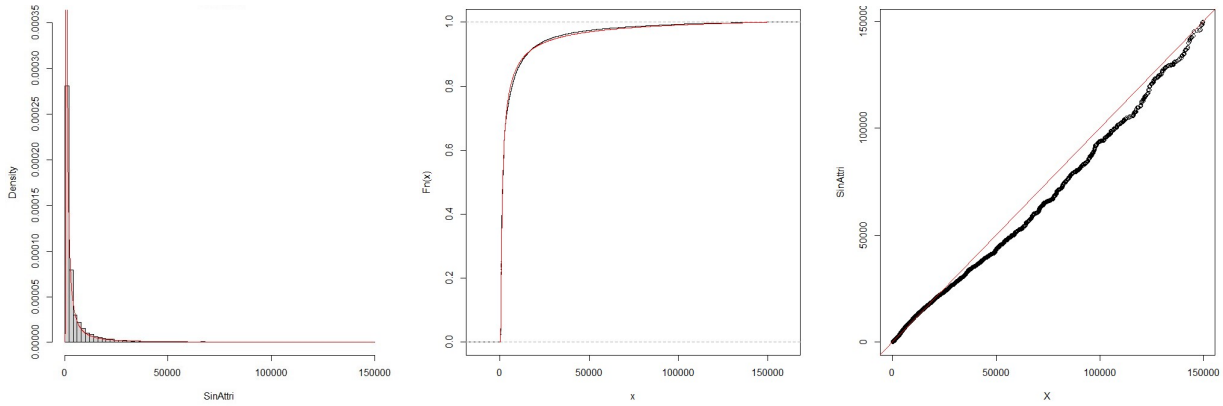

FIGURE 7.16 – Comparaison graphique des distributions empirique et théorique pour la classe 3

L'ajustement à la loi de Burr semble plutôt bon, et ce pour chaque classe ce qui permettra par la suite de faire confiance aux simulations des montants de règlement.

La principale variable relative aux montants des sinistres étant expliquée, il reste à étudier le comportement des montants de recours et des durées de vie relativement à celle-ci.

7.3 Modélisation de la présence ou non de recours

La modélisation de la présence de recours est nécessaire pour appréhender le comportement des sinistres inconnus. On cherche donc dans cette partie à estimer le paramètre de loi (Bernoulli) de la variable binaire $P \in \{0, 1\}$ valant 1 dans le cas d'un recours non nul et 0 sinon.

Le modèle de régression construit utilise un arbre CART, qui prend comme variables explicatives :

- Le montant réglé du sinistre
- Le coût du chantier
- Le type de construction
- La qualité du souscripteur
- La région du chantier
- La présence ou non d'un accord cadre

Seule la variable relative au montant réglé est segmentante dans l'arbre construit. La probabilité modélisée dépend alors seulement de la tranche dans laquelle se trouve la valeur réglée.

Tranche de montant de règlement	Probabilité que le recours soit nul
Moins de 2903 €	0,984
Entre 2903 et 2999 €	0,518
Entre 2999 et 3166 €	0,469
Entre 3166 et 3424 €	0,391
Entre 3424 et 3634 €	0,272
Plus de 3634 €	0,103

TABLE 7.4 – Probabilité qu'un sinistre ait un montant de recours nul conditionnellement à la tranche de montant réglé

7.4 Modélisation conditionnée à l'absence de recours

Lorsque le montant de recours est égal à zéro, la problématique se simplifie puisque l'une des trois variables est alors expliquée. Il ne reste qu'à étudier la structure de corrélation entre les deux variables restantes, à savoir le montant de règlement et la durée de vie du sinistre conditionnellement au fait que le montant de recours soit nul. Par conséquent, le modèle à construire se rapproche fortement du modèle de coût déjà en place, à la différence près que la variable d'intérêt n'est plus la charge nette de recours mais le montant de règlement (variables néanmoins équivalentes lorsque le montant de recours est nul) et que les données utilisées pour l'apprentissage devront être restreintes aux sinistres clos sans recours encaissé.

7.4.1 Ajustement de la loi marginale des montants de règlement

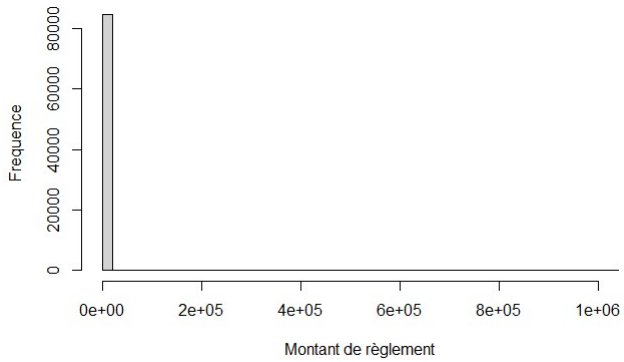

FIGURE 7.17 – Histogramme de la distribution empirique des montants de règlement conditionnés à un recours nul

De la même manière que précédemment, la loi marginale des montants de règlement doit être ajustée pour connaître le comportement de ceux-ci conditionnellement à un montant de recours nul.

A l'instar des règlements des sinistres non conditionnés par la présence de recours, la variable est principalement distribuée sur les faibles montants, malgré la présence de valeurs de règlement très élevées. Les montants des règlements seront alors scindés en deux catégories grâce à la détermination d'une valeur seuil.

7.4.1.1 Identification du domaine d'attraction de la loi

Le diagramme quantile quantile permet l'identification du domaine d'attraction de la loi.

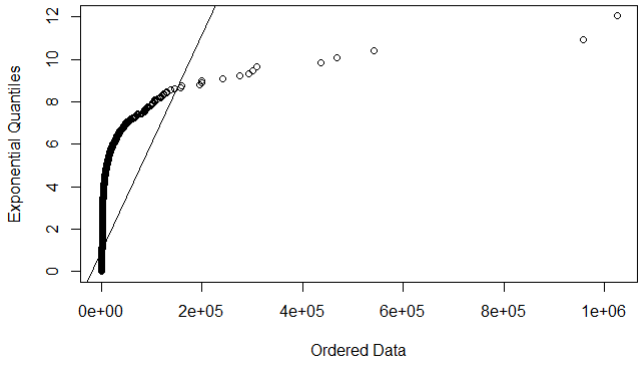

FIGURE 7.18 – QQplot entre la distribution des montants de règlement conditionnés à un recours nul et la loi exponentielle

Le nuage de points représenté s'approche d'une courbe concave. En d'autres termes, la queue de distribution empirique est plus lourde que la distribution exponentielle. La distribution empirique semble donc appartenir au domaine d'attraction de Fréchet.

7.4.1.2 Identification du seuil

Pour l'identification du seuil de grave, le Mean excess plot et le Hill plot sont tracés. On peut alors repérer une « cassure » sur le Mean excess plot. En lisant la courbe de droite à gauche, la contrainte de linéarité semble violée aux alentours de 15 000. Ce seuil est conforté par le Hill plot, puisque l'estimateur semble se stabiliser autour du seuil de 15 000. Cette analyse est limitée à une simple interprétation graphique et est donc subjective. Grâce à la fixation de ce seuil, les montants de règlement peuvent alors être partagés en deux catégories à ajuster.

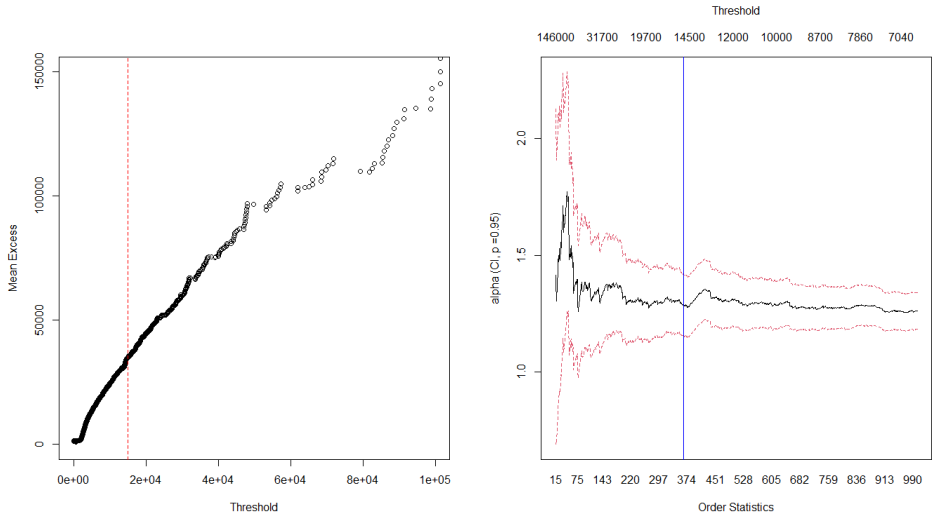

FIGURE 7.19 – Mean excess plot et Hill plot pour la détermination d'un seuil

7.4.1.3 Ajustement des montants de règlement au-dessus du seuil

Une GPD est ajustée pour les montants au-dessus du seuil déterminé. Les paramètres sont estimés par maximum de vraisemblance, et on obtient le paramètre de forme suivant :

	Valeur optimale
ξ	0,7007

TABLE 7.5 – Paramètre de forme optimal de la GPD pour les montants de règlement conditionnés à un recours nul

Les analyses graphiques comparant les distributions empirique et théorique indiquent un bon ajustement de la loi aux données. De plus, le test de Kolmogorov-Smirnov indique que l'on ne peut pas rejeter l'hypothèse selon laquelle les distributions sont identiques. Les paramètres estimés sont alors utilisés dans la suite.

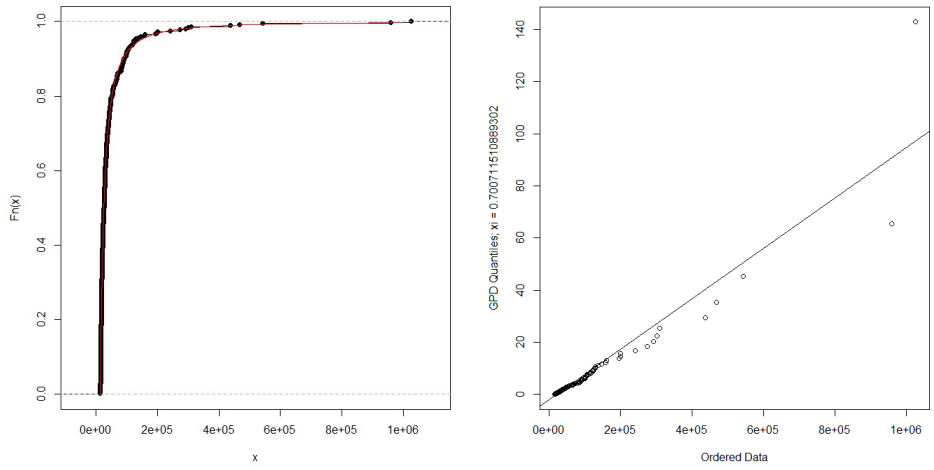

FIGURE 7.20 – Comparaison des distributions empirique et théorique pour les montants de règlement au-dessus du seuil lorsque le recours est nul

7.4.1.4 Ajustement des montants de règlement en dessous du seuil

Pour les montants de règlement inférieurs au seuil, une loi tronquée est ajustée. Pour rappel, les paramètres de plusieurs lois usuelles tronquées sont estimés dans le but de sélectionner la distribution théorique représentant le mieux la variable. Les cinq variables entrant en considération sont la loi Exponentielle, Log-normale, Gamma, Weibull, et Burr. La meilleure loi parmi les cinq est ensuite déterminée selon le critère de minimisation de la moyenne des écarts quadratiques entre les fonctions de répartition empirique et théorique.

De nouveau, la loi de Burr est celle qui ajuste le mieux les données. Les paramètres de celle-ci sont :

	m	s	f
Loi de Burr tronquée	590,1	4,36	0,36

TABLE 7.6 – Paramètres de la loi de Burr tronquée pour les règlements inférieurs au seuil conditionnés à un recours nul

De la même manière que précédemment, une validation graphique peut permettre de consolider l'ajustement effectué. On construit donc plusieurs représentations graphiques comparant les distributions théorique et empirique, à savoir la densité, la fonction de répartition et le QQplot.

L'ajustement à la loi de Burr, malgré le fait qu'il soit le meilleur selon le critère défini précédemment, est de mauvaise qualité. Cette remarque devra être prise en compte dans la méthode d'ajustement de la structure de dépendance. Cependant, cela ne devrait pas poser de problème lors

de l'étape de simulation des sinistres inconnus puisque cette variable sera la première à être simulée, sans prendre en compte le conditionnement à la valeur du montant de recours. Le plus important est donc d'avoir obtenu un ajustement de bonne qualité au global.

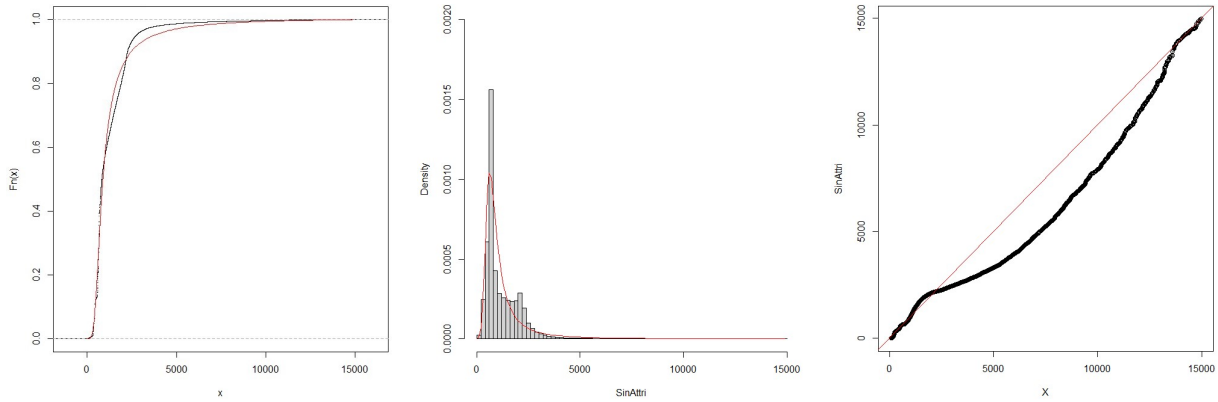

FIGURE 7.21 – Comparaison des distributions empirique et théorique pour les montants de règlement en dessous du seuil lorsque le recours est nul

7.4.2 Ajustement de la loi marginale de la durée de vie des sinistres

Cette nouvelle variable est également distribuée principalement sur des faibles valeurs, avec toutefois la présence de quelques valeurs très élevées.

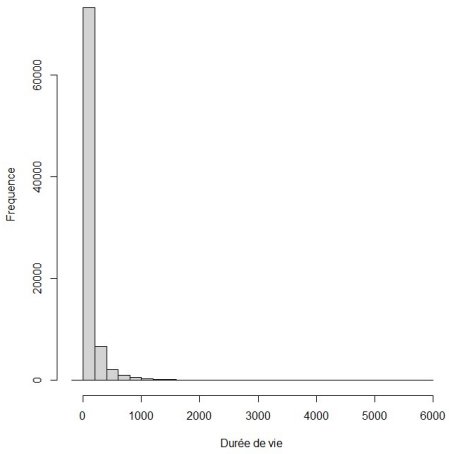

FIGURE 7.22 – Histogramme de la distribution empirique de la durée de vie des sinistres conditionnée à un recours nul

La démarche effectuée pour l'ajustement de la loi marginale de la durée de vie des sinistres conditionnellement à l'absence de recours suit la même procédure que pour les montants de règlement des sinistres. Autrement dit, un seuil est d'abord déterminé dans le but d'ajuster les durées supérieures à une GPD, et les durées inférieures à une loi usuelle tronquée.

7.4.2.1 Identification du domaine d'attraction de la loi

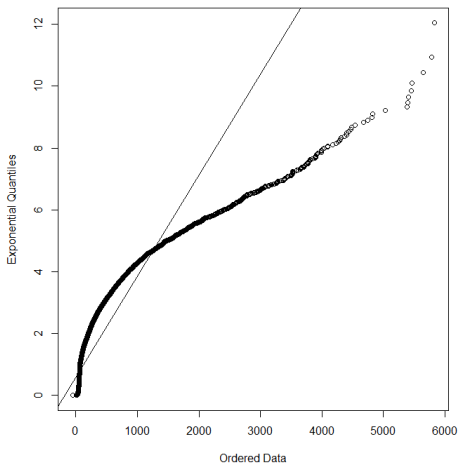

FIGURE 7.23 – QQplot entre la distribution des durées de vie des sinistres conditionnées à un recours nul et la loi exponentielle

Le diagramme quantile quantile permet l'identification du domaine d'attraction de la loi. Le nuage de points représenté s'approche d'une courbe concave. En d'autres termes, la queue de distribution empirique paraît plus lourde que la distribution exponentielle, même si cela est plus léger que pour les montants de règlement. Avec cette information, on peut émettre l'hypothèse que la distribution empirique appartient au domaine d'attraction de Fréchet. Cela devra tout de même être confirmé par la suite.

7.4.2.2 Identification du seuil

Pour l'identification du seuil de grave, le mean excess plot est tracé. On peut alors y repérer une « cassure ». En lisant la courbe de droite à gauche, la contrainte de linéarité semble violée aux alentours de 2000 jours. Cette analyse est limitée à une simple interprétation graphique et est donc subjective. Grâce à la fixation de ce seuil, les durées de vie des sinistres pourront alors être partagées en deux catégories à ajuster. Il est important de noter que le mean excess plot décroit assez fortement après le seuil. Cela peut indiquer que le domaine d'attraction de la loi n'est pas celui de Fréchet, mais de Gumbel voire Weibull. Ce doute sera levé lors de l'estimation du paramètre de la GPD.

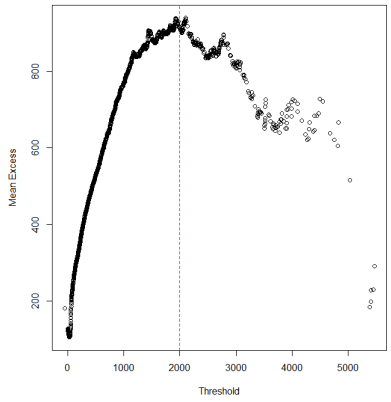

FIGURE 7.24 – Mean excess plot de la durée de vie pour la détermination d'un seuil

7.4.2.3 Ajustement des durées de vie des sinistres au-dessus du seuil

Une GPD est ajustée pour les montants au-dessus du seuil construit. Les paramètres sont estimés par maximum de vraisemblance. On obtient les ajustements suivants :

	Valeur optimale
ξ	-0,1623

TABLE 7.7 – Paramètre de forme optimal de la GPD pour les durées de vie conditionnées à un recours nul

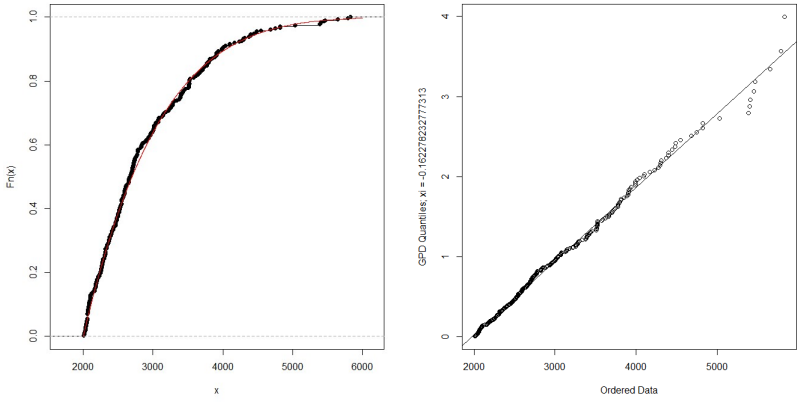

FIGURE 7.25 – Confrontation des distributions empirique et théorique pour les durées de vie supérieures au seuil

La paramètre de forme ξ estimé est négatif, ce qui confirme que le domaine d'attraction de la

loi est celui de Weibull. L'analyse graphique de l'adéquation de la GPD aux données montre un ajustement de bonne qualité.

7.4.2.4 Ajustement des durées de vie des sinistres en dessous du seuil

A l'image de la modélisation des montants de règlement, les durées de vie en dessous du seuil de 2000 jours sont ajustées aux plusieurs mêmes lois usuelles tronquées, à savoir la loi Exponentielle, Log-normale, Gamma, Weibull et Burr. De nouveau, la meilleure distribution théorique est la loi de Burr, avec les paramètres suivants :

	m	s	f
Loi de Burr tronquée	37,98	5,73	0,23

TABLE 7.8 – Paramètres de la loi de Burr ajustée pour les durées de vie inférieures au seuil conditionnées à un recours nul

L'ajustement semble meilleur que pour les montants de règlement. Même s'il n'est pas parfait, il parait assez satisfaisant pour modéliser sereinement la variable correspondant aux durées de vie des sinistres.

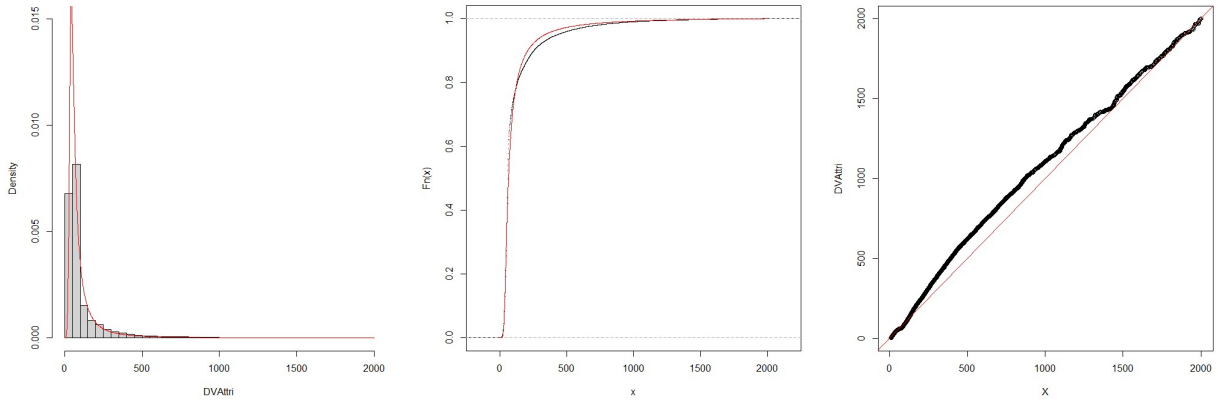

FIGURE 7.26 – Comparaison graphique des distributions empirique et théorique pour les durées de vie des sinistres inférieures au seuil conditionnées à un recours nul

7.4.3 Modélisation de la corrélation entre règlement et durée de vie

Dorénavant, dans le cas supposé où le montant de recours est égal à zéro, on sait de quelle manière se comportent les variables des montants de règlement et des durées de vie des sinistres indépendamment l'une de l'autre. Le but est maintenant de connaître leur comportement conjoint, c'est-à-dire de déterminer l'attitude que va avoir une variable conditionnellement à la valeur de

l'autre. On a pour cela introduit précédemment la théorie des copules, qui permet d'expliquer entièrement la structure de dépendance entre plusieurs variables.

7.4.3.1 Mise en évidence de la dépendance

Avant de pouvoir modéliser la structure de corrélation entre les variables, il est nécessaire de s'assurer que celles-ci présentent une dépendance entre elles. Il est rigoureux de noter que dans le cas de l'indépendance entre les variables, la copule ajustée serait la copule indépendante. L'utilisation des outils présentés dans la partie 7.1.2 permet alors de vérifier que le montant de règlement d'un sinistre est corrélé avec sa durée de vie.

Premièrement, les trois mesures de corrélation globale présentées dans la partie 7.1.2.1 sont calculées. Ces trois indicateurs sont significativement différents de zéro, ce qui rejette l'hypothèse selon laquelle les variables sont indépendantes.

	Coefficient de corrélation entre règlement et durée de vie
Pearson	0,28
Kendall	0,28
Spearman	0,40

TABLE 7.9 – Calcul des différents coefficients de corrélation

Dans un second temps, l'analyse des graphiques présentés dans la partie 7.1.2.2 peut permettre de confirmer la dépendance. Elle est en effet confirmée par ces deux graphiques. Le chi-plot montre une majorité de points en dehors de l'intervalle de confiance, et la courbe du Kendall-plot est différente de la bissectrice (au-dessus, ce qui indique une corrélation positive). Une fois cette corrélation exposée, il ne reste plus qu'à choisir la meilleure structure de dépendance pour les données.

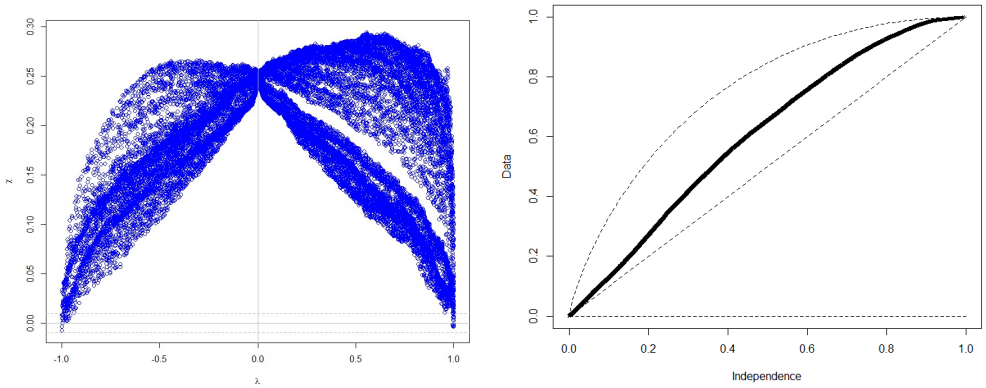

FIGURE 7.27 – Chi-plot et Kendall-plot entre règlement et durée de vie des sinistres

7.4.3.2 Estimation des paramètres et choix de la meilleure copule

Pour commencer, il est nécessaire de choisir quelle méthode d'ajustement utiliser parmi celles présentées dans la partie 7.1.3. En raison de la qualité discutable des ajustements des lois marginales relatives aux montants de règlement et durées de vie des sinistres, la procédure utilisée sera la méthode CML, qui permet l'estimation des paramètres de la copule considérée indépendamment de ceux des marges.

Nous considérons dans la suite six familles de copules différentes, à savoir les copules gaussiennes, de Student, de Gumbel, de Frank, de Joe, et de survie de Clayton. Ces copules appartiennent à la famille des copules elliptiques ou archimédiennes, et sont détaillées en annexe.

Estimation des paramètres

Dans la suite de l'étude, quelques notations sont utilisées :

- Λ : l'ensemble des copules considérées.
- $C_G, C_S, C_{Gu}, C_F, C_J, C_{SC} \in \Lambda$: Respectivement les familles de copules gaussiennes, de Student, de Gumbel, de Frank, de Joe et de survie de Clayton.
- $\Theta_G, \Theta_S, \Theta_{Gu}, \Theta_F, \Theta_J, \Theta_{SC} \in \Lambda$: les domaines de définition des paramètres des copules respectives.
- $\theta_G, \theta_S, \theta_{Gu}, \theta_F, \theta_J, \theta_{SC}$: les paramètres des copules respectives.
- R^{eg} : la variable aléatoire représentant le montant de règlement des sinistres, et $(r_i^{eg})_{1\leq i\leq n}$ les observations.
- D : la variable aléatoire représentant la durée de vie des sinistres, et $(d_i)_{1\leq i\leq n}$ les observations.
- $F_n^{Reg}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_i^{eg} \leq x)$: la fonction de répartition empirique de R^{eg}.
- $F_n^{D}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(d_i \leq x)$: la fonction de répartition empirique de D.
- C_n : la copule empirique des montants de règlement et de la durée de vie.

Soit $C \in \{C_G, C_S, C_{Gu}, C_F, C_J, C_{SC}\}$ et c la densité de copule associée. Soit Θ le domaine de définition du ou des paramètres correspondants. Le paramètre optimal θ^* est obtenu en résolvant :

$$\theta^* = \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{i=1}^{n} c\left(F_n^{Reg}\left(r_i^{eg}\right), F_n^{D}\left(d_i\right); \theta\right)$$

$$\iff \theta^* = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln\left(c\left(F_n^{Reg}\left(r_i^{eg}\right), F_n^{D}\left(d_i\right); \theta\right)\right)$$

Une fois les optimisations effectuées, on définit $\Lambda^* = \{C_G^*, C_S^*, C_{Gu}^*, C_F^*, C_J^*, C_{SC}^*\}$ l'ensemble des copules correspondant aux paramètres optimisés.

Famille	Paramètres
Gaussienne	0,41
Student	$\rho = 0,42 \quad \nu = 6,7$
Gumbel	1,39
Frank	2,77
Joe	1,58
Survie de Clayton	0,71

TABLE 7.10 – Paramètres des différentes copules ajustées entre le montant de règlement et la durée de vie des sinistres

Choix de la meilleure copule

La méthode de choix de la meilleure copule utilisée dans ce mémoire est la même quelle que soit la dimension du modèle. Dans cette partie, la méthode concerne un modèle de dimension 2.

Une fois les paramètres de chaque copule estimés, il ne reste plus qu'à déterminer la meilleure parmi les six proposées. Dans le cadre de cette étude, on décide d'utiliser la minimisation de la moyenne des écarts quadratiques entre la copule empirique et théorique (ou par équivalence la racine carrée de cette valeur). Autrement dit, on résout :

$$C^* = \underset{C \in \Lambda^*}{\operatorname{argmin}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(C_n(r_i^{eg}, d_i) - C(r_i^{eg}, d_i)\right)^2}$$

Famille	Racine de la moyenne des écarts quadratiques
Gaussienne	0,0124
Student	0,0120
Gumbel	0,0090
Frank	0,0106
Joe	0,01330
Survie de Clayton	0,0107

TABLE 7.11 – Critère de choix de la meilleure copule pour la corrélation règlement/durée de vie

La structure de dépendance choisie est la copule de Gumbel.

7.4.3.3 Comparaison des structures de corrélation empirique et théorique

Une fois la structure de dépendance théorique caractérisée, les analyses graphiques permettent de confirmer la bonne adéquation de cette dernière avec les données empiriques. Pour cela, on

simule des observations de la copule de Gumbel ajustée.

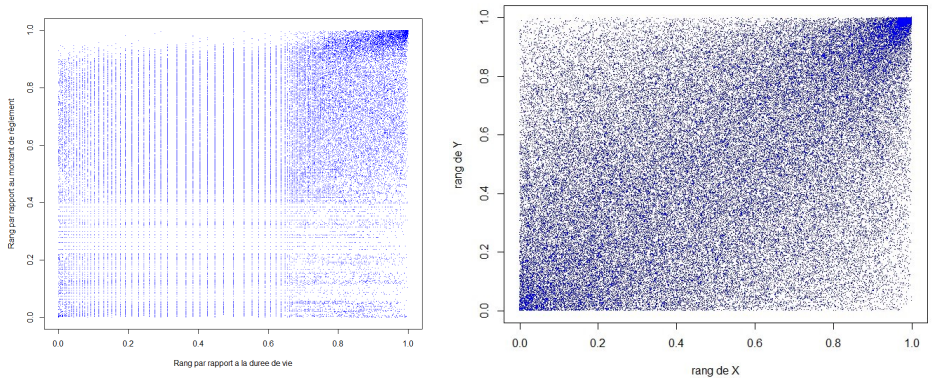

FIGURE 7.28 – Comparaison du rank-rank plot empirique (gauche) et théorique (droite) pour la corrélation règlement/durée de vie

La comparaison du rank-rank plot est rendue délicate par la présence de points alignés verticalement ou horizontalement sur le graphique. Les lignes verticales s'expliquent par le fait que la durée de vie est une variable discrète exprimée en jours. En effet, il peut logiquement arriver que la durée de vie de plusieurs sinistres soit exactement la même, et par conséquent que les valeurs de leurs rangs normalisés soient identiques. Ce phénomène est accentué par le conditionnement à un montant de recours nul, puisque les sinistres sans recours peuvent souvent se clôturer grâce à des démarches plus légères que lorsque des recours sont à exercer. Les observations se concentrent alors sur des petites valeurs en matière de durée de vie. Les lignes horizontales, présentes dans une moindre mesure, n'ont pas la même cause puisque le montant de règlement d'un sinistre peut être assimilé à une variable aléatoire continue. Effectivement, l'échelle et la précision de la variable (au centime près) permettent d'émettre l'hypothèse de continuité de celle-ci. Cependant, il arrive que des sinistres se clôturent sans aucune indemnisation après expertise. Les frais d'expertise restent toutefois à la charge de l'assureur et correspondent la plupart du temps à des montants forfaitaires. Ces valeurs étant constatées plusieurs fois dans la base, cela provoque donc l'alignement horizontal des points correspondants sur le rank-rank plot.

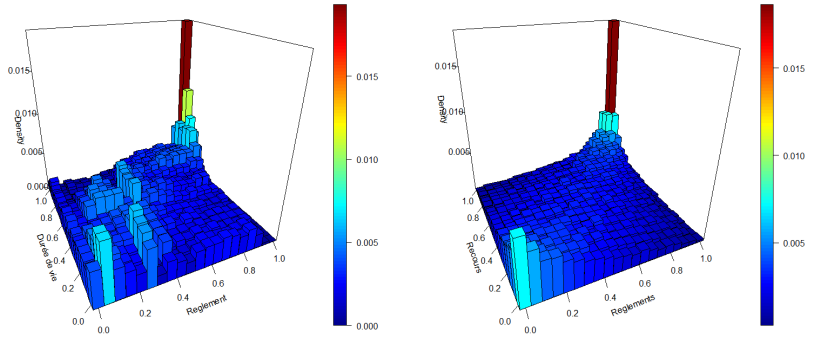

FIGURE 7.29 – Comparaison des densités empirique et théorique des copules ajustées pour la corrélation règlement/durée de vie

L'effet précédemment décrit est effacé par la représentation de la densité par histogramme 3D, le carré $[0,1]^2$ étant divisé en surface de $0,05 \times 0,05$. La comparaison des histogrammes empirique et théorique indique une bonne adéquation de la structure de corrélation.

7.5 Modélisation conditionnée à la présence de recours

Dans le cas où le montant de recours est supposé non nul, il n'y a plus seulement deux mais trois variables restant à expliquer. A la manière de la procédure utilisée dans le cas où le recours est nul, le but est ici de connaître les distributions marginales des trois variables d'intérêt, mais aussi d'adapter une structure de corrélation permettant d'expliquer la dépendance entre ces trois variables. Le cheminement utilisé est alors sensiblement le même que précédemment.

7.5.1 Ajustement des lois marginales

La première étape d'étude des distributions marginales est identique à celle effectuée dans la section précédente. Par conséquent, afin d'éviter une certaine redondance, seuls les principaux résultats des ajustements seront détaillés. Pour rappel, la méthodologie standard mise en place pour l'étude des lois marginales est la suivante :

- Recherche d'un seuil de valeurs « larges » pour la variable ;
- Ajustement des valeurs supérieures à ce seuil à une GPD ;
- Ajustement des valeurs inférieures à ce seuil à une loi usuelle tronquée.

Les résultats obtenus lors de la procédure d'ajustement des montants de règlement et de recours sont résumés dans les tableaux suivants :

	Règlement	Recours
Seuil	150000	150000
Loi des valeurs inférieures au seuil	Burr tronquée	Burr tronquée
Paramètres	$m = 3578 \quad s = 4,97 \quad f = 0,20$	$m = 2242 \quad s = 1,92 \quad f = 0,45$
Loi des valeurs supérieures au seuil	GPD	GPD
Paramètre de forme	0,4414	0,4730

TABLE 7.12 – Résultats des ajustements des variables règlement et recours

Pour la variable de la durée de vie, il a été décidé de ne pas mettre en place de valeur seuil, compte tenu de l'excellent ajustement des données à la loi Log-normale. L'analyse graphique, non intégrée dans ce mémoire, indique en effet que la modélisation semble très adaptée aux données.

	Durée de vie
Seuil	Aucun
Loi	Log-normale
Paramètres	$\mu = 6,13 \quad \sigma = 0,76$

TABLE 7.13 – Résultats des ajustements de la durée de vie

7.5.2 Ajustement de la structure de corrélation règlement-recours-durée de vie

Une fois que chaque distribution marginale est supposée connue, il reste à déterminer la structure de dépendance entre les trois variables lorsque le montant de recours est strictement positif. Comme énoncé au préalable, les solutions existantes concernant les structures de corrélation de dimension supérieure à 2 sont moins diversifiées puisque le sujet est plus difficile à appréhender.

7.5.2.1 Corrélation des variables deux à deux

Avant de pouvoir identifier et estimer une structure de corrélation entre les variables, il est nécessaire de s'assurer que celles-ci présentent une dépendance entre elles. Pour illustrer cette dépendance de chaque couple, les trois différentes mesures de corrélation présentées auparavant sont calculées.

	Règlement-Durée de vie	Règlement-Recours	Recours-Durée de vie
Pearson	0,29	0,99	0,27
Kendall	0,25	0,90	0,21
Spearman	0,36	0,97	0,32

TABLE 7.14 – Coefficients de corrélation entre les variables deux à deux

Les trois indicateurs montrent une dépendance significative entre les variables deux à deux. L'étude de la structure de dépendance des variables deux à deux, bien que n'étant pas l'objectif final de cette partie, est tout de même intéressante pour représenter une partie de la corrélation. La partie 7.4.3.2, estimant les paramètres et choisissant la meilleure copule, est alors réitérée pour les trois possibles couples de variables.

Au vu des représentations par histogramme 3D des densités empiriques, les dépendances des couples règlement/durée de vie et recours/durée de vie semblent relativement similaires. En revanche, le couple règlement/recours semble plus fortement corrélé. De nouveau, la copule de Gumbel est celle qui modélise le mieux la corrélation, et ce pour les trois couples.

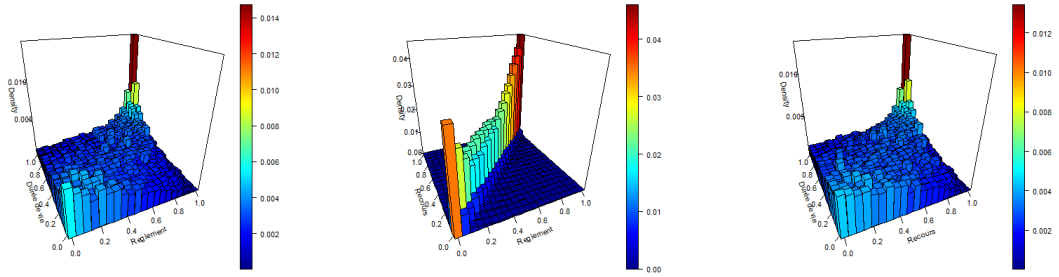

FIGURE 7.30 – Densités des copules empiriques entre les variables deux à deux

	Famille de copules	Paramètre
Règlement-Durée de vie	Gumbel	1,31
Règlement-Recours	Gumbel	9,6
Recours-Durée de vie	Gumbel	1,26

TABLE 7.15 – Copules bi-variées ajustées pour chaque couple de variables

Construire une structure de corrélation des variables deux à deux permet de représenter une partie de la dépendance, mais ne permet pas de l'expliquer en totalité. Une structure de corrélation intégrant les trois variables doit alors être envisagée. Il est alors nécessaire d'introduire les copules archimédiennes hiérarchiques, un type de copules qui présente des avantages en dimensions supérieures à 2.

7.5.2.2 Les copules archimédiennes

Les copules archimédiennes sont, avec les copules elliptiques, les deux principales catégories de copules utilisées. Introduites par Genest et Mackay (1986), elles regroupent plusieurs familles de copules dont les plus usuelles sont détaillées en annexe.

Définition 7.5.1. *Soit* $(u_1, ..., , u_n) \in [0,1]^n$. *La copule archimédienne* C *de générateur* $\phi : [0,1] \to [0, +\infty]$ *est définie par l'égalité :*

$$C(u_1, \ldots, u_n) = \phi^{-1}(\phi(u_1) + \ldots + \phi(u_n))$$

Le générateur ϕ *doit être choisi de classe* $\mathcal{C}^2$ *de sorte que* $\phi(1) = 0, \phi'(u) \leq 0$ *et* $\phi''(u) > 0$.

Définition 7.5.2. *Le générateur* ϕ *d'une copule archimédienne est dit strict si* $\phi(0) = +\infty$. *Dans ce cas, on parle de* ***copule archimédienne stricte****.*

Malgré leur facilité de construction et leurs propriétés intéressantes, les copules archimédiennes se heurtent à une faiblesse majeure en grande dimension. Leur paramètre (de dimension 1) est censé

décrire précisément la structure de dépendance entre toutes les variables, quelle que soit la dimension considérée. Or, l'une des propriétés des copules archimédiennes est qu'elles sont échangeables, c'est-à-dire que :

$$C(u_1, \ldots, u_n) = C\left(u_{\pi(1)}, \ldots, u_{\pi(n)}\right)$$

pour toute permutation π de $(1, ..., n)$.

Ceci implique alors que la structure de dépendance est la même pour toutes les variables prises deux à deux. Le manque de flexibilité des copules archimédiennes nous pousse alors à chercher un autre type de structure, plus adapté à nos variables, qui n'ont clairement pas le même degré de dépendance entre elles. D'un côté, le règlement est très fortement corrélé au montant des recours. De l'autre, la durée de vie est faiblement corrélée aux deux autres variables. Il faut alors trouver une structure de corrélation permettant de prendre en compte cette particularité.

7.5.2.3 Les copules archimédiennes hiérarchiques

Les copules archimédiennes hiérarchiques (CAH) ou imbriquées constituent une solution pour assouplir la structure de dépendance des copules archimédiennes. Proposées par Joe (2001), elles permettent d'obtenir des degrés de corrélation différents entre les variables prises deux à deux.

Une CAH est construite en imbriquant plusieurs copules archimédiennes les unes dans les autres. On présentera dans cette partie des exemples de CAH de dimension 4. Une généralisation s'effectue facilement dans le cas de dimensions supérieures. Une CAH peut être dite totalement imbriquée, c'est-à-dire composée de $n-1$ niveaux pour n variables aléatoires. On peut par exemple avoir la structure suivante :

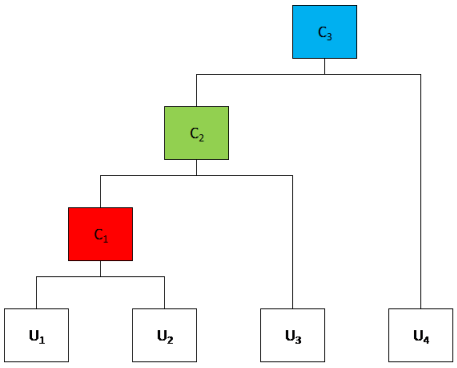

FIGURE 7.31 – Copule archimédienne hiérarchique totalement imbriquée

La copule C est alors donnée par la formule :

$$C(u_1, u_2, u_3, u_4) = C_3(C_2(C_1(u_1, u_2), u_3), u_4)$$

Dans les cas où le nombre de niveaux est inférieur à $n-1$ pour n variables aléatoires, la CAH est dites partiellement imbriquée. Le type de structure ci-dessous fait alors partie de cette catégorie.

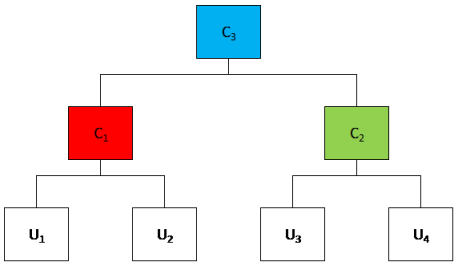

FIGURE 7.32 – Copule archimédienne hiérarchique partiellement imbriquée

Cette fois, la copule C est donnée par :

$$C\left(u_1, u_2, u_3, u_4\right) = C_3\left(C_1\left(u_1, u_2\right), C_2\left(u_3, u_4\right)\right)$$

Ce type de structure permet alors de proposer une solution au problème initial lorsque les variables n'admettent pas les mêmes structures de corrélation deux à deux. De plus, la construction des CAH est intuitive et plutôt simple d'interprétation. Cependant, elle ne permet pas d'apporter une solution à toutes les difficultés que l'on peut rencontrer lors de l'étude des corrélations en grande dimension. D'abord, la souplesse apportée aux copules archimédiennes par la hiérarchisation n'est pas nécessairement suffisante. En effet, dans le premier exemple, les paires de variables (U_1, U_3) et (U_2, U_3) ont obligatoirement la même structure de corrélation. Ensuite, les formules de C définies précédemment ne sont pas assurément des copules. Les conditions nécessaires pour les générateurs sont inconnues et les conditions suffisantes ne sont pas toujours faciles à vérifier.

Exemples : Soient $(\phi_i)_{1\leq i\leq n}$ les générateurs de copules de Gumbel de paramètre θ_i. Une condition suffisante pour que C soit une copule est que la suite $(\theta_i)_{1\leq i\leq n}$ soit décroissante avec $\theta_i > 1$ pour tout i.

De la même manière, si $(\phi_i)_{1\leq i\leq n}$ sont des générateurs de copules de Clayton de paramètre θ_i. Une condition suffisante pour que C soit une copule est que la suite $(\theta_i)_{1\leq i\leq n}$ soit décroissante avec $\theta_i > 0$ pour tout i.

7.5.2.4 Estimation des paramètres et choix de la meilleure CAH

Dans notre situation, la copule archimédienne hiérarchique que l'on cherche à construire est de dimension 3. La structure est alors forcément une CAH totalement imbriquée (la copule se résumerait sinon à une copule archimédienne classique). Compte tenu des conditions à vérifier pour que la hiérarchie construite engendre bien une fonction copule, on se limitera à des CAH composées d'une unique famille de copules. Les paramètres de la CAH seront alors estimés pour des structures composées de copules de Gumbel, Clayton, Frank, Joe et Ali-Mikhail-Haq (AMH).

Les conditions à vérifier pour que la fonction construite soit une copule sont :

- $\theta_i > 1 \quad \forall i = 1, ..., n$ pour les copules de Gumbel et Joe ;

- $\theta_i > 0 \quad \forall i = 1, ..., n$ pour les copules de Clayton et Frank ;
- $0 < \theta_i < 1 \quad \forall i = 1, ..., n$ pour les copules AMH ;
- Les θ_i sont décroissants en remontant dans la hiérarchie.

Par la suite, les notations suivantes seront utilisées :

- Λ : l'ensemble des CAH considérées ;
- $C_{Gu}, C_F, C_J, C_C, C_{AMH} \in \Lambda$: respectivement les familles de CAH de Gumbel, Frank, Joe, Clayton et AMH ;
- $\Theta_{Gu}, \Theta_F, \Theta_J, \Theta_C, \Theta_{AMH} \in \Lambda$: les domaines de définition des paramètres des copules respectives ;
- $\theta_{Gu}, \theta_F, \theta_J, \theta_C, \theta_{AMH}$: les paramètres des copules respectives ;
- R^{eg} : la variable aléatoire représentant le montant de règlement des sinistres, et $(r_i^{eg})_{1 \leq i \leq n}$ les observations ;
- R^{ec} : la variable aléatoire représentant le montant de recours des sinistres, et $(r_i^{ec})_{1 \leq i \leq n}$ les observations ;
- D : la variable aléatoire représentant la durée de vie des sinistres, et $(d_i)_{1 \leq i \leq n}$ les observations ;
- $F_n^{Reg}(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \mathbb{1}(r_i^{eg} \leq x)$: la fonction de répartition empirique de R^{eg} ;
- $F_n^{Rec}(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \mathbb{1}(r_i^{ec} \leq x)$: la fonction de répartition empirique de R^{ec} ;
- $F_n^{D}(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \mathbb{1}(d_i \leq x)$: la fonction de répartition empirique de D ;
- C_n : la copule empirique des montants de règlement, de recours et de la durée de vie.

Soit $C \in \{C_{Gu}, C_F, C_J, C_C, C_{AMH}\}$ et c la densité de copule associée. Soit Θ le domaine de définition des paramètres correspondants. Les paramètres optimaux θ^* sont obtenus grâce à la méthode CML en résolvant :

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} c\left(F_n^{Reg}\left(r_i^{eg}\right), F_n^{Rec}\left(r_i^{ec}\right), F_n^{D}\left(d_i\right); \theta\right)$$

$$\iff \theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln\left(c\left(F_n^{Reg}\left(r_i^{eg}\right), F_n^{Rec}\left(r_i^{ec}\right), F_n^{D}\left(d_i\right); \theta\right)\right)$$

Une fois les optimisations effectuées, on définit $\Lambda^* = \{C_{Gu}^*, C_F^*, C_J^*, C_C^*, C_{AMH}^*\}$ l'ensemble des copules correspondant aux paramètres optimisés.

De nouveau, le choix de la meilleure copule se faire par minimisation de la racine carrée de la moyenne des écarts quadratiques entre la copule empirique et la copule théorique. Autrement dit,

on résout :

$$C^* = \underset{C \in \Lambda^*}{\operatorname{argmin}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(C_n(F_n^{Reg}(r_i^{eg}), F_n^{Rec}(r_i^{ec}), F_n^{D}(d_i)) - C(F_n^{Reg}(r_i^{eg}), F_n^{Rec}(r_i^{ec}), F_n^{D}(d_i)) \right)^2}$$

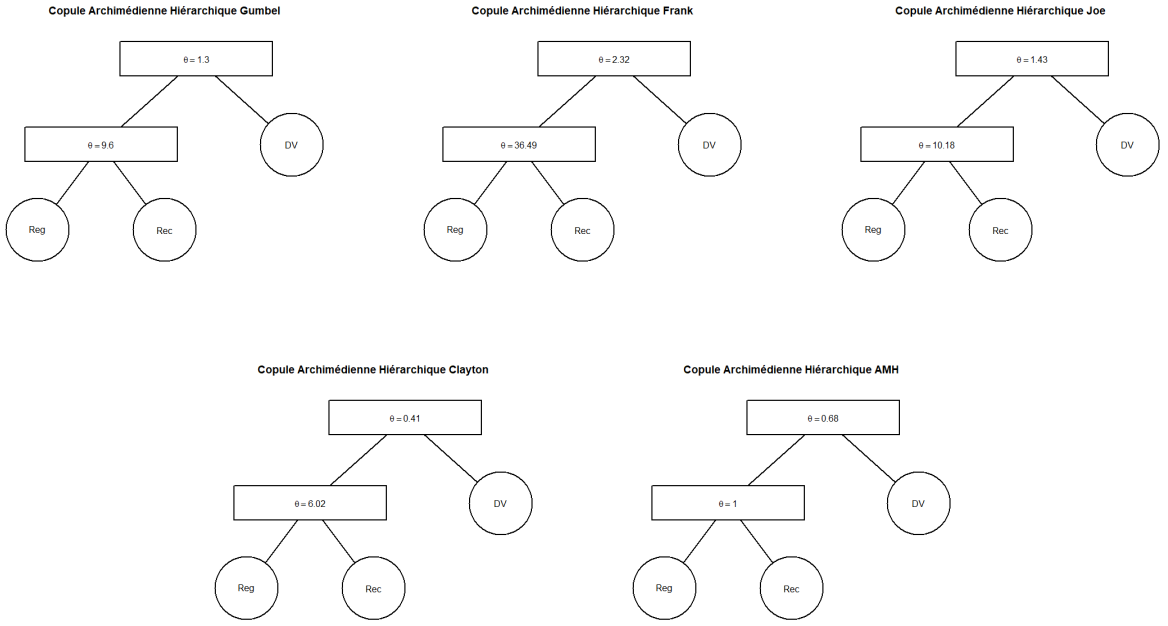

FIGURE 7.33 – Différentes CAH ajustées

Sans surprise compte tenu du fait que les meilleures copules bi-variées sont les copules de Gumbel pour chaque couple de variables, la meilleure copule au sens du critère défini est la CAH de Gumbel.

7.5.2.5 Représentation graphique de l'ajustement

La représentation graphique est plus difficile à appréhender lorsque la dimension est supérieure à 2. En dimension 3, le diagramme des rangs permet tout de même d'illustrer la copule théorique simulée et la copule empirique, et donc de présumer de la qualité d'ajustement. Ces diagrammes peuvent être représentés avec plusieurs perspectives du nuage de points 3D. L'ajustement semble alors relativement bon.

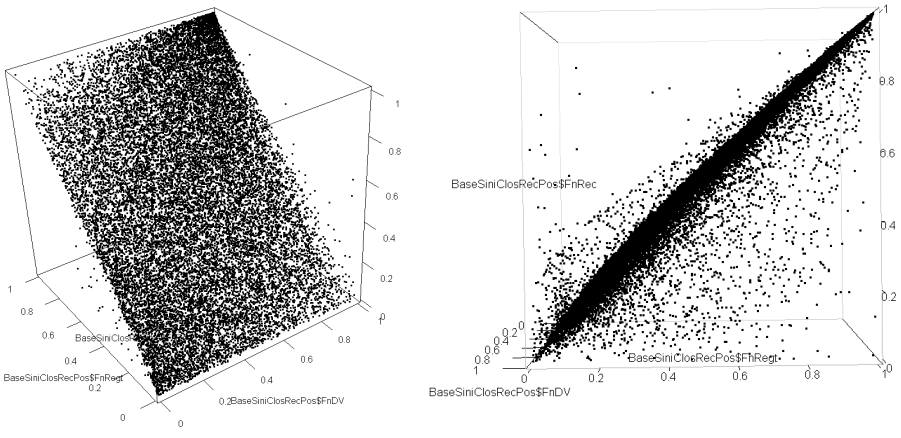

FIGURE 7.34 – Diagramme des rangs de la copule empirique

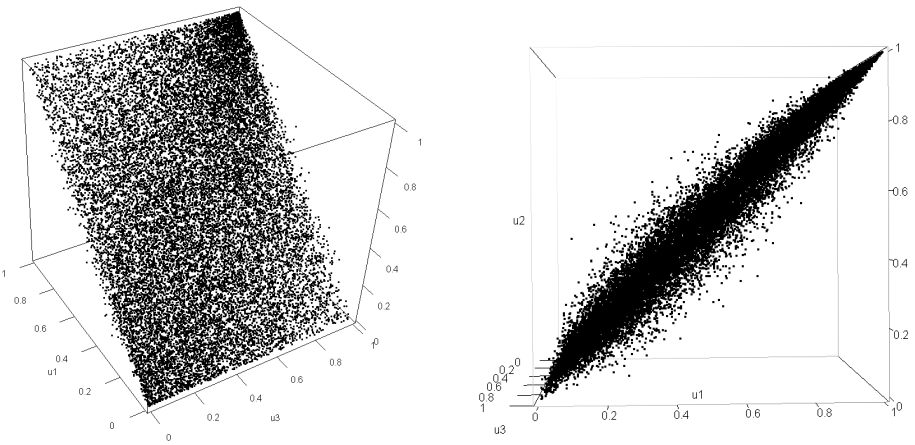

FIGURE 7.35 – Diagramme des rangs de la copule de Gumbel théorique

Chapitre 8

Simulations et résultats

L'ensemble des modèles décrits dans ce mémoire font appel à des simulations de scénarios. Ces dernières emploient alors des méthodes spécifiques aux outils actuariels utilisés.

8.1 Méthodes de simulation

8.1.1 Simulation des distributions ajustées

Le modèle de coût fait appel à la simulation de distributions théoriques ajustées. Les logiciels matriciels comme R proposent généralement des *packages* permettant de simuler rapidement des observations pour une loi usuelle donnée. La structure de dépendance par copule se construisant sur des lois uniformes sur $[0,1]$, il est alors nécessaire de pouvoir traduire des observations uniformes en des observations correspondant aux distributions ajustées. On rappelle alors la définition de l'inverse généralisé d'une fonction de répartition :

Définition 8.1.1. *Soit X une variable aléatoire de fonction de répartition F. Alors par définition,*

$$F(x) = \mathbb{P}(X \leq x)$$

F est continue à droite et admet des limites à gauche. On définit ***l'inverse généralisée*** *de F par :*

$$F^{-1}(y) = \inf\{x \mid F(x) \geq y\}$$

La notion d'inverse généralisée est utilisée dans la propriété fondamentale suivante :

Propriété 8.1.1. *La variable $U = F(X)$ est distribuée selon une loi uniforme sur $[0,1]$.*

En effet, on a $\mathbb{P}(U \leq u) = \mathbb{P}(F(X) \leq u) = \mathbb{P}(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$.

On peut alors écrire la relation $F^{-1}(U) = X$. Il suffit donc d'appliquer cette fonction F^{-1} à une variable aléatoire de loi $\mathcal{U}(0,1)$ pour simuler la loi initiale désirée. Cette technique est la méthode de simulation par inversion de la fonction de répartition.

L'inverse généralisée d'une fonction de répartition n'admet pas toujours de formule explicite, en particulier lorsque la loi ajustée est un mélange de plusieurs lois. Il est alors possible d'utiliser un algorithme de résolution numérique (du type Newton-Raphson) pour calculer les valeurs de cette fonction.

8.1.2 Simulation des copules

Les structures de corrélation construites entre les variables doivent également pouvoir être simulées, afin de respecter l'ajustement. Il existe alors des méthodes pour simuler les observations de lois uniformes relatives aux structures de dépendance par copule. Une fois ces distributions uniformes spécifiées, il suffit ensuite d'utiliser la propriété décrite dans la section précédente en inversant la fonction de répartition de la loi modélisée pour obtenir les variables aléatoires de distributions désirées.

En d'autres termes, si l'on veut simuler une observation x d'un vecteur aléatoire $X = (X_1, ..., X_n)$ de distributions marginales $F_1, ..., F_n$ et de copule C, la démarche est la suivante :

- On simule $u = (u_1, ..., u_n)$ de copule C ;
- On applique l'inverse généralisée des fonctions de répartition marginales aux composantes de ce vecteur :

$$x = (F_1^{-1}(u_1), ..., F_n^{-1}(u_n))$$

La principale difficulté réside dans la simulation de la copule C. Il peut exister des méthodes pour simuler les lois uniformes de copules particulières. Une méthode est par exemple proposée pour les copules archimédiennes en dimension 2. Si l'on considère une copule archimédienne C de générateur ϕ, la démarche de simulation est la suivante :

- Simuler R et S de loi $\mathcal{U}(0,1)$ indépendantes ;
- Trouver T, tel que l'équation $S = T - \dfrac{\phi(T)}{\phi'(T)}$ soit vérifiée ;
- Poser $U = \phi^{-1}(R \times \phi(T))$ et $V = \phi^{-1}((1-R) \times \phi(T))$.

U et V suivent alors des lois $\mathcal{U}(0,1)$ liées par la copule C.

Pour les copules archimédiennes hiérarchiques totalement imbriquées, utilisées dans le cadre de ce mémoire, l'algorithme général utilisé est le suivant :

On note $C_{i|J}(. \mid u_j, j \in J)$ avec J un sous-ensemble de $\{1, ..., n\}$ et $i \in \{1, ..., n\}$ la fonction de répartition de U_i, conditionnellement aux valeurs prises par les variables aléatoires U_j, $j \in J$.

- On détermine u_1 de loi $\mathcal{U}(0,1)$. Dans notre cas, cette valeur est déterminée en appliquant la fonction de répartition théorique de la variable à l'observation simulée des montants de règlement.
- On simule une réalisation de u_2 par inversion de $C_{2|1}(. \mid u_1)$;
- ...
- On simule une réalisation u_n par inversion de $C_{n|1,...,n-1}(. \mid u_1, ..., u_{n-1})$.

Selon le type de copule considéré, cet algorithme n'est pas forcément le plus efficace. Il se révèle raisonnablement efficace lorsque les fonctions $C_{k|1,...,k-1}(. \mid u_1, ..., u_{k-1})$ ont des expressions explicites. Dans le cadre de ce mémoire, ces fonctions n'admettent pas d'expression explicite, ce qui oblige à passer par des algorithmes d'optimisation numérique qui nécessitent un temps de calcul considérable.

8.1.3 Simulation des processus de Hawkes

L'intensité d'un processus de Hawkes étant variable dans le temps et auto-excitée, il n'est pas simple de simuler une trajectoire issue de ce processus. Ogata (1981) propose un algorithme qui génère les temps de survenance pour un processus de Hawkes simple sur un intervalle $[0, T]$, sous condition de fournir l'intensité de base λ_0 ainsi que les paramètres α et β.

Initialisation :

- Renseigner les valeurs de λ_0, α, β et T ;
- λ^* prend la valeur de λ_0 ;
- Simuler u de loi $\mathcal{E}(\lambda^*)$;
- Si $u \leq T$ alors $s = u$, $t_1 = s$, i prend la valeur 1, et λ^* prend la valeur de l'intensité en s conditionnellement à l'instant de survenance t_1 ; sinon l'algorithme se termine.

L'étape d'initialisation crée le premier temps de survenance (s'il est inférieur à T) qui est une réalisation d'une loi exponentielle de paramètre λ_0. La routine générale permet ensuite de simuler les autres instants de survenance jusqu'au temps T, en utilisant un test de rejet.

Routine générale : Tant que $s \leq T$:

- Simuler u de loi $\mathcal{E}(\lambda^*)$;
- s prend la valeur $s + u$;
- Si $s \leq T$:
  1. Simuler D de loi $\mathcal{U}(0,1)$;
  2. Si $D \leq \dfrac{\lambda(s \mid t_1, ..., t_i)}{\lambda^*}$ alors $t_{i+1} = s$, λ^* prend la valeur de l'intensité en s conditionnellement aux instants de survenance $t_1, ..., t_{i+1}$, et i prend la valeur $i + 1$;
  3. Sinon, λ^* prend la valeur de l'intensité en s conditionnellement aux instants de survenance $t_1, ..., t_i$.

Lorsque l'objectif est de projeter une trajectoire du processus déjà en cours, comme ce peut être le cas pour la majorité des contrats, il suffit de renseigner l'intensité théorique au temps considéré ainsi que les instants de survenance connus et d'appliquer la routine générale.

8.2 Schéma récapitulatif de la démarche

Le modèle proposé pour la projection du portefeuille de la garantie Dommages Ouvrage utilise des simulations pour déterminer la sinistralité inconnue qui interviendra d'ici la date de fin de couverture. La simulation de la sinistralité future d'un contrat peut alors être schématisée.

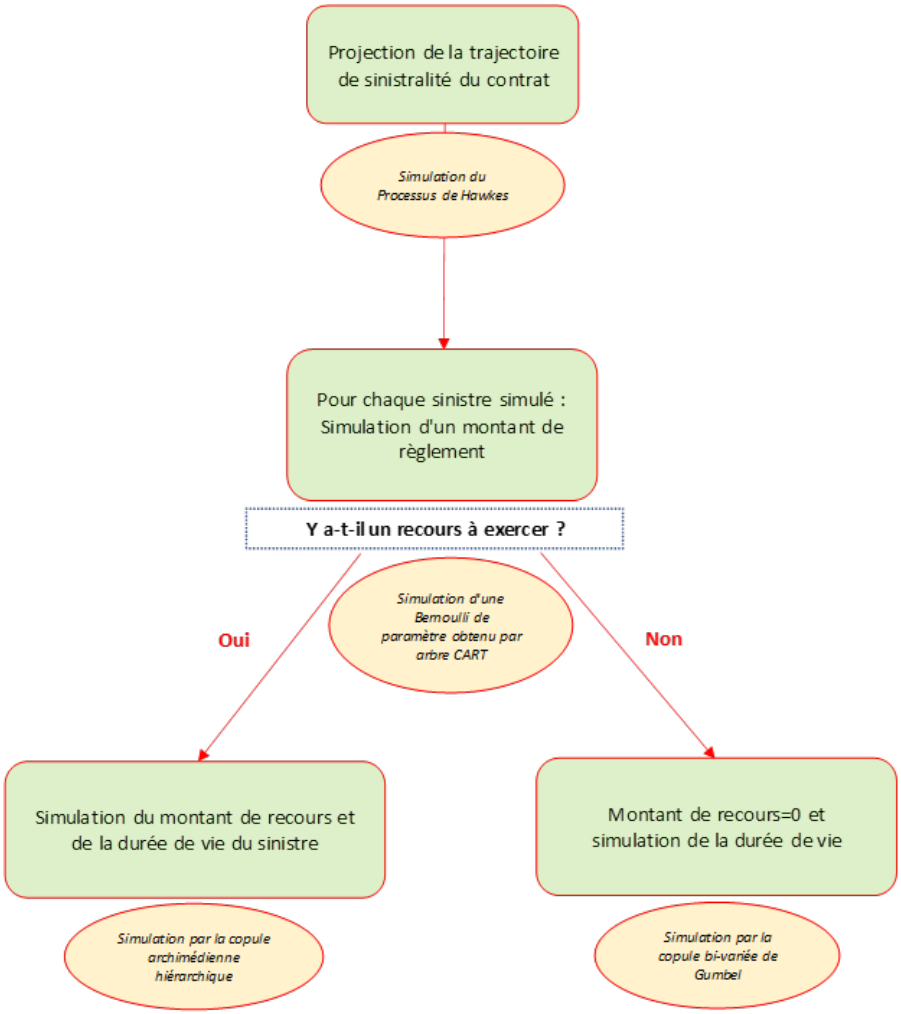

FIGURE 8.1 – Schéma récapitulatif de la démarche de projection des sinistres inconnus

Cette simulation est effectuée pour chaque contrat pour projeter l'intégralité du portefeuille. Plusieurs scénarios de projection sont effectués pour pouvoir converger en moyenne vers la sinis-

tralité ultime du contrat.

8.3 Temps de calcul

Pour effectuer la projection, les programmes dédiés utilisent les méthodes de simulation décrites précédemment. Il est alors intéressant d'étudier les temps de calcul que consacrent les programmes à chaque partie de la procédure.

8.3.1 Phase d'apprentissage

La phase d'apprentissage comprend l'ensemble des étapes de classification des contrats et de différentes estimations paramétriques des modèles. Elle requiert 3 minutes et 20 secondes de calcul pour être mise en œuvre.

8.3.2 Phase de simulation

Pour la phase de projection, la simulation de 1000 scénarios est assez consommatrice en temps de calcul, puisqu'elle nécessite 13 heures. Cette durée correspond à 45 secondes pour chaque scénario. Il faut savoir qu'un scénario projette environ 42000 contrats. La projection d'un contrat nécessite alors environ 0,001 seconde.

A l'intérieur de cette projection, pour un scénario :

- Il faut en moyenne 1,4 secondes pour projeter la fréquence de sinistralité inconnue du portefeuille, c'est-à-dire la trajectoire de 42000 contrats. Cela correspond à 60000 sinistres simulés en moyenne.
- Pour attribuer à chacun de ces sinistres un montant de règlement, de recours et une durée de vie, l'algorithme nécessite 40 secondes, soit 0,00067 seconde par sinistre.

8.4 Résultats

8.4.1 Résultat de la projection par DOC

Une fois l'ensemble des scénarios simulés, il est possible de visualiser la sinistralité projetée par DOC, et de la comparer avec celle constatée à l'heure actuelle. A cet effet, on considère le ratio sinistres à primes.

Le modèle proposé réussit à projeter la sinistralité, même pour les années pour lesquelles les contrats n'ont pas observé beaucoup de sinistres. En revanche, le S/P projeté pour les 3 à 4 dernières années est nettement plus faible que pour les années précédentes. L'hypothèse de sous-estimation

de la sinistralité peut être évoquée pour ces mêmes années, qui pourrait être causée par la très faible sinistralité connue en comparaison avec les processus de Hawkes modélisés.

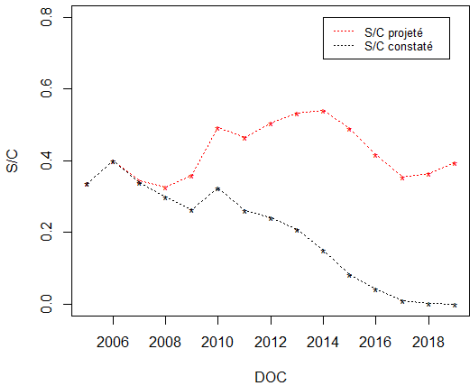

FIGURE 8.2 – Confrontation des S/P projetés et constatés

8.4.2 Convergence de la sinistralité

Etant donné que le modèle de projection fonctionne sur la simulation de scénarios, il est important de vérifier que la sinistralité projetée converge. L'étude de la convergence permet également de connaître le nombre de scénarios nécessaires pour obtenir une estimation robuste.

8.4.3 Convergence de la sinistralité par DOC

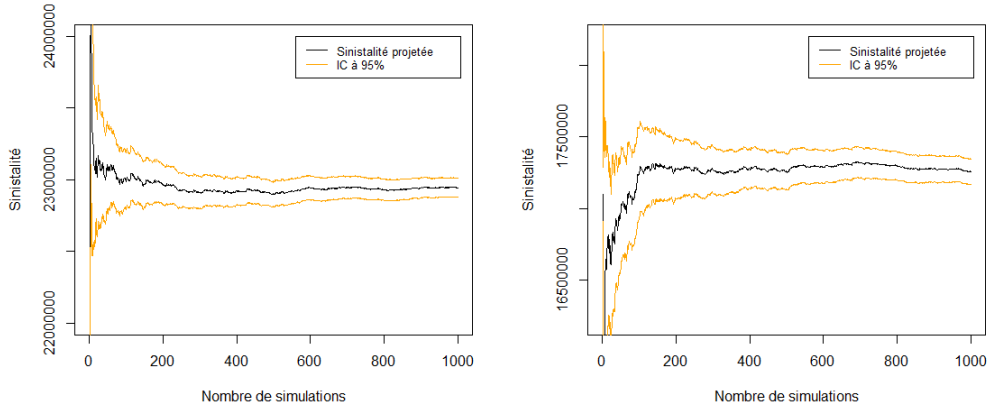

FIGURE 8.3 – Convergence de la sinistralité pour les DOC 2010 et 2017

La sinistralité agrégée au niveau de la DOC des contrats converge assez rapidement. Cette

vitesse de convergence est plus grande pour les DOC anciennes (2010), puisque peu de sinistres inconnus sont simulés, contrairement aux DOC plus récentes (2017). Pour les DOC les plus récentes, la simulation de 200 scénarios semble être suffisante.

8.4.4 Convergence de la sinistralité par contrat

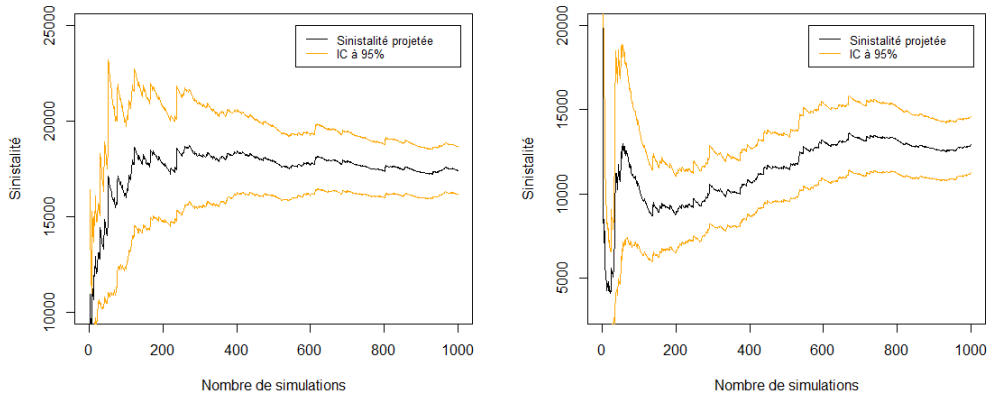

FIGURE 8.4 – Convergence de la sinistralité pour des contrats des DOC 2010 et 2017

Pour ce qui est de la convergence au niveau contrat, le nombre de simulation nécessaire est bien plus grand. Pour un contrat de 2010, la sinistralité semble converger autour de 400 simulations. Pour un contrat de 2017, l'estimateur semble se stabiliser autour de 700 simulations.

Chapitre 9

Conclusion

L'assurance construction est une branche de l'assurance non-vie qui comporte des garanties décennales, c'est-à-dire qui couvrent l'assuré pendant une période de dix ans. La sinistralité ultime d'un contrat, et par extension son résultat, est alors connue plus d'une dizaine d'années après sa souscription. Il convient donc pour l'assureur de pouvoir évaluer le plus rapidement possible la sinistralité future des contrats qu'il a en portefeuille, dans le but de réagir rapidement en cas de dérive de la sinistralité, et d'adapter les nouvelles souscriptions en conséquence le cas échéant. Cependant, les méthodes macro couramment utilisées ne permettent pas de prendre en compte les caractéristiques des chantiers spécifiques à chaque année de souscription. L'aspect temporel de la sinistralité est également important pour considérer l'inflation au plus juste, cette dernière ayant un impact non négligeable sur le résultat. Enfin, la garantie Dommages Ouvrage, produit d'assurance sur laquelle se focalise ce mémoire, fonctionne en lien avec la garantie responsabilité civile décennale pour l'exercice de recours. Ces derniers prennent une place primordiale dans le fonctionnement de la garantie et l'absence de considération de ceux-ci peut engendrer un manque de précision du modèle.

Pour répondre à ces problématiques concernant la projection ligne à ligne de la sinistralité de la garantie Dommages Ouvrage, ce mémoire s'est appuyé sur les précédents travaux effectués par Martin (2019), en proposant des alternatives aux modèles proposés. Dans un premier temps, une méthode alternative au modèle de fréquence a été proposée. Ce modèle se veut plus simple, en réduisant considérablement le nombre de paramètres nécessaire pour expliquer la fréquence de sinistres d'un contrat. L'utilisation de processus temporels permet de retirer la segmentation de la fréquence par DOC, tout en gardant l'information de l'instant de survenance des sinistres. L'antériorité de sinistralité des contrats est également préservée dans le modèle, puisque le processus temporel ajusté est un processus de Hawkes. Pour rappel, les processus de Hawkes sont des processus auto-excitants, c'est-à-dire que l'intensité du processus augmente au moment des sinistres, avant de décroitre exponentiellement jusqu'à l'observation d'un nouveau sinistre. Autrement dit, la probabilité d'observer un nouveau sinistre est plus importante si un ou plusieurs sinistres ont été observés sur ce contrat récemment. L'utilisation de ce type de processus est cohérente, dans le sens où la sinistralité observée donne de l'information sur la qualité des travaux, et donc sur la probabilité d'observer des futurs sinistres sur ce contrat. Dans un second temps, les montants des

règlements et des recours des sinistres ont été intégrés au modèle de coût comme des variables à part entière, remplaçant la variable de la charge nette de recours du sinistre initialement utilisée. La structure de dépendance ajustée par copule bivariée a alors évolué vers une structure à trois variables, afin de conserver la variable correspondant à la durée de vie des sinistres. Pour cela, les copules archimédiennes hiérarchiques nous ont permis d'obtenir un bon ajustement, grâce à leur flexibilité lorsque le degré de corrélation entre les variables est significativement différent.

Le modèle proposé, bien qu'il réponde à la problématique initiale, admet tout de même quelques faiblesses. Tout d'abord, le modèle de coût construit nécessite l'utilisation de plusieurs optimisations numériques. Ces optimisations sont principalement utilisées pour l'inversion de fonctions de répartition, pour simuler les copules ajustées ainsi que les lois marginales. Ces opérations informatiques sont complexes et nécessitent alors un temps de calcul non négligeable. Ensuite, le processus temporel ajusté pour expliquer la fréquence est un processus de Hawkes simple, c'est-à-dire que sa fonction d'intensité est stationnaire. Cette hypothèse n'est pas complètement vérifiée en pratique puisque l'intensité de la sinistralité observée sur la période définie est en réalité croissante sur les trois premières années, relativement constante ensuite et décroissante sur les trois dernières années. Des écarts pourraient alors être constatés entre les prédictions et les observations sur la projection complète des trajectoires de fréquence. Enfin, au contraire du modèle de fréquence en place, le processus de Hawkes utilisé pour expliquer la fréquence capte mal la déformation du portefeuille puisque l'apprentissage se fait uniquement sur des trajectoires entièrement observées.

L'ensemble de ces travaux fournissent une méthode de projection de la sinistralité de la garantie Dommages Ouvrage qui se veut performante. Cependant, les limites du modèle énoncées précédemment pourraient faire l'objet d'un complément d'étude. La simulation d'observations respectant la structure de dépendance ajustée pourrait être travaillée afin de limiter le recours aux algorithmes d'optimisation numérique, en explorant une méthode alternative à la simulation par inversion de la fonction de répartition. Un approfondissement du modèle de fréquence par processus de Hawkes est aussi envisageable. L'utilisation d'un processus de Hawkes avec une intensité de base variable permettrait d'obtenir un meilleur ajustement et diminuerait donc l'incertitude des projections. Cette intensité de base variable pourrait par exemple être supposée constante par morceaux (d'intervalle égal à un an). Il serait également intéressant que l'apprentissage prenne en compte les trajectoires n'étant pas encore arrivées à maturité, afin d'utiliser l'ensemble des informations disponibles pour les contrats de la base de données. Ces améliorations rendraient alors plus complexe la démarche d'estimation des paramètres, la fonction de vraisemblance associée au processus devenant plus difficilement exploitable. L'algorithme de simulation des trajectoires associées au processus devrait lui aussi être modifié en conséquence. Enfin, on peut penser que le modèle de fréquence est en capacité d'être transféré à la projection de la garantie responsabilité civile décennale. Une entreprise souscrivant une assurance pour plusieurs opérations de construction, la sinistralité d'un chantier pourrait alors également être expliquée par la sinistralité des autres chantiers, en utilisant des processus de Hawkes influencés par plusieurs types d'évènements.

Bibliographie

ARTICLE L241-1 DU CODE DES ASSURANCES : Codifié par le décret n° 76-666 du 16 juillet 1976 et modifié par la loi n°2015-990 du 6 août 2015

ARTICLE L242-1 DU CODE DES ASSURANCES : Codifié par le décret n° 76-666 du 16 juillet 1976 et modifié par l'ordonnance n°2005-658 du 8 juin 2005

ARTICLE L243-3 DU CODE DES ASSURANCES : Codifié par le décret n° 76-666 du 16 juillet 1976 et modifié par l'ordonnance n°2005-658 du 8 juin 2005

ARTICLE R343-7 DU CODE DES ASSURANCES : Créé par DÉCRET n°2015-513 du 7 mai 2015

ARTICLE 1792 DU CODE CIVIL : Codifié par la loi 1804-03-07 promulguée le 17 mars 1804 et modifié par Loi n°78-12 du 4 janvier 1978

ARTICLE 143-10 DU REGLEMENT N° 2015-11 DU 26 NOVEMBRE 2015 RELATIF AUX COMPTES ANNUELS DES ENTREPRISES D'ASSURANCE : Autorité des normes comptables. Homologué par arrêté du 28 décembre 2015 publié au Journal Officiel du 30 décembre 2015 et du 3 janvier 2016

ARTICLE 143-13 DU REGLEMENT N° 2015-11 DU 26 NOVEMBRE 2015 RELATIF AUX COMPTES ANNUELS DES ENTREPRISES D'ASSURANCE : Autorité des normes comptables. Homologué par arrêté du 28 décembre 2015 publié au Journal Officiel du 30 décembre 2015 et du 3 janvier 2016

ARTICLE 143-14 DU REGLEMENT N° 2015-11 DU 26 NOVEMBRE 2015 RELATIF AUX COMPTES ANNUELS DES ENTREPRISES D'ASSURANCE : Autorité des normes comptables. Homologué par arrêté du 28 décembre 2015 publié au Journal Officiel du 30 décembre 2015 et du 3 janvier 2016

C. BOURRY : Évaluation des provisions techniques et du capital économique associé au risque de réserve en assurance construction. Mémoire d'actuariat, ISFA, 2016.

B. DEPLANTE : Provisionnement et tarification en Dommages Ouvrage. Mémoire d'actuariat, ISFA, 2010.

A. CAMBON : Elaboration d'un modèle interne partiel concernant le risque de souscription non-vie pour tenir compte des spécificités d'une société spécialisée dans les branches longues. Mémoire

d'actuariat, ISUP, 2010.

M. MARTIN : Méthode de projection ligne à ligne de la sinistralité de la garantie RCDO en construction permettant de prendre en compte la déformation du portefeuille. Mémoire d'actuariat, ISFA, 2019.

P. BOUVIER : Application des copules à la finance de marché. Université du Québec à Montréal, 2010.

G. MAZO. Construction et estimation de copules en grande dimension. Université de Grenoble, 2014.

C. GENEST and J. BOIES (2003) : Testing dependence with Kendall Plots. The American Statistician, 44, 275-284.

N. I. FISHER and P. SWITZER (1985) : Chi-plot for assessing dependence. Biometrica, 72, 253-265.

C. GENEST and R. J. MACKAY (1986) : Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. La Revue Canadienne de Statistique. 14, 145-159.

H. Joe (2001) : Multivariate models and dependence concepts. Chapman and Hall/CRC.

Y. Ogata (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics, 30, 243-261.

Y. Ogata (1981). On Lewis' simulation method for point processes. IEEE Transactions on Information Theory, 27, 23-31.

M. HOFERT and D. PHAM : Densities of nested Archimedean copulas. Journal of Multivariate Analysis, 118, 37-52.

FEDERATION FRANCAISE DU BATIMENT : Vers une crise majeure de l'assurance construction ? 2018.

FEDERATION FRANCAISE DE L'ASSURANCE : L'assurance Française, données clés 2018. 2019.

Annexe A

Quelques copules usuelles

A.1 Les copules elliptiques

Une copule est dite elliptique si elle est la copule d'une loi elliptique. Pour rappel, une loi continue est dite elliptique de paramètre de position $\mu \in \mathbb{R}^d$ et de matrice de forme définie positive $\Sigma \in Mat_d(\mathbb{R})$ si sa densité f peut s'écrire, pour tout $x \in \mathbb{R}^d$:

$$f(x) = (\det \Sigma)^{-\frac{1}{2}} g\left((x-\mu)\Sigma^{-1}(x-\mu)^T\right)$$

où g est une fonction à valeurs positives vérifiant $\int_{\mathbb{R}^d} g(xx^T)dx = 1$.

A.1.1 La copule Gaussienne

On note ϕ la fonction de répartition de la loi $\mathcal{N}(0,1)$ univariée et Φ_Σ la fonction de répartition du vecteur gaussien centré réduit de matrice de corrélation $\Sigma \in Mat_d(\mathbb{R})$. La copule gaussienne s'écrit, pour tout $(u_1, ..., u_d) \in [0,1]^d$:

$$C(u_1, \ldots, u_d) = \Phi_\Sigma\left(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d)\right)$$

En dimension $d = 2$, en notant ρ le coefficient de corrélation linéaire entre les deux variables, elle admet les propriétés suivantes :

- La copule gaussienne n'admet pas de dépendance forte des extrêmes : $\lambda_L = \lambda_U = 0$;
- Le τ de Kendall est donné par la formule : $\tau = \frac{2}{\pi}\arcsin(\rho)$;
- Le ρ de Spearman est donné par la formule : $\rho_S = \frac{6}{\pi}\arcsin(\frac{\rho}{2})$.

C'est une copule très populaire puisqu'elle est très facile à simuler et propose une bonne flexibilité grâce à sa matrice de corrélation de dimension d. Cependant, elle ne présente pas de dépendance des extrêmes ce qui peut la rendre inadaptée pour représenter la corrélation des queues de distribution.

A.1.2 La copule de Student

La copule de Student est la copule sous-jacente à une distribution multi-variée de Student. Elle est construite à partir de la loi de Student centrée réduite multi-variée. On note t_ν la fonction de répartition de la loi de Student univariée à ν degrés de liberté et $t_{\Sigma,\nu}$ la distribution de Student multivariée à ν degrés de liberté de matrice de corrélation $\Sigma \in Mat_d(\mathbb{R})$. La copule de Student s'écrit, pour tout $(u_1, ..., u_d) \in [0,1]^d$:

$$C\left(u_1, \ldots, u_d\right) = t_{\Sigma,v}\left(t_v^{-1}\left(u_1\right), \ldots, t_v^{-1}\left(u_d\right)\right)$$

La copule de Student admet les propriétés suivantes :

- Lorsque le degré de liberté ν tend vers $+\infty$, la copule de Student tend vers la copule gaussienne ;

  En dimension $d = 2$, en notant ρ le coefficient de corrélation linéaire entre les deux variables :
- Elle admet le même coefficient de dépendance forte des extrêmes à gauche et à droite $\lambda_L = \lambda_U = 2 - 2t_{\nu+1}\left(\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right)$;
- Le τ de Kendall est donné par la formule : $\tau = \frac{2}{\pi}\arcsin(\rho)$.

La copule de Student permet, contrairement à la copule gaussienne, de représenter une dépendance au niveau des valeurs extrêmes. Néanmoins, comme pour la copule gaussienne, la dépendance est symétrique (c'est à dire que $C(u_1, ..., u_d) = C(1 - u_1, ..., 1 - u_d)$) ce qui limite son utilisation lorsque la corrélation est asymétrique. A cet effet, les copules archimédiennes offrent une alternative intéressante.

A.2 Les copules archimédiennes

Pour rappel, en posant $u = (u_1, \ldots, u_d) \in [0,1]^n$, une copule archimédienne C est définie par un générateur $\phi : [0,1] \to [0, +\infty]$ et est donnée par la formule :

$$C\left(u_1, \ldots, u_d\right) = \phi^{-1}\left(\phi\left(u_1\right) + \ldots + \phi\left(u_d\right)\right)$$

Avec ϕ de classe $\mathcal{C}^2$ vérifiant $\phi(1) = 0, \phi'(u) \leq 0$ et $\phi''(u) > 0$.

A.2.1 La copule de Gumbel

La copule de Gumbel est une copule archimédienne définie par le générateur :

$$\phi(t) = (-\ln(t))^\theta \qquad \theta \geq 1$$

On a alors :

$$C_\theta(u) = \exp\left(-\left((-\ln(u_1))^\theta + ... + (-\ln(u_d))^\theta\right)^{\frac{1}{\theta}}\right)$$

Cette copule permet de représenter des variables positivement corrélées, et vérifie les propriétés suivantes :

- C_θ tend vers la copule indépendante lorsque θ tend vers 1 ;
- C_θ tend vers la borne de Fréchet-Hoeffding M lorsque θ tend vers $+\infty$;

En dimension $d = 2$:

- Le τ de Kendall est donné par la formule : $\tau = \dfrac{\theta - 1}{\theta}$;
- Elle admet un coefficient de dépendance forte des extrêmes à droite $\lambda_U = 2 - 2^{\frac{1}{\theta}}$, mais pas à gauche $\lambda_L = 0$.

Cette copule asymétrique a le pouvoir de représenter une structure de dépendance plus accentuée sur la queue supérieure, ce qui peut être un avantage dans le secteur de l'assurance pour représenter les valeurs extrêmes.

A.2.2 La copule de Clayton

La copule de Clayton est une copule archimédienne définie par le générateur :

$$\phi(t) = \frac{t^{-\theta} - 1}{\theta} \qquad \theta \in [-1, +\infty[$$

On a alors :

$$C_\theta(u) = \left(u_1^{-\theta} + ... + u_d^{-\theta} - d + 1\right)^{-\frac{1}{\theta}}$$

Cette copule permet de représenter des variables positivement corrélées, et vérifie les propriétés suivantes :

- C_θ tend vers la copule indépendante lorsque θ tend vers 0 ;

- C_θ tend vers la borne de Fréchet-Hoeffding M lorsque θ tend vers $+\infty$;

En dimension $d = 2$:

- Le τ de Kendall est donné par la formule : $\tau = \dfrac{\theta}{\theta + 2}$;
- Elle admet un coefficient de dépendance forte des extrêmes à gauche $\lambda_L = 2^{-\frac{1}{\theta}}$, mais pas à droite $\lambda_U = 0$.

Contrairement à la copule de Gumbel, la copule de Clayton permet d'instaurer une dépendance des extrêmes pour la queue inférieure. Cette dépendance des extrêmes peut être inversée pour obtenir une dépendance de queue à droite en considérant la copule de survie de Clayton définie en dimension 2 par :

$$C_\theta^S(u, v) = u + v - 1 + C_\theta(1 - u, 1 - v)$$

A.2.3 La copule de Frank

La copule de Frank est une copule archimédienne définie par le générateur :

$$\phi(t) = -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right) \qquad \theta \in \mathbb{R}^*$$

On a alors :

$$C_\theta(u) = -\frac{1}{\theta}\ln\left(1 + \frac{\prod_{i=1}^{d}(e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{d-1}}\right)$$

Cette copule permet aussi bien de représenter des corrélation positives que négatives, et vérifie les propriétés suivantes :

- C_θ tend vers la copule indépendante lorsque θ tend vers 0 ;
- C_θ tend vers la borne de Fréchet-Hoeffding M lorsque θ tend vers $+\infty$;
- C_θ tend vers la borne de Fréchet-Hoeffding W lorsque θ tend vers $-\infty$;

En dimension $d = 2$:

- Le τ de Kendall est donné par la formule : $\tau = 1 - \dfrac{4(1 - D_1(\theta))}{\theta}$ où $D_k(\theta) = \frac{k}{\theta}\displaystyle\int_0^\theta \frac{t^k}{e^t - 1}dt$;
- Le ρ de Spearman est donné par la formule : $\rho_S = 1 - \dfrac{12(D_1(\theta) - D_2(\theta))}{\theta}$;
- Cette copule ne présente pas de dépendance de queue ($\lambda_L = \lambda_U = 0$).

Contrairement à la copule de Gumbel et de Clayton, la copule de Frank permet de représenter les corrélations positives et négatives. L'absence de dépendance de queue peut cependant la rendre inadaptée pour modéliser la dépendance au niveau des extrêmes.

A.2.4 La copule de Joe

La copule de Joe est une copule archimédienne définie par le générateur :

$$\phi(t) = -\ln\left(1-(1-t)^{\theta}\right) \qquad \theta \geq 1$$

On a alors :

$$C_\theta(u) = 1 - \left(1 - \prod_{i=1}^{d}\left(1-(1-u_i)^{\theta}\right)\right)^{\frac{1}{\theta}}$$

Cette copule permet de représenter des corrélation positives et vérifie les propriétés suivantes :

- C_θ tend vers la copule indépendante lorsque θ tend vers 1 ;
- C_θ tend vers la borne de Fréchet-Hoeffding M lorsque θ tend vers $+\infty$;

En dimension $d = 2$:

- Le τ de Kendall est donné par la formule : $\tau = 1 + \frac{4}{\theta}\int_0^1 \frac{(1-t^\theta)\ln(1-t^\theta)}{t^{\theta-1}}dt$;
- Elle admet un coefficient de dépendance forte des extrêmes à droite $\lambda_U = 2 - 2^{\frac{1}{\theta}}$, mais pas à gauche $\lambda_L = 0$.

La copule de Joe se rapproche de la copule de Gumbel puisqu'elle représente des corrélations positives et présente les mêmes coefficients de dépendance des extrêmes.

A.2.5 La copule d'Ali-Mikhail-Haq

La copule de d'Ali-Mikhail-Haq est une copule archimédienne définie par le générateur :

$$\phi(t) = \ln\frac{1-\theta(1-t)}{t} \qquad \theta \in [-1,1]$$

On a alors :

$$C_\theta(u) = \frac{1-\theta}{\prod_{i=1}^{d}\frac{1-\theta(1-u_i)}{u_i} - \theta}$$

Cette copule permet aussi bien de représenter des corrélations positives que négatives, et vérifie les propriétés suivantes en dimension $d = 2$:

- Le τ de Kendall est donné par la formule :

$$\tau = \frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2\ln(1-\theta)}{3\theta^2}$$

;

- Le ρ de Spearman est donné par la formule :

$$\rho_S = \frac{12(1+\theta)\displaystyle\int_1^{(1-\theta)} \frac{\ln t}{1-t}dt - 24(1-\theta)\ln(1-\theta)}{\theta^2} - \frac{3(\theta+12)}{\theta}\,;$$

- Si $\theta = 1$, elle admet un coefficient de dépendance forte des extrêmes à gauche $\lambda_L = 0,5$;
- Elle n'admet pas de coefficient de dépendance forte des extrêmes à droite $\lambda_U = 0$.

La copule d'Ali-Mikhail-Haq permet la représentation de corrélations positives et négatives. Dans la plupart des cas ($\theta \neq 1$), elle ne présente pas de dépendance de queue, ce qui peut la rendre inadaptée pour modéliser la dépendance au niveau des extrêmes.

Annexe B

Les lois de probabilité positives et continues utilisées

B.1 La loi exponentielle

Soient $\lambda \in \mathbb{R}_+^*$ et X une variable aléatoire qui suit la loi exponentielle de paramètre λ. On note $X \sim \mathcal{E}(\lambda)$.

La densité de X a pour support l'intervalle $]0, +\infty[$ et est donnée par :

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x>0\}}$$

Sa fonction de répartition est donnée par :

$$F_X(x) = (1 - e^{-\lambda x}) \mathbb{1}_{\{x>0\}}$$

L'espérance et la variance de X sont données par :

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Remarque : La loi exponentielle est peu flexible puisqu'elle ne possède qu'un seul paramètre.

B.2 La loi log-normale

Soient $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$, la variable aléatoire X suit une loi log-normale (Log-$\mathcal{N}(\mu, \sigma^2)$) si $Y = \ln(X)$ suit une loi $\mathcal{N}(\mu, \sigma^2)$.

La densité de X est définie sur $]0,+\infty[$ et est donnée par :

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) = \frac{1}{x} f_Y(\ln(x))$$

où f_Y est la densité de Y.

Sa fonction de répartition est définie sur $]0,+\infty[$ et est donnée par :

$$F_X(x) = F_Y(\ln(x))$$

où F_Y est la fonction de répartition de Y.

L'espérance et la variance de X sont données par :

$$\mathbb{E}[X] = e^{\mu+\frac{\sigma^2}{2}}$$
$$Var(X) = \left(e^{\sigma^2}-1\right) e^{2\mu+\sigma^2}$$

B.3 La loi de Weibull

Soient k et $\lambda \in \mathbb{R}_+^*$, X une variable aléatoire qui suit la loi de Weibull de paramètre de forme k et de paramètre d'échelle λ. On note $X \sim Weibull(k,\lambda)$.

La densité de X a pour support l'intervalle $]0,+\infty[$ et est donnée par :

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k} \mathbb{1}_{\{x>0\}}$$

Sa fonction de répartition est donnée par :

$$F_X(x) = (1-e^{-(\frac{x}{\lambda})^k})\mathbb{1}_{\{x>0\}}$$

L'espérance et la variance de X sont données par :

$$\mathbb{E}[X] = \lambda\Gamma\left(1+\frac{1}{k}\right)$$
$$Var(X) = \lambda^2\left(\Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2\right)$$

où Γ est la fonction Gamma d'Euler.

Remarque : Si $X^k \sim \mathcal{E}(\lambda)$, alors X suit une loi de Weibull de paramètres k et λ.

B.4 La loi Gamma

Soient k et $\theta \in \mathbb{R}_+^*$, X une variable aléatoire qui suit la loi Gamma de paramètre de forme k et de paramètre d'échelle θ. On note $X \sim \Gamma(k,\theta)$.

La densité de X a pour support l'intervalle $]0, +\infty[$ et est donnée par :

$$f_X(x) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k}\mathbb{1}_{\{x>0\}}$$

Il n'existe pas de formule explicite pour définir sa fonction de répartition, elle est donc donnée par :

$$F_X(x) = \int_0^x \frac{t^{k-1}e^{-\frac{t}{\theta}}}{\Gamma(k)\theta^k}dt\mathbb{1}_{\{x>0\}}$$

L'espérance et la variance de X sont données par :

$$\mathbb{E}[X] = k\theta$$

$$Var(X) = k\theta^2$$

Remarque : Si $k = 1$, alors X suit une loi exponentielle de paramètre $\frac{1}{\theta}$.

B.5 La loi de Burr

Soient m, s et $f \in \mathbb{R}_+^*$, X une variable aléatoire qui suit la loi de Burr de paramètre de localisation m, de paramètre de dispersion s et de paramètre de famille f. On note $X \sim Burr(m, s, f)$.

La densité de X a pour support l'intervalle $]0, +\infty[$ et est donnée par :

$$f_X(x) = fs\frac{\left(\frac{x}{m}\right)^{s-1}}{m\left(1 + \left(\frac{x}{m}\right)^s\right)^{f+1}}\mathbb{1}_{\{x>0\}}$$

Sa fonction de répartition est donnée par :

$$F_X(x) = 1 - \left(1 + \left(\frac{x}{m}\right)^s\right)^{-f}\mathbb{1}_{\{x>0\}}$$

L'espérance et la variance de X sont données par :

$$\mathbb{E}[X] = m\frac{\Gamma\left(\frac{2}{s}\right)\Gamma\left(f - \frac{2}{s}\right)}{s\Gamma(f)}$$

$$Var(X) = m^2\frac{\Gamma\left(\frac{3}{s}\right)\Gamma\left(f - \frac{3}{s}\right)}{s\Gamma(f)}$$

Remarque : La loi de Burr est flexible grâce à ses trois paramètres et s'ajuste généralement bien aux données.

Annexe C

Généralités sur les arbres CART

Les arbres CART (*Classification And Regression Trees*), développés par Breiman et al (1984) constituent une méthode utilisée pour regrouper des individus hétérogènes en classes homogènes afin de résumer l'information d'une grande base de données. C'est une méthode de classification et de régression non paramétrique pouvant être utilisée comme telle mais étant aussi à la base de nombreux autres algorithmes de *Machine Learning*.

Un arbre CART est constitué de plusieurs éléments :

- Une racine, qui contient l'ensemble de la population à segmenter ;
- Des branches, qui contiennent les règles de division qui permettent de segmenter la population ;
- Des nœuds, qui contiennent les sous-populations de l'arbre CART créées à partir de la racine ;
- Des feuilles, qui représentent les nœuds terminaux et contiennent les sous-populations homogènes créées, et qui fournissent une estimation de la variable quantitative d'intérêt.

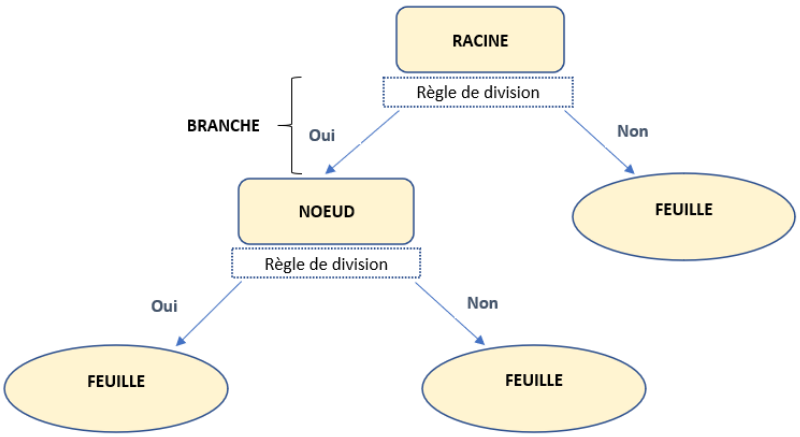

FIGURE C.1 – Représentation d'un arbre CART

Un arbre CART se lit de la racine vers les feuilles. Une règle de division apparaît à chaque ramification. Cette règle admet une réponse binaire (oui/non) et n'est basée que sur une variable explicative. Chaque individu de la population initiale appartient à une unique feuille.

La construction d'un arbre CART se décompose en deux phase. La première étape consiste à construire un arbre maximal, avant d'effectuer une étape d'élagage qui formera un sous-arbre de l'arbre maximal appelé arbre optimal.

C.1 Construction de l'arbre maximal

On considère les notations suivantes :

- $i \in \{1, ..., n\}$: l'identifiant de l'observation ;
- $j \in \{1, ..., k\}$: l'identifiant du facteur de risque (covariable) ;
- Y_i : la valeur de la variable à expliquer du i^{eme} individu ;
- $X_i = (X_i^1, ..., X_i^k)$: le vecteur des facteurs de risque de l'individu i ;
- $\chi = (\chi_1, ..., \chi_k)$: l'espace des covariables.

L'algorithme de construction de l'arbre maximal démarre de la racine qui regroupe l'ensemble de la population à segmenter. La procédure cherche alors à déterminer une règle de division $\{X^j \in P\}$ contre $\left\{X^j \in \overline{P}\right\}$ concernant le facteur de risque j et la partition $\left\{P, \overline{P}\right\}$ de χ_j et résultant sur des nœuds N_1 et N_2 qui minimisent une certaine fonction de coût.

- Pour un algorithme de régression, la fonction à minimiser est la variance intra-groupes qui résulte de la segmentation d'un nœud père en deux nœuds fils N_1 et N_2. La variance d'un nœud N est donnée par $V(N) = \frac{1}{\#N} \sum_{i \in N} (Y_i - \overline{Y}_N)^2$ où $\overline{Y}_N$ est la moyenne des Y_i appartenant au nœud N.
  Il faut alors minimiser :
  $$V_{intra}(N_1, N_2) = \frac{1}{n} \sum_{i \in N_1} (Y_i - \overline{Y}_{N_1})^2 + \frac{1}{n} \sum_{i \in N_2} (Y_i - \overline{Y}_{N_2})^2 = \frac{\#N_1}{n} V(N_1) + \frac{\#N_2}{n} V(N_2)$$
- Pour un algorithme de classification (où les classes sont notées $\{1, ..., L\}$) , c'est l'impureté des nœuds fils qui est considérée, généralement via l'indice de Gini défini par $\phi(N) = \sum_{c=1}^{L} p_N^c (1 - p_N^c)$, où p_N^c est la proportion d'observations de classe c dans le nœud N.
  Il faut alors minimiser :
  $$G(N_1, N_2) = \frac{\#N_1}{n} \phi(N_1) + \frac{\#N_2}{n} \phi(N_2)$$

Une fois la racine divisée en deux sous-groupes, on réitère plusieurs fois le procédé sur chacun des deux nœud fils créés par la procédure, jusqu'à atteindre un critère d'arrêt qui met fin à l'algorithme. Cette condition d'arrêt peut par exemple être d'interrompre le procédé lorsque le nœud contient moins d'un certain nombre d'observations, ou lorsque les individus d'un même nœud possèdent les mêmes valeurs de facteurs de risque. L'arbre obtenu est alors appelé arbre maximal.

C.2 Elagage

L'étape d'élagage qui s'ensuit s'appuie sur l'arbre maximal construit afin d'obtenir un arbre dit optimal. Cette phase consiste en la recherche du meilleur sous-arbre élagué de l'arbre maximal, qui découle de l'idée que l'arbre maximal possède une très grande variance et un faible biais contrairement à l'arbre limité à sa racine qui possède un fort biais et une faible variance. La procédure d'élagage considère l'ensemble des sous-arbres construits à partir de l'arbre maximal et cherche à minimiser une erreur d'ajustement pénalisée ou non par le nombre de feuilles de l'arbre (complexité du modèle). Cette erreur considérée peut par exemple être l'erreur de validation croisée.

C.3 Principaux avantages et inconvénients

L'algorithme CART possède comme principal avantage sa grande simplicité d'interprétation. Le résultat peut en effet être facilement lu et interprété par une personne non spécialiste grâce à l'affichage clair des règles de segmentation. De plus, c'est une méthode statistique consistante, non-paramétrique et adaptée à la gestion d'un grand nombre de variables explicatives.

En opposition avec ces points forts, l'algorithme CART admet quelques faiblesses. On peut citer comme inconvénient l'élagage de l'arbre maximal lorsque ce dernier admet un grand nombre de feuilles. Le nombre de sous-arbre étant exponentiel avec le nombre de feuilles, le recours à un algorithme récursif est nécessaire et peut conduire à un arbre optimal qui n'est pas l'optimum global. Son principal point faible reste tout de même son manque de robustesse. En effet, l'arbre optimal obtenu peut varier fortement avec une faible variation du jeu de données initial. Les solutions peuvent alors être de faire appel à des stratégies d'agrégation comme le *bagging* ou le *boosting*.

Annexe D

Les Modèles Linéaires Généralisés

Les Modèles Linéaires Généralisés (MLG), décrits pour la première fois sous cette appellation par Nelder et Wedderburn (1972), sont une famille de modèles qui permettent d'étudier la nature de la liaison entre une variable à expliquer Y et un ensemble de variables explicatives $X = (X_1, ..., X_k)$.

On considère les notations suivantes :

- $i \in \{1, ..., n\}$: l'identifiant de l'observation ;
- $j \in \{1, ..., k\}$: l'identifiant du facteur de risque (covariable) ;
- Y_i : la valeur de la variable à expliquer du i^{eme} individu ;
- $X_i = (X_i^1, ..., X_i^k)$: le vecteur des facteurs de risque de l'individu i.

D.1 Famille exponentielle

Pour utiliser les modèles linéaires généralisés, on suppose tout d'abord que la loi de probabilité de la variable à expliquer Y appartient à la famille exponentielle. Une variable aléatoire appartient à la famille exponentielle si sa densité f peut s'écrire sous la forme :

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \qquad (1)$$

Exemple : La loi Gamma fait partie de la famille exponentielle. En effet, sa densité s'écrit :

$$\begin{aligned} f(y, k, \lambda) &= \frac{y^{k-1} e^{-\frac{y}{\lambda}}}{\Gamma(k)\lambda^k} \\ &= \exp\left(-\frac{y}{\lambda} + \ln\left(y^{k-1}\right) - \ln\left(\Gamma(k)\lambda^k\right)\right) \\ &= \exp\left(\frac{-\frac{y}{k\lambda} + \ln\left(\frac{1}{k\lambda}\right)}{\frac{1}{k}} + k\ln(k) - \ln(\Gamma(k)) + \ln\left(y^{k-1}\right)\right) \end{aligned}$$

On identifie alors dans l'expression (1) :

$\theta = \frac{1}{k\lambda}$
$b(\theta) = \ln(\theta)$
$\varphi = k$
$a(\varphi) = -\frac{1}{\varphi}$
$c(y, \varphi) = \varphi \ln(\varphi) - \ln(\Gamma(\varphi)) + \ln\left(y^{\varphi-1}\right)$

D.2 Prédicteur linéaire

Tout comme pour un modèle linéaire classique, les réponses prédites par le modèle le sont à partir d'une combinaison linéaire des variables explicatives. Ce prédicteur linéaire, noté η est donné par la formule :

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_i^j$$

D.3 Fonction de lien

En revanche, contrairement aux modèles linéaires classiques, les valeurs des prédicteurs linéaires η_i ne correspondent pas à la prédiction moyenne des observations, mais à une transformation de celle-ci. Autrement dit, si on note g une fonction inversible (que l'on appelle **fonction de lien**), la variable Y_i est expliquée par la relation :

$$g\left(\mathbb{E}[Y_i]\right) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_i^j$$

$$\Leftrightarrow \mathbb{E}[Y_i] = g^{-1}\left(\eta_i\right)$$

D.4 Estimation des paramètres

Les paramètres β_j du modèle sont estimés par maximum de vraisemblance. Cette méthode ne permet généralement pas d'obtenir de formule explicite et l'utilisation d'un algorithme de résolution numérique est nécessaire.

D.5 Principaux avantages et inconvénients

Les modèles linéaires généralisés permettent de modéliser des réponses diverses (dans les ensembles $\mathbb{R}, \mathbb{R}^+, \mathbb{N}$, etc.), en quantifiant l'impact des différents facteurs de risque. Contrairement aux modèles linéaires classiques, ils disposent d'une flexibilité intéressante. En effet, la loi de la variable à expliquer n'est plus limitée à la loi normale, et celle-ci n'est plus obligatoirement estimée directement par le prédicteur linéaire, mais à travers une fonction de lien. Ces modèles se basent cependant sur une hypothèse forte : les variables explicatives X_i doivent être indépendantes entre elles deux à deux.

Annexe E

Théorie des valeurs extrêmes

La théorie des valeurs extrêmes, utilisée dans ce mémoire pour l'ajustement des différentes lois, permet d'appréhender le comportement de variables aléatoires au niveau des queues de distribution. Les principaux éléments théoriques seront présentés dans cette annexe. Dans la suite, on pose :

- $X_1, ..., X_n$ un ensemble d'observations de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition inconnue F ;
- $M_n = \max(X_1, ..., X_n)$.

Avant d'entrer dans le détail de la théorie des valeurs extrêmes, on définit les trois distributions suivantes qui sont fondamentales dans l'étude des queues de distributions :

Définition E.0.1.

$$\begin{array}{ll} \textit{La distribution de Fréchet } (\alpha > 0) : & \Phi_\alpha(x) = \begin{cases} 0 & \textit{si } x \leq 0 \\ \exp\{-x^{-\alpha}\} & \textit{si } x > 0 \end{cases} \\ \textit{La distribution de Weibull } (\alpha > 0) : & \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & \textit{si } x \leq 0 \\ 1 & \textit{si } x > 0 \end{cases} \\ \textit{La distribution de Gumbel :} & \Lambda(x) = \exp\{-e^{-x}\} \quad x \in \mathbb{R} \end{array}$$

E.1 Loi du maximum

Lorsque l'on s'intéresse aux valeurs extrêmes que peut prendre une variable aléatoire, on peut chercher à connaître le comportement du maximum. Sous certaines conditions, la loi du maximum peut être caractérisée grâce au théorème de Fisher-Typpett.

Théorème E.1.1. *S'il existe des suites de réels $a_n > 0$ et b_n, telles que*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \underset{n\to\infty}{\longrightarrow} G(x)$$

pour une distribution non-dégénérée G, alors G est du même type que la distribution suivante :

$$G_\xi(x) = \exp\left(-(1+\xi x)_+^{-1/\xi}\right).$$

La distribution G_ξ est appelée distribution des extrêmes généralisée (GEV).

Le paramètre ξ, appelé paramètre de forme, donne une information sur l'épaisseur de la queue de distribution et l'appartenance à un domaine d'attraction de loi particulier :

- Si $\xi > 0$, la distribution est à « queue épaisse » et appartient au domaine d'attraction de la loi de Fréchet ;
- Si $\xi = 0$, la distribution est à « queue intermédiaire » et appartient au domaine d'attraction de la loi de Gumbel ;
- Si $\xi < 0$, la distribution est à « queue fine » et appartient au domaine d'attraction de la loi de Weibull.

Définition E.1.1. *Soit G une distribution des extrêmes généralisée. On dit que F appartient au domaine d'attraction de G s'il existe deux suites (a_n) et (b_n) telle que :*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \left[F\left(a_n x + b_n\right)\right]^n \underset{n\to\infty}{\longrightarrow} G(x)$$

E.2 Loi des dépassements de seuil

Plutôt que de considérer le maximum d'un échantillon, on peut s'intéresser aux lois des dépassements du seuil u_n, c'est à dire aux observations $(X_i - u_n)_+$ strictement positives. Il est alors nécessaire d'introduire la distribution de Pareto généralisée (GPD), qui caractérise la loi des dépassements de seuil.

Définition E.2.1. *La distribution de Pareto généralisée $GPD(\beta, \xi)$ est définie pour $x \geq 0$ si $\xi \geq 0$ et $x \in [0, -\beta/\xi]$ si $\xi < 0$ par :*

$$G_{\xi,\beta}(x) = \begin{cases} 1 - [1 + \xi(x/\beta)]_+^{-1/\xi} & si\ \xi \neq 0 \\ 1 - e^{-x/\beta} & si\ \xi = 0 \end{cases}$$

Comme pour la distribution des extrêmes généralisée (GEV), la distribution de Pareto généralisée (GPD) possède un paramètre de forme ξ. Il existe un lien étroit entre ces deux distributions. Cela permet de disposer de plusieurs approches pour déterminer ce paramètre de forme, soit en passant par la GPD, soit en passant par la GEV, en fonction du type de données disponibles.

E.3 Méthodes de détermination d'un seuil

Dans la pratique, l'étape la plus sensible est celle du choix du seuil u pour définir les dépassements. En effet, utiliser un seuil peu élevé permet d'augmenter le nombre de données, mais

augmente le risque que l'approximation par la loi Pareto généralisée soit mauvaise. Au contraire, utiliser un seuil trop élevé limite le nombre de données et les estimateurs sont moins précis. Il est donc nécessaire de trouver un équilibre entre ces deux cas extrêmes, en utilisant les propriétés de la GPD.

E.3.1 Le graphique des dépassements moyens

Le graphique des dépassements moyens (*mean excess plot*) se base sur la propriété suivante :

Propriété E.3.1. *Soit $Y_u = X - u|X > u$ la variable aléatoire représentant les dépassements du seuil u. Alors, si $Y_u \sim GPD(\sigma_u, \xi)$, la fonction de dépassement de seuil au-delà de $v > u$ définie par*

$$E(X - v \mid X > v) = \frac{\sigma_u + \xi(v - u)}{1 - \xi}$$

est linéaire en v avec une pente $\dfrac{\xi}{1-\xi}$.

La procédure consiste donc à tracer la fonction de dépassement moyen pour plusieurs seuils. Si la variable aléatoire suit une loi de Pareto Généralisée pour le seuil u, alors le graphique doit être approximativement linéaire au-delà de ce seuil. Cela permet alors de sélectionner le seuil le plus bas à partir duquel la loi suit une GPD. C'est la méthode qui a été privilégiée lors des travaux effectués pour ce mémoire.

E.3.2 Le graphique de stabilité du paramètre d'échelle

Le graphique de stabilité du paramètre d'échelle se base sur les propriétés suivantes :

Propriété E.3.2. *Soit $Y_u = X - u|X > u$. Si $Y_u \sim GPD(\sigma_u, \xi)$, alors pour tout seuil $v \geq u$, $Y_v \sim GPD(\sigma_u + \xi(v - u), \xi)$. En particulier, le paramètre de forme ξ ne dépend pas du seuil.*

Propriété E.3.3. *En définissant $\sigma^* = \sigma_v - \xi_v$, alors σ^* ne dépend plus de $v \geq u$ si la variable aléatoire suit une GPD pour le seuil u.*

La procédure consiste donc à estimer $\widehat{\sigma}^*$ et $\widehat{\xi}$ et à les représenter sur un graphique pour différents seuils. Si la variable aléatoire suit une GPD pour le seuil u, alors les paramètres (σ^*, ξ) seront approximativement constants pour les seuils v plus grands.

Au-delà des méthodes de détermination de seuil présentées ci-dessus, ce dernier peut être confirmé en vérifiant la stabilité des estimateurs du paramètre de forme ξ. Différents estimateurs de ce paramètre existent. On peut par exemple citer l'estimateur de Pickands, l'estimateur de Dekkers, Einmahl et de Haan lorsque $\xi \in \mathbb{R}$, ou encore l'estimateur de Hill lorsque $\xi > 0$.

Annexe F

V de Cramer et Rapport de corrélation

Cette annexe a pour objectif de présenter deux indicateurs de corrélation utilisés dans le corps de ce mémoire.

F.1 Le V de Cramer

Le V de Cramer est un indicateur permettant de mesurer la corrélation entre deux variables qualitatives X_1 et X_2.

Bien que le test du χ^2 permette de savoir si les variables sont dépendantes, il ne donne pas d'indication sur l'intensité de cette relation. Le V de Cramer y remédie et permet également de s'affranchir de l'influence néfaste des grands échantillons sur le test du χ^2. Basé sur ce dernier, le coefficient de Cramer est donné, pour un échantillon de taille n par la formule :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \times \min(K_1 - 1, K_2 - 1)}}$$

avec K_1 et K_2 le nombre de modalités respectives de X_1 et X_2.

Il est important de préciser que le terme $\chi^2_{max} = n \times \min(K_1 - 1, K_2 - 1)$ représente la valeur maximale que peut prendre la statistique de test du χ^2.

La valeur du V de Cramer est comprise entre 0 et 1. Plus il est proche de 0, plus il y a indépendance entre les deux variables. Au contraire, plus il est proche de 1, plus la corrélation est importante.

F.2 Le rapport de corrélation

Le rapport de corrélation est un indicateur permettant de mesurer la corrélation entre une variable qualitative X à k modalités (notées $J_1, ..., J_k$) et une variable quantitative Y.

Ce coefficient s'appuie sur l'analyse de variance à un facteur. Dans ce cadre, une équation décompose la variance totale en une variance inter-classes et une variance intra-classes. On a alors :

$$Var_{Totale} = Var_{Inter} + Var_{Intra}$$

où :

$$Var_{Totale} = \frac{1}{n} \sum_{j} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

$$Var_{Intra} = \frac{1}{n} \sum_{j} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

$$Var_{Inter} = \frac{1}{n} \sum_{j} n_j \, (\bar{y}_j - \bar{y})^2$$

avec :

- y_{ij} la i^{eme} observation de la classe J_j ;
- n_j l'effectif de la classe J_j ;
- $\bar{y}_j$ la moyenne de la classe J_j ;
- n l'effectif total ;
- $\bar{y}$ la moyenne de l'ensemble de l'effectif.

Le rapport de corrélation est le rapport entre la variance inter-classes et la variance totale :

$$\widehat{\eta}^2 = \frac{Var_{Inter}}{Var_{Totale}} = \frac{Var_{Inter}}{Var_{Inter} + Var_{Intra}}$$

Il est compris entre 0 et 1 et mesure le poids de la variance entre les modalités dans la variance totale. Par conséquent, plus ce coefficient est proche de 0 (cas où toutes les modalités de X ont la même moyenne pour la variable Y), plus les variables sont indépendantes. Au contraire, plus ce coefficient est proche de 1 (cas où les individus d'une même modalité ont la même valeur de Y, et ce pour toutes les modalités de X), plus les variables sont liées.

Table des figures

Liste des tableaux