Mémoire présenté le :

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

Par : Antoine MISERAY

Titre : Analyse du coût des sinistres dans le cadre de la modélisation ligne à ligne d'un portefeuille d'assurance automobile

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 signature

Membres présents du jury de l'ISFA

Secrétariat :

Bibliothèque :

Entreprise :

Nom : Prim'Act

Signature :

Directeur de mémoire en entreprise :

Nom : Frédéric PLANCHET

Signature :

Invité :

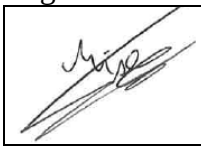
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat



Mémoire d'Actuaire ISFA

Analyse du coût des sinistres dans le cadre de la modélisation ligne à ligne d'un portefeuille d'assurance automobile

Réalisé par Antoine Miseray

Tuteur entreprise : Frédéric Planchet Tuteur ISFA : Yahia Salhi

Lyon, le 21 février 2015

Mots Clés

Méthode ligne à ligne, Provision pour Sinistres A Payer (PSAP), *Chain Ladder*, *Mack*, modèles de censure, estimateur Kaplan-Meier, copule, dépendance, mesures de risque, besoin en fonds propres, algorithme de Marshall-Olkin.

Résumé

En assurance [1], les résultats liés à l'estimation des provisions réglementaires sont directement conditionnés par le choix des méthodes à employer. Il existe un nombre important de méthodes de provisionnement et la réglementation n'impose pas aux organismes assureurs l'utilisation d'une méthode particulière. Cependant, le régulateur prêtera une attention toute particulière à ce que le montant des provisions constituées couvre la sinistralité à venir.

La plupart des organismes assureurs utilisent en pratique des méthodes de provisionnement déterministes comme *Chain Ladder* se basant sur des données agrégées. Ces méthodes ne permettent pas de déterminer l'erreur de prédiction associée aux réserves. Pour pallier à cela, des méthodes stochastiques existent. Cependant, ces méthodes de provisionnement qui se basent sur des données de sinistres agrégées sont beaucoup moins adaptées dans les cas où nous disposons d'un portefeuille de sinistres avec une forte volatilité des montants.

L'objectif de ce mémoire, et en particulier la première partie du travail, est la mise en œuvre d'une méthode ligne à ligne pour le provisionnement des sinistres. Il sera ensuite intéressant de comparer ces résultats avec les provisions obtenues à l'aide des triangles et avec les provisions "dossier" [2]. L'intérêt est d'avoir recours à une méthode habituellement employée pour la réalisation de tables de mortalité (les modèles de censure) et de quantifier dans quelle mesure l'application de ces modèles en assurance non-vie permettrait d'appréhender l'estimation des provisions de sinistres avec davantage de précision. Pour ce faire, il est nécessaire de posséder un certain nombre d'informations pour chacun des sinistres présents dans la base. Les données relatives aux sinistres ne sont donc plus regroupées par année de survenance et par année de développement mais chaque sinistre est considéré individuellement. Enfin, il sera intéressant de faire quelques analyses sur les caractéristiques des sinistres en tenant compte de la nature de la garantie comme variable d'hétérogénéité.

Une fois la modélisation des coûts de sinistres effectuée à l'aide d'un modèle de censure, il sera intéressant dans un second temps de modéliser la dépendance entre les trois branches de risque étudiées dans ce mémoire : la garantie Responsabilité Civile (RC) corporelle, la RC matérielle et la garantie Dommage. Au regard de sa complexité, la modélisation des dépendances entre branches de risque d'un portefeuille d'assurance non vie est souvent effectuée sur la base d'hypothèses simplificatrices. La théorie des copules permet aujourd'hui de représenter la dépendance entre branches de risques de manière intelligible. L'objet de la seconde partie de ce mémoire est de présenter comment utiliser les copules afin d'estimer le besoin en fonds propres d'une société d'assurance non vie en considérant uniquement le risque de fluctuation des sinistres. Pour ce faire, une méthodologie adaptée de recherche des dépendances entre les trois garanties étudiées est suivie. Les copules qui modélisent le mieux ces dépendances sont ensuite retenues en s'appuyant sur des tests d'ajustement. Enfin, des simulations réalisées à partir de l'algorithme de Marshall-Olkin permettront de quantifier l'impact de la prise en compte des dépendances sur le besoin en fonds propres de la société, estimé par deux mesures de risque (la *Value at Risk* et la *Tail Value at Risk*).

1. Plus particulièrement, dans le cadre de ce mémoire, en IARD : Incendie, Accidents et Risques Divers.
2. Provisions constituées dossier par dossier.

Keywords

Line by line method, outstanding claims reserve, Chain Ladder, Mack, censored models, Kaplan-Meier estimator, copula, dependency, risk measures, capital requirement, Marshall-Olkin algorithm.

Abstract

For non-life insurance companies, results of statutory reserves are directly conditioned by the method chosen. Currently, there are many reserving methods and the regulation do not require insurers to use a specific method. However, the regulator is making sure that the amount of the reserve established covers the future claims costs.

Most of insurers use reserving methods like Chain Ladder which is based on aggregate data. The disadvantage of these methods is that they do not estimate the prediction error of the reserve estimation. In order to solve this problem, stochastic methods can be used. However, these methods based on aggregate data are less adequate in case of a portfolio with a high volatility of claims.

The aim of this thesis, and in particular the first part, is to apply a line by line method in order to estimate the reserve of claims. Then, it will be interesting to compare these results to reserves estimated with the Chain Ladder method and individual *"dossier"* reserves. The interest is to use a method generally used for mortality tables development (censored models) and to quantify how the application of these models to the field of non-life insurance would allow to understand more precisely the estimation of these reserves. This requires to gather plenty of information for each claim of the data base and to consider claims individually. Eventually, it will be interesting to perform some analysis based on characteristics of claims considering the guarantee like a heterogeneity variable.

After the modelling of claims costs with censored models, using Kaplan-Meier estimator, it is interesting to model the dependence between three French guarantees : *Responsabilité Civile corporelle, Responsabilité Civile matérielle* and *dommage*. Because it is complex, the modelling of dependent guarantees of a non-life insurance portfolio has often been based on a set of simplified assumptions. Now, copula theory can represent dependence between risks in an intelligible way. The aim of the second part of this thesis is to present how it is possible to use copula theory in order to assess the capital requirement of a non-life insurer only adjusted for the risk of volatility in insurance losses. First, a specific methodology is developed to detect dependencies embedded in the portfolio. Secondly, copulas that better reflect those links are selected using different adjustment tests. Eventually, simulations obtained with a Marshall-Olkin algorithm are used to quantify the impact of dependencies on the capital requirement of the company, the latter being assessed thanks to two risk measures (the Value at Risk and the Tail Value at Risk).

Remerciements

Je tiens tout particulièrement à remercier mon référent d'entreprise, M. Frédéric Planchet, associé chez Prim'Act, pour son expertise en la matière et ses conseils avisés mais également pour sa disponibilité et son investissement tout au long de l'élaboration de ce mémoire.

Je remercie également mon référent d'école, M. Yahia Salhi, pour ses conseils ainsi que pour son suivi régulier.

Merci à l'ensemble des collaborateurs Prim'Act pour leurs conseils et aux dirigeants qui m'ont permis de travailler dans d'excellentes conditions.

Mes remerciements s'adressent également aux équipes pédagogiques et administratives de l'Institut de Science Financière et d'Assurances.

Enfin, je remercie ma famille et mes proches pour m'avoir soutenu et fait confiance tout au long de mes études universitaires.

Je vous souhaite de prendre autant de plaisir en lisant ce mémoire que j'ai pu en avoir moi-même lors de sa réalisation.

Sommaire

Introduction

Ce mémoire s'appuie sur des données d'assurance IARD. Le portefeuille étudié concerne plus particulièrement la sinistralité automobile d'une compagnie pour les années de survenance allant de 1993 à 2009. Ce portefeuille regroupe plusieurs garanties mais le choix des analyses s'est tourné principalement sur les garanties Responsabilité Civile corporelle [3], Responsabilité Civile matérielle [4] et Dommage.

La difficulté associée à l'estimation des provisions réglementaires découle directement des choix relatifs aux méthodes employées. En France, il n'y a pas d'obligation concernant la ou les méthodes à employer, à condition de pouvoir justifier le choix effectué auprès de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR). Le premier objectif de ce mémoire est de mettre en place une méthode de provisionnement ligne à ligne et d'être en mesure de pouvoir la comparer avec une méthode plus classique et plus utilisée en pratique, *Chain Ladder*.

Cette dernière méthode repose sur des données agrégées. Il sera donc intéressant de comparer deux méthodes qui s'appuient initialement sur les mêmes données mais qui reposent sur une approche significativement différente. Enfin, le fait de pouvoir disposer des montants de provision "dossier" [5] devrait permettre de juger objectivement de la fiabilité de ces méthodes et ainsi retenir la plus adaptée pour provisionner les branches de risque étudiées.

La méthode développée dans la première partie du mémoire repose sur des modèles habituellement utilisés pour la création de tables de mortalité : les modèles de censure. L'intérêt est donc de déterminer dans quelle mesure l'utilisation de ces modèles sur des données ligne à ligne peut permettre d'améliorer l'estimation des Provisions pour Sinistres A Payer (PSAP) [6]. Aussi, l'étude des branches ayant des caractéristiques différentes (branches "longues"/branches "courtes") devrait potentiellement nous permettre de dégager des avantages quant à l'utilisation de telle ou telle méthode pour une catégorie de branche donnée.

Afin de mettre en place cette méthode d'estimation, un travail important a été réalisé sur la base de données. En effet, il a été nécessaire de mettre en place plusieurs tests de cohérence dans le but d'éliminer les valeurs aberrantes et ainsi aboutir à une base de données fiable. De plus, un travail de mise en forme des données a été également indispensable pour pouvoir appliquer des modèles de censure. Etant donné le volume important de la base de données, ce travail a été réalisé à l'aide du logiciel SAS. Ensuite, le logiciel R a été utilisé pour le travail de modélisation.

Afin de se prémunir de résultats défavorables non anticipés, les sociétés d'assurance

3. Responsabilité Civile (RC) au titre de dommages corporels.
4. Responsabilité Civile (RC) au titre de dommages matériels.
5. Provisions constituées dossier par dossier.
6. Cette notion sera définie dans le chapitre 6, page 73, et est détaillée dans l'article de Magali KELLE [2007] [1].

doivent justifier d'un niveau minimum de fonds propres. Alors que les autorités de contrôle exigent une solvabilité maximale, les actionnaires cherchent à limiter le capital introduit dans l'entreprise afin de maximiser le rendement de leur investissement. Ceci implique de déterminer un optimum qui prend en compte les potentielles dépendances entre les risques qui découlent de l'activité de l'entreprise d'assurance. Auparavant, soit les risques étaient considérés comme indépendants entre eux, soit leur dépendance était admise en leur imposant de suivre la même loi de distribution. Maintenant, la théorie des copules permet de modéliser les liaisons entre risques de manière plus réaliste [7].

Ainsi, le second axe de développement de ce mémoire concerne l'étude de la dépendance entre les différentes garanties étudiées. Ce travail propose une analyse des charges multidimensionnelles à travers l'utilisation de la théorie des copules. En effet, les branches de notre portefeuille sont plus ou moins corrélées entre elles. Il s'agit de modéliser le besoin en fonds propres *via* deux mesures de risque [8] pour ensuite quantifier, à l'aide de simulations réalisées selon l'algorithme de Marshall-Olkin, de quelle manière ce dernier est impacté par la prise en considération des dépendances modélisées par des copules sélectionnées préalablement.

Une analyse de la dépendance pourra donc certainement permettre d'apprécier une méthode de calcul du besoin en fonds propres d'une entreprise d'assurance donnée, qui prendra en compte cette dépendance, plutôt qu'une méthode qui considérera les branches de risque comme étant indépendantes entre elles [9]. Pour ce faire, un modèle théorique sera mis en évidence et une application, développée au chapitre 7, page 94, va permettre d'illustrer l'importance de la prise en compte de la dépendance entre les trois branches de risque étudiées.

7. Nous verrons pourquoi dans la seconde partie de ce mémoire.
8. La *Value at Risk* et la *Tail Value at Risk*.
9. A noter que le provisionnement réalisé sur la base d'une espérance (premier volet de ce mémoire) ne dépend pas de la structure de dépendance (l'espérance de la somme est égale à la somme des espérances), d'où l'intérêt de passer par une approche "besoin en capital économique" pour illustrer l'importance de la prise en compte de la structure de dépendance entre branches de risque.

Première partie

Contexte

Chapitre 1

Présentation des données

L'objet de ce chapitre est d'effectuer d'une part, une brève présentation des aspects importants en lien avec l'assurance automobile française et d'autre part, une description détaillée du portefeuille étudié dans le but de mettre en avant les données à partir desquelles seront développés les modèles. Pour ce faire, la section 1.1, page 14, va rappeler les caractéristiques essentielles du marché de l'assurance automobile français et définir le périmètre des garanties sur lesquelles nous allons travailler. La section 1.2, page 17, présentera les caractéristiques des données étudiées (évolution des cotisations et présentation des garanties étudiées).

1.1 Une base de données constituée de l'état des règlements cumulés des garanties d'un contrat sinistré d'assurance automobile

1.1.1 Description des garanties présentes dans la base de données

En France, la garantie Responsabilité Civile est la seule assurance obligatoire en automobile. La définition énoncée dans le paragraphe suivant s'appuie sur celle proposée par la Fédération Française des Sociétés d'Assurances (FFSA) [2].

La garantie Responsabilité Civile permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un de ses passagers (blessures ou décès d'un piéton, d'un passager, ou d'un occupant d'un autre véhicule ; dégâts aux autres voitures, deux-roues, immeubles, etc.). Cette garantie Responsabilité Civile couvre les conducteurs autorisés ou non autorisés. Cependant, après avoir indemnisé les victimes, l'assureur peut disposer d'un recours à l'encontre des conducteurs non autorisés.

La garantie Responsabilité Civile regroupe la garantie Responsabilité Civile corporelle et la garantie Responsabilité Civile matérielle. De ce fait, les sinistres qui touchent la garantie Responsabilité Civile sont ensuite segmentés en deux catégories :

- les sinistres relatifs à la garantie RC matérielle pour lesquels la charge sinistre est en général rapidement connue et les sinistres rapidement clôturés ;
- les sinistres relatifs à la garantie RC corporelle pour lesquels le délai de clôture de sinistres peut être long car il dépend de l'état de consolidation médicale du tiers.

Il existe d'autres garanties non obligatoires, elles se regroupent en 3 catégories :

- les garanties qui couvrent les dommages subis par le véhicule (dommage matériel, vol, incendie, bris de glace, vandalisme) ;

- les dommages subis par le conducteur du véhicule (dommage corporel) ;
- les garanties de service (assistance et protection juridique).

Les applications réalisées dans ce mémoire portent uniquement sur les garanties Responsabilité Civile corporelle et matérielle ainsi que sur la garantie Dommage.
Néanmoins, des statistiques descriptives vont permettre de connaître la composition de notre portefeuille.

1.1.2 Présentation de la base de données des sinistres

La base des sinistres est constituée de l'ensemble des sinistres survenus entre le 01/01/1993 et le 31/12/2009, vus au 31/12/2009. Elle présente un historique de 17 années de survenance avec des données détaillées par sinistre offrant une vision de la sinistralité journalière.

Toutefois, s'agissant d'une vision de la base sinistre arrêtée au 31/12/2009, les données portant sur les derniers exercices de survenance présentent un niveau d'information moins complet (non stabilisé) que les exercices anciens, ainsi les derniers exercices et notamment 2009 n'intègrent pas tous les *IBNR*[1].

Ceci explique notamment le fait que l'exercice 2009 semble présenter une sinistralité plus faible, avec un taux de sinistres sans suite encore faible pour les branches RC (le marquage des sinistres en sans suite nécessitant des délais liés aux expertises). Ce propos est illustré à l'aide d'un tableau présentant les nombres de sinistres, par garanties et par années de survenance, selon qu'ils soient sans suite ou non[2].

1.1.3 Description des données

Nous disposons des données relatives aux garanties suivantes :

- Les garanties obligatoires ;
  - Responsabilité Civile matérielle (code : A),
  - Responsabilité Civile corporelle (code : A).
- Les garanties facultatives ;
  - Dommage (code : C),
  - Vol (code : E),
  - Incendie (code : D),
  - Bris de glace (code : K),
  - Recours (code : F),
  - Tempête (code : T),
  - Climatique (code : W),
  - Catastrophe Naturelle (code : M),
  - Attentat (code : S),
  - Acte Vandalisme (code : V),

1. *Incurred But Not Reported*. Le lecteur intéressé pourra se référer à un article du site Internet de l'Institut des Actuaires [3].
2. Ce tableau est présent en annexe .1, page 135.

  - Protection Juridique (code : J).

Le travail de modélisation et les applications vont se baser sur la sinistralité relative aux garanties suivantes :

- Les garanties obligatoires ;
  - Responsabilité Civile matérielle,
  - Responsabilité Civile corporelle.
- Les garanties facultatives ;
  - Dommage.

1.1.4 Présentation des variables

Chaque ligne de notre base de données correspond à l'état des règlements cumulés sur une garantie donnée d'un contrat sinistré.
Pour chaque ligne, nous disposons des variables suivantes :

- Le numéro de sinistre ;
- Le code de la garantie ;
- La date de survenance du sinistre ;
- La date de déclaration du sinistre ;
- La date de clôture du sinistre ;
- L'indicateur corporel/matériel ;
- L'indicateur sans suite ;
- L'indicateur de clôture ;
- Le code du département du sinistre ;
- Le montant des règlements cumulés au 31/12/2009 ;
- Le montant des frais d'experts et honoraires versés au 31/12/2009 ;
- Le montant des frais d'épave au 31/12/2009 ;
- Le montant des recours encaissés au 31/12/2009 ;
- Le stock de provision de sinistres au 31/12/2009 ;
- Le stock de prévisions de recours au 31/12/2009 ;
- Le montant des abandons de recours enregistrés au 31/12/2009.

1.1.5 Autres données techniques disponibles

Afin de mener à bien notre étude, les données suivantes ont été mises à disposition en complément de la base de données des sinistres :

- Le montant total des cotisations vues au 31/12/2009 pour les 17 exercices concernés des branches Vol, Responsabilité Civile, Dommage ;
- Les triangles (par années de survenance/développement) pour les garanties Dommage, Responsabilité Civile corporelle, Responsabilité Civile matérielle ; triangles des paiements, des provisions, des nombres de sinistres, des recours encaissés, des S/C brut de recours pour chaque exercice de 1993 à 2009.

Il sera intéressant d'utiliser ces triangles afin d'appliquer des méthodes comme *Chain Ladder*. Cette dernière permet d'obtenir une charge ultime (somme des règlements et de la PSAP) qui pourra ensuite être comparée à celle obtenue à partir de la méthode de provisionnement ligne à ligne mise en place dans ce mémoire.

1.2 Analyse des garanties étudiées

1.2.1 Stabilité des cotisations

Bien qu'il soit question dans ce mémoire de l'étude de la sinistralité et non de l'étude des cotisations, cette brève analyse des cotisations va nous permettre de connaître les principales évolutions de notre portefeuille au cours du temps.

En effet, à travers le chapitre 6, page 73, résultant d'une application, il est utile de savoir de quelle manière évolue les nombres de sinistres en fonctions des années de survenance sans avoir recours aux triangles relatifs aux nombres de sinistres qui présentent des anomalies [3]. De plus, il est raisonnable d'apparenter les évolutions des nombres de sinistres à celles des cotisations, à condition d'avoir un tarif stable dans le temps chez l'assureur.

Le but de cette section est de mettre en avant l'évolution des cotisations au cours des 17 exercices de survenance. Les deux graphiques *infra* décrivent l'évolution des cotisations pour les garanties Responsabilité Civile et Dommage.
Il est important de préciser que, sur le graphique suivant, les cotisations de la garantie RC regroupent les cotisations relatives aux garanties RC matérielle et corporelle.

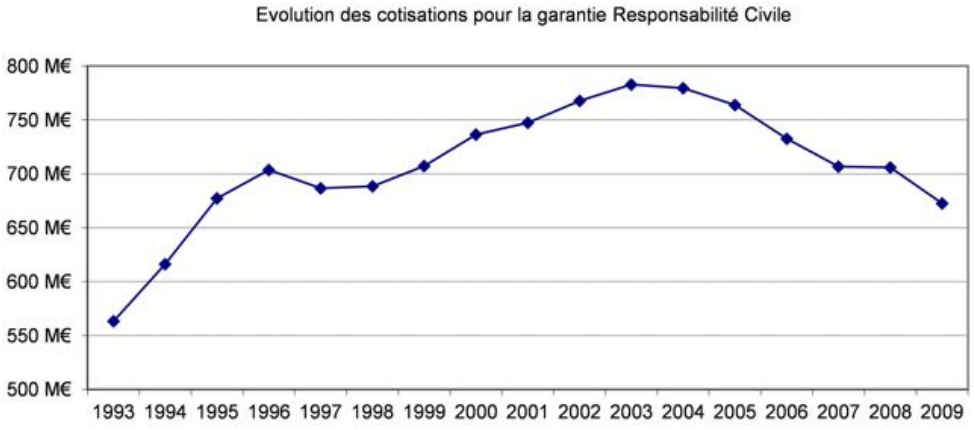


FIGURE 1.1 – Evolution des cotisations pour la garantie Responsabilité Civile

3. Ces anomalies sont détaillées chapitre 6, page 77, lors de l'application relative aux données étudiées.

D'après le graphique ci-dessus, on remarque que l'évolution des cotisations paraît cohérente avec les évolutions tarifaires relatives au marché de l'automobile français sur la même période. On constate notamment une diminution des cotisations depuis 2004 [4]. Toutefois, il n'est pas possible de conclure sur l'évolution du tarif. En effet, des cotisations peuvent baisser avec un tarif en hausse si l'assureur perd des clients.
Le portefeuille et le marché évoluent donc dans le même sens et d'après ces évolutions, le portefeuille d'assurés est resté globalement stable au cours du temps [5].

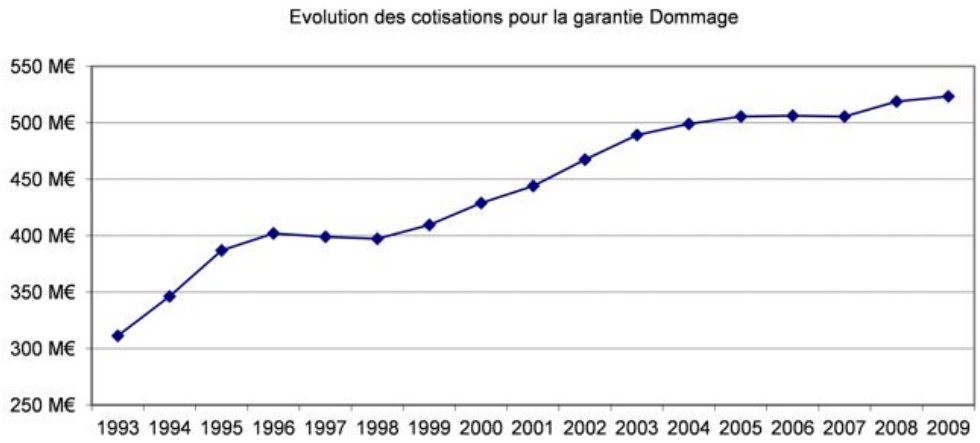


FIGURE 1.2 – Evolution des cotisations pour la garantie Dommage

La garantie Dommage est quant à elle facultative. Ainsi, il n'est pas évident de déterminer si l'évolution des cotisations est davantage liée à l'évolution du nombre d'assurés plutôt qu'à l'effet tarifaire.
Cependant, sur le marché automobile français, le poids des souscriptions de la garantie Dommage est resté globalement stable depuis 2005 alors que les tarifs ont augmenté. Ce constat semble être cohérent avec l'évolution observée sur les cotisations du portefeuille étudié. Afin d'illustrer ce propos, voici les taux de souscription des garanties facultatives pour les années 2005 à 2008.

Taux de souscription des garanties dans les contrats des véhicules de 1ère catégorie (1)
(en pourcentage du nombre de contrats)
Source : FFSA

	2005	2006	2007	2008
Vol et incendie	81	81	81	82
Bris de glaces	88	88	88	89
Dommages tous accidents	59	60	61	62
Poids des garanties facultatives dans le chiffre d'affaires automobile	**60**	**61**	**61**	**62**

(1) Véhicules 4 roues à moteur, de poids inférieur à 3,5 tonnes

FIGURE 1.3 – Taux de souscription des garanties facultatives

On constate que les garanties facultatives pèsent plus que les garanties obligatoires.

4. Source : site internet de la FFSA [2].
5. En réalité, le portefeuille a connu une légère croissance globalement régulière, de l'ordre de 1,6 % par an entre 1996 et 1999.

1.2.2 Responsabilité Civile dommages corporels

L'objet des graphiques présentés *infra* n'est pas de différencier les années de survenance. Cependant, ils permettent d'atteindre l'objectif recherché : dégager des tendances qui seront mises en évidence différemment lors de l'application réalisée dans le chapitre 6, page 73.

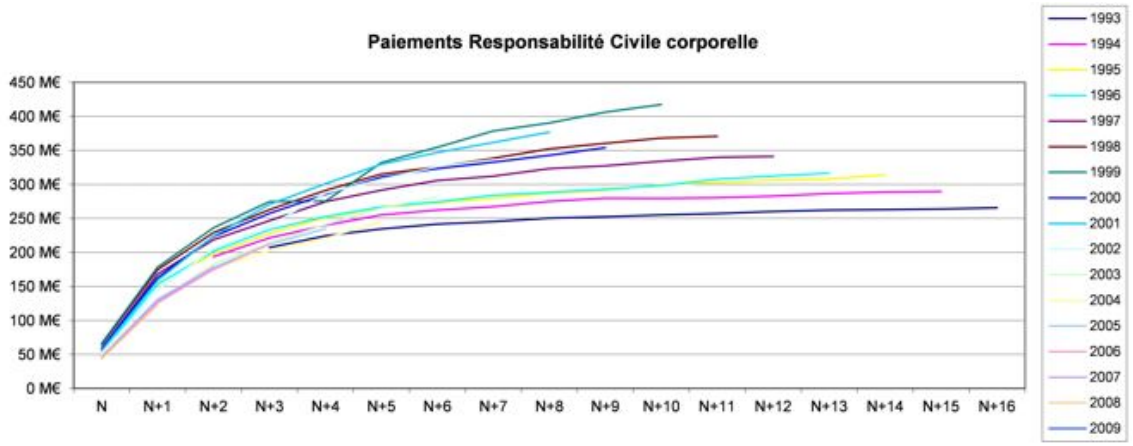


FIGURE 1.4 – Evolution des paiements pour la garantie Responsabilité Civile corporelle

Ce graphique met parfaitement en évidence le fait que la garantie Responsabilité Civile corporelle est une branche dite "longue", caractérisée par des délais de développement assez longs. En effet, pour les différentes années de survenance, on observe une stabilisation des paiements relativement lente au cours du temps.

1.2.3 Responsabilité Civile dommages matériels

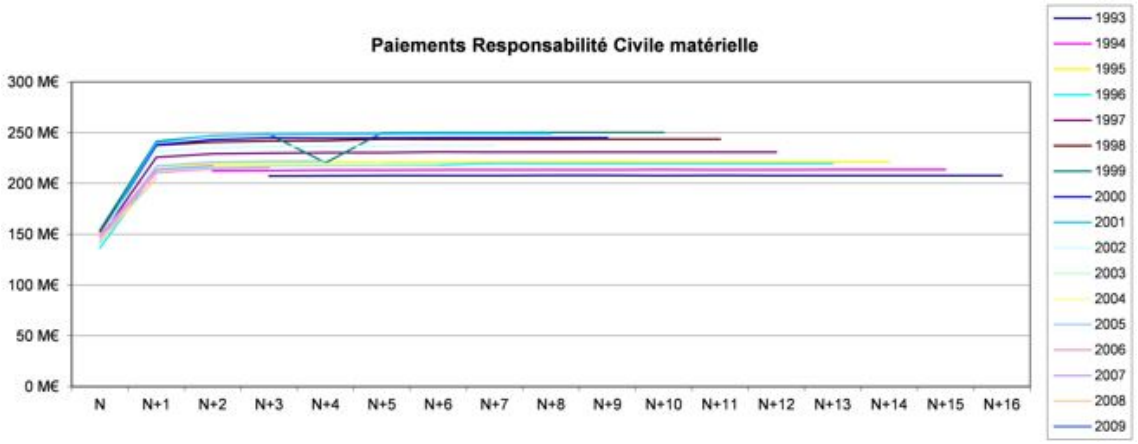


FIGURE 1.5 – Evolution des paiements pour la garantie Responsabilité Civile matérielle

Contrairement à la garantie Responsabilité Civile corporelle, nous sommes ici en présence d'une branche dite "courte". En effet, nous constatons que les sinistres sont réglés pour la plupart au bout d'un an, d'où une stabilisation rapide des paiements.

1.2.4 Dommage

De même que la garantie Responsabilité Civile matérielle, la garantie Dommage est une branche “courte”.

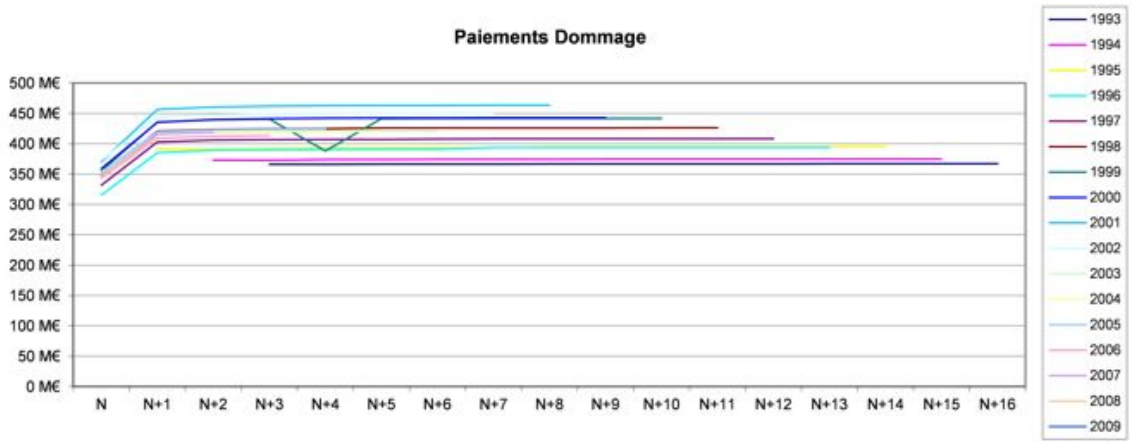


FIGURE 1.6 – Evolution des paiements pour la garantie Dommage

1.3 Conclusion

Nous pouvons conclure que le portefeuille considéré semble représentatif des évolutions observées sur le marché automobile français. Ce constat est utilisé dans les sections relatives aux applications afin d'effectuer des comparaisons et approximations en présence de données manquantes ou erronées.

Dépourvu d'information complémentaire sur le nombre exact d'assurés présents dans le portefeuille au cours des différentes années de survenance, nous ne pouvons étudier la stabilité du parc automobile assuré qu'à travers le volume de cotisations versées. Nous avons vu que ce dernier suit globalement les évolutions du marché et nous amène donc à penser que le portefeuille considéré est globalement stable.

De plus, sur les trois garanties étudiées, une d'entre elles est caractérisée de branche “longue”, la garantie Responsabilité Civile corporelle. Au cours des différentes études menées, il sera intéressant d'effectuer des comparaisons portant sur ces garanties et de distinguer les caractéristiques des branches “longues” et des branches “courtes”. Autrement dit, il sera pertinent de mener les études présentes dans ce mémoire en utilisant les différentes garanties comme variable d'hétérogénéité.

Les différences entre branches “courtes” et branches “longues” seront mises en évidence notamment au cours de l'application relative aux calculs de Provisions pour Sinistres A Payer (PSAP) [6] .

6. Cette notion sera définie dans le chapitre 6, page 73. Elle est également détaillée dans l'article de Magali KELLE [2007] [1].

Chapitre 2

Gestion de la base de données

2.1 Problématique d'une base de données volumineuse

Les données reçues étaient initialement contenues dans des fichiers texte. L'idée première était d'effectuer l'ensemble du travail à l'aide du logiciel R, seulement il n'a pas été possible d'importer l'ensemble des données sous R en raison du volume de la base de données. En effet, même s'il était possible d'importer l'ensemble des fichiers texte sous R, le logiciel (installé sur un ordinateur standard [1]) ne supportait pas la concaténation de ces fichiers permettant d'obtenir une base unique.

R ne supporte pas l'import de tables contenant plus de 2 millions de lignes environ, en considérant un ordinateur de performance moyenne avec 4 giga de RAM. Néanmoins, en augmentant les performances de l'ordinateur, l'ensemble du travail aurait pu être réalisé sous R.

La solution la plus évidente pour gérer une base de données aussi lourde que celle-ci a été l'utilisation du logiciel SAS. En effet, le recours à ce logiciel reconnu pour effectuer de la gestion de bases de données très conséquentes a permis d'importer les fichiers texte et ainsi concaténer l'ensemble pour obtenir une base de données SAS regroupant les sinistres des 17 années de survenance, soit 23 957 186 lignes. Les requêtes ont malgré tout été très coûteuses en temps de traitement.

2.2 Traitement des données sous SAS, modélisation et calculs sous R

Le fait d'avoir une seule et même base de données a permis de réaliser de nombreux tests de cohérence dans le but d'éliminer les valeurs aberrantes. Il a été nécessaire d'éliminer de la base les sinistres pour lesquels la date de survenance était supérieure à la date de déclaration ainsi que ceux dont la date de clôture était inférieure à la date de déclaration ou date de survenance.

En effectuant ce travail, 73 lignes aberrantes ont été éliminées (une ligne représentant l'état des règlements cumulés au 31/12/2009 sur une garantie donnée d'un contrat sinistré). De plus, les sinistres pour lesquels des montants négatifs étaient observés pour les variables suivantes ont été supprimés (47 lignes) :

1. Processeur Intel(R) Core(TM) 3,30 GHz, 4 Go de RAM et un système d'exploitation de 64 bits.

- règlements sinistres ;
- frais honoraires et prestations ;
- montant abandon recours ;
- montant de recours.

Ces lignes devaient certainement être des erreurs de saisie ou des annulations. Aucun doublon n'ayant été observé dans la base, le plus judicieux a été d'éliminer les lignes pour lesquelles l'une de ces quatre variables avait un montant négatif sachant qu'elles représentent un faible pourcentage des lignes totales qui constituent la base de données des sinistres (47 lignes sur 23 957 186).

Enfin, des incohérences ont été détectées dans la base au niveau de la variable $sin\_clos\_O\_N$ qui permet de connaitre l'état d'un sinistre (clos ou en cours). Ainsi, pour certains sinistres (nombre non négligeable) nous avons :

- $sin\_clos\_O\_N = O$ et une date de clôture non renseignée ;
- $sin\_clos\_O\_N = N$ et une date de clôture renseignée.

Après réflexion, il a été décidé, pour une ligne où une date de clôture est renseignée, de forcer $sin\_clos\_O\_N = O$ et pour une ligne où une date de clôture est non renseignée, de forcer $sin\_clos\_O\_N = N$ (en faisant cela, on part du principe que la variable date de clôture est bien renseignée contrairement à la variable $sin\_clos\_O\_N$).

Il faut noter que ce travail de préparation de la base de données, ainsi que l'élimination des données aberrantes, sont très importants pour réaliser un travail de modélisation sur des données les plus fiables possibles. En effet, le travail que l'on va effectuer par la suite sur les données, ainsi que les résultats que nous obtiendrons, vont être directement conditionnés par la qualité de notre base de données.

Un travail de transformation de la base de données a également été réalisé. En effet, pour l'étude de la structure de dépendance entre branches de risque, il a été nécessaire de transformer la base existante en une base de données qui apporte l'information sur les coûts cumulés associés à chaque garantie impactée par un sinistre donné. Dans cette nouvelle base, une ligne représente non plus l'état des règlements cumulés sur une garantie donnée d'un contrat sinistré mais l'état des règlements cumulés pour l'ensemble des garanties impacté par un sinistre donné.
Cette transformation de base a pu être réalisée grâce aux numéros des sinistres (variable clé de la base).

Une fois le traitement des valeurs aberrantes effectué et la totalité des bases créées, un export des données au format *.txt* a été réalisé :

- une base avec l'ensemble des données ;
- une base par garantie.

Certains modules comme "SAS IML", permettant le calcul intégral, n'étant pas disponibles au sein de la société, il a été décidé d'utiliser R pour réaliser le travail de modélisation à proprement dit. De plus, SAS est extrêmement performant pour la gestion de bases de données volumineuses. Toutefois, l'utilisation de R reste préférable pour effectuer du calcul

matriciel et intégral parce que ces calculs sont plus faciles à implémenter sous R.

Concernant l'implémentation des méthodes relatives à la théorie des copules, l'utilisation du package *copula*[2] a été très appréciée. Enfin, il faut noter que de nombreuses ressources bibliographiques sont disponibles pour le logiciel R[3]. Nous pouvons, par exemple, citer l'ouvrage édité par Arthur Charpentier[4] qui expose notamment des méthodes d'estimation pour la théorie des copules.

Ce qui est intéressant ici est d'avoir été confronté à devoir travailler avec deux logiciels complémentaires, ce qui permet d'appréhender les atouts et les limites de chacun d'entre eux.

2.3 Construction de statistiques *"as if"*

Les données de la base sinistres mises à disposition sont composées de montants de sinistres non revalorisés. Afin d'établir une certaine homogénéité entre les différents coûts de sinistres, il a été retenu de ramener tous les montants payés à une même date de calcul : l'année 2009.

Nous allons donc modifier les données pour constituer ce que l'on appelle une statistique *"as if"*, c'est-à-dire comparable à l'année d'évaluation. Il faut pour cela indexer les sinistres et leur appliquer les conditions d'assurance de l'année d'évaluation.

L'idée principale est qu'un sinistre de 100 000 euros en 1993 n'est certainement pas comparable à un sinistre de 100 000 euros en 2009. En effet, l'évolution économique et l'inflation ont un impact sur les coûts des sinistres. On va donc indexer les sinistres par rapport à un indice représentatif. L'indice des prix à la consommation[5] a été retenu pour effectuer ce travail.

On aurait également pu affiner le travail de revalorisation des sinistres en utilisant un indice spécifique à chaque garantie (par exemple l'indice du coût de la construction pour des sinistres incendies en assurance de particuliers). Seulement, ce travail aurait été très coûteux en temps et il aurait été délicat de trouver un indice approprié pour chacune des garanties constituant le portefeuille.

Soit :

- n l'année de cotation ;
- I_k l'indice de l'année k ;
- S_k la valeur d'un sinistre de l'année k ;
- S_k^n la valeur *"as if"* d'un sinistre de l'année k vu l'année n ;

Alors on a : $S_k^n = S_k \text{ x } \dfrac{I_n}{I_k}$

En guise d'exemple, on a :

$$\text{Sinistre } \textit{"as if"} = \text{Coût sinistre en 1993 x } \frac{\text{Indice Prix à la consommation 2009}}{\text{Indice Prix à la consommation 1993}}$$

2. L'utilisation de ce package est détaillée sur Internet et référencé dans la bibliographie : [4].
3. Le lecteur intéressé pourra se référer à un site Internet documenté [5].
4. Computational Actuarial Science with R [2015] [6].
5. Cet indice est disponible sur le site Internet de l'INSEE [7].

Notre base de données des sinistres ne nous permet pas de connaître les différents règlements associés aux différentes dates de règlement mais uniquement les montants cumulés au 31/12/2009. Ainsi, la correction ne pourra donc être que partielle.
Par exemple, pour un montant cumulé d'un sinistre clos au 31/12/2003, il s'agit de remettre en valeur de l'année 2009 le montant cumulé tandis que l'idéal serait de corriger les différents montants réglés chaque année pour les mettre en euros 2009 et ensuite en faire la somme.

Les statistiques *"as if"* sont ainsi obtenues pour chaque montant de la base sinistre. Ces montants correspondent à ce qui serait payé si les sinistres étaient survenus en 2009. C'est à partir de ces montants que l'ensemble du travail, ainsi que les statistiques descriptives seront réalisés. Le travail de préparation des données de la base sinistres est ainsi terminé.

En ce qui concerne les triangles de paiements, plusieurs solutions ont été envisagées après la consultation du mémoire de Ilan Habib et Stéphane Riban [2012] [8].
Pour être cohérent avec le travail réalisé sur les données ligne à ligne, il a été décidé de supposer que l'inflation future est égale à l'inflation passée, ce qui conduit à ne pas effectuer de retraitement sur les triangles.

Ces justifications et retraitements vont permettre de pouvoir comparer les PSAP obtenues par la méthode de *Chain Ladder* de celles relatives à l'application d'une méthode de provisionnement ligne à ligne[6].

6. Ce travail est réalisé au cours du chapitre 6, page 73.

Deuxième partie

Présentation théorique des modèles utilisés

Chapitre 3

Description du modèle sous-jacent à nos données

3.1 Décomposition de la charge totale X

Tout d'abord, il a été décidé de retenir pour l'évaluation des Provisions pour Sinistres A Payer, ainsi que pour l'étude de la structure de dépendance, un triplé de variables aléatoires correspondant aux coûts des branches Dommage, Responsabilité Civile corporelle et Responsabilité Civile matérielle.

Les garanties Responsabilité Civile corporelle et matérielle sont obligatoires contrairement à la garantie Dommage qui est facultative.
Afin d'étudier des garanties pour lesquelles l'assuré est couvert, il a été décidé de retenir uniquement les sinistres pour lesquels le coût relatif à la branche Dommage est strictement positif. Ainsi, nous sommes certains que d'une part, l'assuré a souscrit à la garantie Dommage et que d'autre part, ces sinistres ont occasionnés pour cette garantie un remboursement non nul de la part de l'assureur pour l'assuré.

En ce qui concerne les garanties Responsabilité Civile corporelle et matérielle, nous sommes également certains que l'assuré est couvert par ces garanties. Cependant, certains sinistres peuvent ne pas générer de frais pour l'une des deux garanties ou les deux en même temps.

Nous sommes donc en présence de deux types de variables aléatoires : une qui caractérise les coûts des sinistres relatifs à la garantie Dommage est une variable aléatoire continue. Les deux variables relatives aux coûts des sinistres des garanties Responsabilité Civile corporelle et matérielle sont des variables aléatoires ayant une composante discrète et une composante continue.
La composante discrète est en fait une masse de Dirac au point zéro et caractérise les sinistres pour lesquels aucun remboursement n'a été effectué. Il s'agit donc des sinistres qui n'ont pas impacté les garanties Responsabilité Civile corporelle et/ou Responsabilité Civile matérielle.

Il est donc possible de décomposer la charge totale X relative aux trois garanties étudiées en une charge X_1 relative à la garantie Dommage et des charges X_2 et X_3 relatives aux garanties Responsabilité Civile corporelle et matérielle et d'écrire : $X = X_1 + X_2 + X_3$.

3.2 Caractérisation de la loi de X_2 et de X_3

- La fonction de répartition de X_2 et de X_3 :

La fonction de répartition des charges liées aux garanties Responsabilité Civile corporelle et matérielle est définie de la manière suivante :

$$F_p(x) = p\,\delta_0(x) + (1-p)\overbrace{\int_0^x f_X(u)du}^{F_X(x)} \tag{3.1}$$

avec $F_p(0) = P(X=0) = p$ et $\mathbb{P}(X \leq x | X > 0) = \dfrac{\mathbb{P}(0 < X \leq x)}{1-p} = \int_0^x f_X(u)du.$

L'inverse généralisée de la fonction de répartition est définie de cette manière :

$$F_p^{-1}(y) = \inf\{x / F_p(x) \geq y\} \tag{3.2}$$

Il est possible d'estimer la fonction de répartition F dans un cadre paramétrique :

- $\hat{p} = \dfrac{\#(i/X_i = 0)}{n}$, $\hat{p}$ est égal au rapport entre le nombre de sinistres ayant un montant de règlement égal à zéro et le nombre de sinistres qui constitue l'échantillon.

- si $f = f_\theta$, $\hat{\theta}$ est égal à l'Estimateur du Maximum de Vraisemblance (EMV) pour l'échantillon $(X_i | X_i > 0)_{1 \leq i \leq n}$. Il s'agit donc de réaliser l'estimation des variables à partir des montants de sinistres strictement supérieurs à zéro seulement.

L'inverse généralisé de la fonction de répartition peut s'écrire de cette façon :

$$F_p^{-1}(y) = \begin{cases} 0 & \text{si } y \leq p, \text{ puisque } F_p(0) = p\ ; \\ F_X^{-1}\left(\dfrac{y}{1-p}\right) & \text{si } y > p, \text{ car pour avoir } F_p(x) \geq y, \text{ il faut que } x > 0. \end{cases} \tag{3.3}$$

L'écriture de l'inverse généralisé de la fonction de répartition va nous permettre de calculer les rangs lors de l'application relative à la modélisation de la structure de dépendance entre les coûts relatifs aux trois branches de risque étudiées [1].

3.3 Les modèles de censure

Ce modèle a pour objectif d'estimer la Provision pour Sinistres A Payer (PSAP) à l'aide d'une approche sinistre par sinistre. Contrairement à des méthodes se basant sur des données agrégées dans un triangle de liquidation, il s'agit ici de considérer les sinistres de façon individuelle afin d'estimer un montant de provision relatif à chacun d'entre eux pour ensuite les agréger et ainsi déterminer une estimation de la charge ultime.

1. Cette application est réalisée dans le chapitre 7, page 94.

3.3.1 Spécificités des données et hypothèses relatives aux modèles de censure

Lors de l'application classique d'un modèle de censure, des durées de vie sont étudiées. Dans notre cas, les durées de vie vont être assimilées à des montants de sinistres.
Il est nécessaire de justifier l'utilisation d'un modèle de censure sur nos données. Les montants de sinistres sont donc des variables positives, notées X, qui ont deux particularités :

- la caractérisation la plus intéressante de leur loi est le taux de risque ; c'est généralement le taux qui est modélisé ;
- la variable d'intérêt X n'est pas toujours observée, mais une autre variable C, appelée censure, donne l'information sur X. En effet, pour certains sinistres, l'évènement étudié (la clôture) ne se produit pas pendant la période d'observation et en conséquence, certaines données sont censurées.

On retrouve lors de l'application d'un modèle de durée, les notions de censures et de troncatures. Ces dernières proviennent du fait que l'on n'a pas accès à la totalité de l'information. En effet, plutôt que d'observer des réalisations indépendantes et identiquement distribuées (iid) de X, on observe la réalisation de la variable aléatoire X soumise à diverses perturbations, indépendantes ou non du phénomène étudié. Ces phénomènes peuvent, dans l'ensemble, être regroupés en deux grandes catégories : les censures et les troncatures [2].
Les données étudiées dans ce mémoire sont caractérisées par la présence d'une censure à droite uniquement.

En effet, nous sommes en présence d'un portefeuille particulier où le montant de franchise appliqué est le même pour chacune des trois garanties étudiées. Une étude des coûts relatifs à ces trois branches de risque a permis de vérifier cela. Il a donc été légitime, pour notre étude, de s'épargner la prise en compte de données tronquées à gauche [3].

- Censure à droite :

Soit C une variable aléatoire positive. Le montant de sinistre X est dit censuré à droite si, à la place de X, on observe C à la date d'observation et si on sait que $X > C$.
Dans le cas général, une donnée est censurée à droite si la seule information dont on dispose est qu'à une date donnée l'événement ne s'est pas encore produit. En considérant le cas particulier étudié dans ce mémoire, un sinistre est censuré à droite si la seule information dont on dispose est qu'à une date d'observation donnée, la clôture ne s'est pas encore produite. Ainsi, à la date d'observation, nous ne connaissons pas le coût unitaire total d'un sinistre mais seulement le montant cumulé déjà réglé à cette date.

3.3.2 Cadre de l'étude

Afin d'être en mesure d'appliquer un modèle de censure sur nos données, nous allons devoir disposer de la totalité des informations relatives aux sinistres. Autrement dit, nous allons devoir, à partir de nos données, retracer la vie complète des sinistres de l'ensemble de notre portefeuille. Notre base satisfait cette exigence puisque nous disposons des informations suivantes pour chaque sinistre :

2. Les troncatures se distinguent des censures dans le sens où elles concernent l'échantillonnage lui-même.
3. Il y a troncature à gauche dès lors que la variable d'intérêt, ici X, n'est pas observable lorsqu'elle est inférieure à un seuil donné $C > 0$ (une constante).

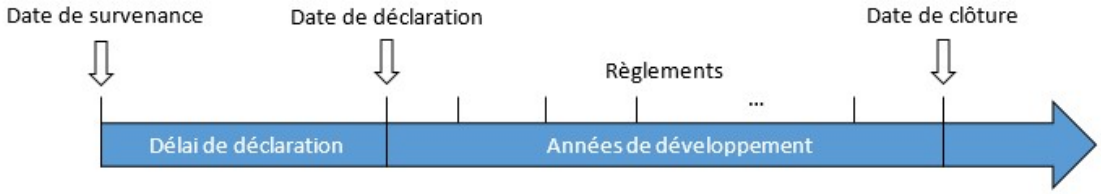


FIGURE 3.1 – Vie complète des sinistres

Notre base nous permet de connaître le montant cumulé des règlements pour chacun des sinistres à la date d'observation. Nous ne connaissons pas les dates et montants des différents règlements individuels.

L'utilisation d'une telle modélisation implique d'être capable de différencier les deux états possibles d'un sinistre : "clos" et "en cours".

- Les sinistres clos :

Les sinistres clos sont, par définition, caractérisés par une date de clôture antérieure à la date d'observation (date de notre étude : 31/12/2009).
Pour chaque sinistre clos, nous disposons de l'ensemble des informations ; à savoir : le montant cumulé des règlements concernant un sinistre (somme des montants réglés entre la date de déclaration et la date de clôture du sinistre) et la durée qui sépare la date de déclaration de la date de clôture. Les sinistres clos peuvent être représentés comme ceci :

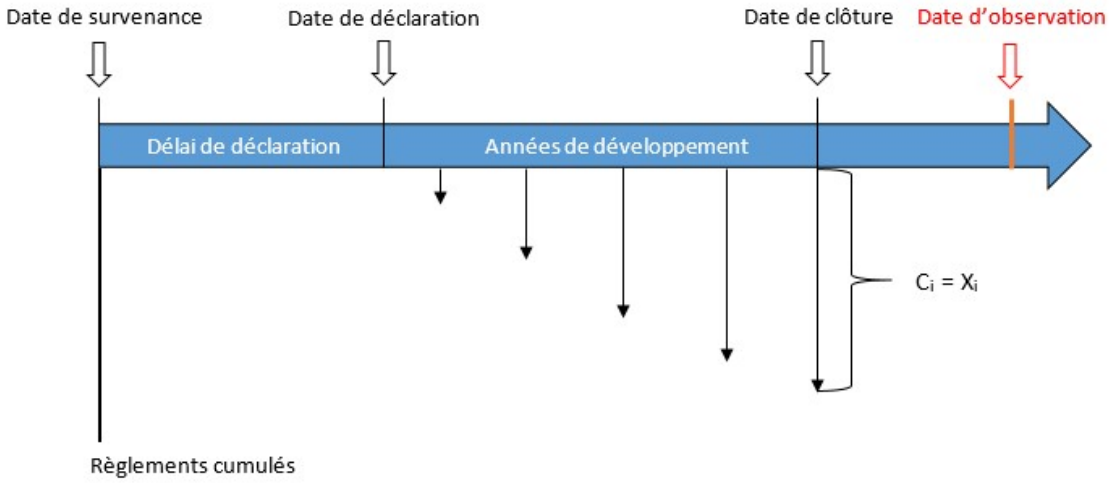


FIGURE 3.2 – Vie complète des sinistres clos avant la date d'observation

- Les sinistres en cours :

Les sinistres en cours sont, par définition, caractérisés par une date de clôture postérieure à la date d'observation. Nous ne connaissons donc pas encore la date de clôture pour ces sinistres. Dans ce cas, nous connaissons uniquement le montant cumulé des règlements effectués à la date d'observation C_i mais pas le coût total du sinistre X_i. La différence entre X_i et C_i constitue le montant de provision à estimer à l'aide de notre modèle de censure. Cette situation peut être représentée de la façon suivante :

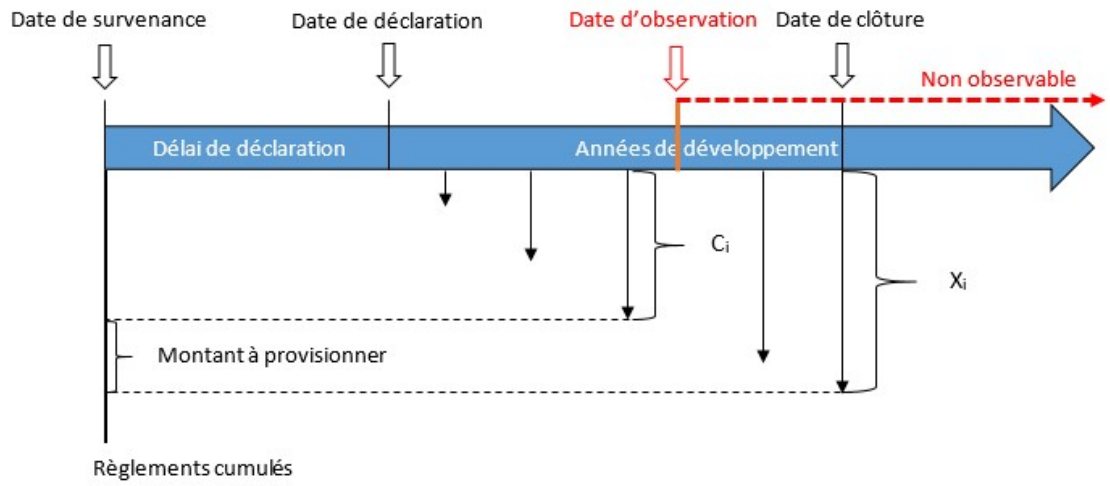


FIGURE 3.3 – Vie complète des sinistres encore en cours à la date d'observation

Pour les sinistres en cours, le coût total ainsi que la durée totale de développement sont inconnues, c'est-à-dire censurés.

3.3.3 Notations et définitions relatives aux modèles de censure

Ces notations seront utilisées tout au long du mémoire. Elles sont reprises des travaux de Noémie Rose [2009] [9], Sandra Gotlib [2012] [10] ainsi que du cours de Frédéric Planchet [2013-2014] [4] et de l'ouvrage de Frédéric Planchet et Pierre Thérond [2006] [11] :

- i le *ième* sinistre ;
- C_i le montant cumulé des règlements déjà versés pour le *ième* sinistre et connu à la date d'observation ;
- X_i le coût total du *ième* sinistre ;
- D_i l'indicatrice de l'état du *ième* sinistre (clos ou en cours) à la date de l'étude : $D_i = \mathbb{1}_{Date\ de\ clôture\ \leq\ Date\ d'observation} = \mathbb{1}_{X_i\ \leq\ C_i}$.

A partir de ces notations, il est possible de caractériser les sinistres d'un portefeuille selon qu'ils soient clos ou encore en cours :

Sinistres clos	**Sinistres en cours**
Date de clôture $\leq$ Date d'observation	Date de clôture $>$ Date d'observation
$D_i = 1$	$D_i = 0$
X_i connu	X_i inconnu
$X_i = C_i$	$X_i > C_i$

TABLE 3.1 – Tableau récapitulatif de l'état d'un sinistre : clos ou en cours

Tout d'abord, nos données contiennent une variable censurée : le coût total des sinistres. C'est à partir de cette constatation qu'il a été jugé nécessaire d'utiliser les modèles de

4. Support de cours ISFA : Modèles de durée.

censure afin d'estimer les coûts de nos sinistres non clos à la date d'observation. Les modèles de censure constituent un outil mis en œuvre dans de nombreux domaines de l'assurance vie afin de modéliser par exemple la durée de vie humaine, la durée de l'arrêt de travail. Ces modèles sont moins connus en assurance non-vie et c'est dans ce cadre qu'il sera utilisé.

Comme énoncé précédemment, ces modèles présentent plusieurs spécificités et reposent sur des hypothèses comme le caractère aléatoire ainsi que la positivité de la variable censurée étudiée. Une autre particularité de cette approche réside dans l'introduction des fonctions de survie et de hasard qui vont nous permettre de définir entièrement la loi de la variable d'étude (de la même façon que la fonction de densité et de répartition).

- **Fonction de densité :**

$f(x), x \in \mathbb{R}_+$ la densité de probabilité de X. $f(x)$ est définie comme la dérivée de $F(x)$. C'est la probabilité d'avoir un évènement entre x et $(x + dx)$:

$$f(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X < x + dx)}{dx} = F^{'}(x) = -S^{'}(x)$$

- **Fonction de répartition :**

$F(x) = \mathbb{P}(X \leq x) = \int_0^x f(t)\, dt$ la fonction de répartition de X, qui mesure la probabilité de "mourir" au plus tard en x. Dans notre situation, cette fonction mesure la probabilité que le coût total d'un sinistre soit inférieur ou égal à un montant x.

- **Fonction de survie :**

Sous condition que la densité de la variable X existe, nous pouvons définir la fonction de survie $S(x)$ de la manière suivante :

$$S(x) = \mathbb{P}(X > x) = 1 - F(x) = \int_x^{+\infty} f(t)\, dt$$

D'après cette définition, nous pouvons déduire l'égalité suivante : $S^{'}(x) = -f(x)$.
La fonction $S(x)$ mesure la probabilité de survivre en x et dans le cadre de notre, étude elle représente la probabilité que le coût total d'un sinistre soit supérieur à x. Notons que la fonction de survie $S(x)$ est décroissante et continue : $S(0) = 1$ et $S(+\infty) = 0$.

Chacune de ces trois fonctions caractérise entièrement la loi de X ; d'autres fonctions peuvent également spécifier la loi de X :

- **Fonction de survie conditionnelle :**

La fonction de survie conditionnelle $S_u(x)$ représente la probabilité que le montant total d'un sinistre soit supérieur à $x + u$ sachant que ce dernier est supérieur à u :

$$S_u(x) = \mathbb{P}(X > u + x | X > u) = \frac{\mathbb{P}(X > u + x)}{\mathbb{P}(X > u)} = \frac{S(u + x)}{S(u)} = \exp(-\int_u^{u+x} h(t)dt)$$

- **Fonction de hasard :**

La fonction de hasard $h(x)$, également appelé taux de risque instantané, représente la probabilité que le coût total d'un sinistre soit compris entre x et $x + dx$, sachant que ce dernier est supérieur à un seuil x :

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{S(x)} = -\frac{S^{'}(x)}{S(x)} = -\frac{d}{dx}\ln(S(x)) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X < x + dx | X \geq x)}{dx}$$

$h(x)$ est donc le taux instantané de clôture d'un sinistre pour un montant x (même si $h(x)$ n'est pas nécessairement inférieur à 1).

De cette définition, on déduit directement cette relation entre la fonction de survie et la fonction de hasard :

$$S(x) = \exp(-\int_0^x h(t)\,dt) \qquad \text{ou} \qquad h(x) = \frac{d[-\ln(S(x))]}{dt}$$

- **Fonction de hasard cumulée :**

$$H(x) = \int_0^x h(t)\,dt = -\ln[S(x)]$$

On obtient la relation suivante entre fonction de densité, fonction de hasard et fonction de hasard cumulée :

$$f(x) = h(x)\exp[-H(x)]$$

- **Espérance résiduelle :**

Soit $r()$ la durée de vie moyenne restante (*expected residual life*). Cette fonction représente l'espérance du montant qu'il reste encore à payer pour un sinistre donné, sachant que ce dernier a déjà atteint un montant c à la date d'observation.

$$r(c) = \mathbb{E}[X - c | X > c] = \frac{1}{S(c)} \int_c^{+\infty} S(t)dt$$

3.3.4 L'approche non paramétrique

L'estimateur de la fonction de survie :

L'estimateur de la fonction de survie le plus utilisé lorsqu'aucune hypothèse ne veut être faite, *a priori* sur la forme de la loi de survie, est l'estimateur de *Kaplan-Meier*.

Cet estimateur est donné par l'expression suivante :

$$\hat{S}(x) = \prod_{i | x_i \leq x} \left(1 - \frac{d_i}{n_i}\right) \tag{3.4}$$

avec d_i qui vaut 1 lorsque le sinistre est clos, 0 lorsque le sinistre est encore en cours (censure) et n_i le nombre de sinistres non clos à des montants x_i, c'est à dire les sinistres ayant un montant cumulé supérieur à x_i.

Le lecteur intéressé trouvera en annexe .2, page 138, une présentation heuristique de l'estimateur de la fonction de survie et en annexe .3, page 138, une définition plus formelle permettant d'obtenir la dérivation de l'estimateur de *Kaplan-Meier*.

Les principales hypothèses et leur signification :

Tout d'abord, il est fondamental de vérifier la validité des hypothèses portant sur la régularité des fonctions manipulées. Ensuite, deux points doivent absolument être pris en compte lors de l'utilisation de l'estimateur de *Kaplan-Meier* : le premier point concerne l'hypothèse de censure non informative et le second relève de l'homogénéité de la population étudiée [5].

La variance de l'estimateur de *Kaplan-Meier* :

Afin d'apprécier la précision de l'estimation effectuée de la fonction de survie $S(x)$, il est indispensable d'estimer la variance de l'estimateur $\hat{S}(x)$:

$$\text{Var}[\hat{S}(x)] = \hat{S}(x)^2 \times \hat{\sigma}^2_{lS_t} \quad avec \quad \hat{\sigma}^2_{lS_t} = \sum_{i|x_i \leq x} \frac{d_i}{(n_i - d_i)n_i} \tag{3.5}$$

La démonstration relative à la variance de l'estimateur de *Kaplan-Meier* est présentée en annexe .4, page 141.

De plus, la mise en place d'un Intervalle de Confiance sur la fonction de survie est très importante et sera réalisée dans la partie pratique de ce mémoire (au chapitre 6, page 78). De plus, une description détaillée de la construction de cet intervalle est présente en annexe .5, page 143.

3.3.5 L'approche paramétrique

Cette section s'inscrit dans la continuité du travail de Noémie Rose [9] ainsi que de celui de Sandra Gotlib [10].

Modélisation et calcul de la Provision pour Sinistres A Payer (PSAP) :

- Modélisation du coût relatif à un sinistre :

L'objectif des modèles paramétriques est d'évaluer la loi suivie par nos données et d'en estimer les paramètres. La première étape consiste à ajuster une loi à nos coûts de sinistres. La loi retenue est celle qui maximise la log-vraisemblance. Ensuite, à l'aide de la méthode du maximum de vraisemblance, les paramètres de la loi sont estimés et en sont déduits fonctions de survie et de hasard qui caractérisent complétement la loi des coûts de sinistres.

La variable X est désormais connue grâce à l'estimation de l'échantillon. Il faut ensuite évaluer la loi suivie par la provision individuelle relative à chaque sinistre.

- Modélisation de la provision individuelle relative à un sinistre :

Nous cherchons maintenant à estimer la loi de la provision individuelle pour permettre d'évaluer la provision totale. Par définition, la provision relative à un sinistre est égale à l'espérance de la charge résiduelle du sinistre, charge résiduelle définie par :

$$P = X - c | X \geq c \quad \text{avec} \quad c > 0 \tag{3.6}$$

5. Des détails et explications à propos de ces deux hypothèses sont présentés dans le cours de modèles de survie de Gilbert Colletaz [2012].

c étant le montant de sinistre déjà réglé à la date d'observation.
La loi de X étant connue, celle de $X - c$ l'est également et nous cherchons maintenant à connaître la loi conditionnelle de $X - c$ sachant que le coût total du sinistre est supérieur ou égal au montant déjà réglé pour ce sinistre. Il en vient que l'espérance de la variable P représente la provision à constituer pour un sinistre.

Chaque sinistre engendre une provision qui lui est propre. De ce fait, cette partie de l'étude se fait de façon individuelle et est caractérisée de méthode de provisionnement ligne à ligne. Il s'agit de déterminer la loi conditionnelle au dépassement d'un certain seuil par la variable d'intérêt X à l'aide de la fonction de survie conditionnelle. Soit c, le seuil dépassé par la variable X, nous avons alors :

$$\begin{aligned} S_{\theta,c}(x) &= \mathbb{P}(X > c + x | X > c) \\ &= \frac{\mathbb{P}(X > c + x)}{\mathbb{P}(X > c)} \\ &= \frac{S_\theta(c + x)}{S_\theta(c)} \end{aligned} \tag{3.7}$$

De plus, la fonction de survie de la provision individuelle relative à un sinistre $P = X - c | X > c$ que l'on note S_α est définie par :

$$\begin{aligned} S_\alpha(p) &= \mathbb{P}((X - c | X > c) > p) \\ &= \mathbb{P}(X - c > p | X > c) \\ &= \frac{\mathbb{P}(X > p + c)}{\mathbb{P}(X > c)} \\ &= S_{\theta,c}(p) \end{aligned} \tag{3.8}$$

D'après ces définitions, la fonction de survie de la provision individuelle à constituer correspond à la fonction de survie du coût d'un sinistre conditionnellement au fait que ce coût dépasse un seuil qui est en fait la somme des règlements déjà effectués pour ce sinistre.

- Modélisation de la provision totale relative à l'ensemble des sinistres :

La loi de la provision individuelle d'un sinistre est maintenant supposée connue, il reste à établir celle de la provision totale appelée Provision pour Sinistres A Payer (PSAP). Afin de connaître le montant de la provision pour l'ensemble du portefeuille, il s'agit dans un premier temps de déterminer la loi suivie par cette somme pour ensuite en prendre l'espérance.

La Provision pour Sinistres A Payer est égale à l'espérance de la charge résiduelle totale à payer. Cette charge est une somme de variables aléatoires indépendantes mais non identiquement distribuées :

$$\Lambda = \sum_{i \in I_e} P_i \tag{3.9}$$

avec I_e l'ensemble des sinistres en cours dans le portefeuille à la date d'observation (sinistres censurés pour lesquels $D_i = 0$).

En fonction de la loi de la provision individuelle, différentes méthodes peuvent être envisagées dans le but de déterminer la distribution de Λ :

- soit il est possible d'effectuer un calcul explicite (c'est par exemple le cas si les provisions individuelles suivent des lois Normales) ;
- soit nous nous situons dans le domaine d'application du Théorème Central Limite (la condition de Lyapounov ou de Lindeberg étant supposée satisfaite [6]).

Dans le premier cas, le calcul de l'espérance et de la variance de P (lorsque ces deux quantités existent) est possible à partir de la fonction de survie associée :

$$E_{\theta,c}(P) = \int_0^\infty S_{\theta,c}(p)dp \tag{3.10}$$

$$V_{\theta,c}(P) = 2\int_0^\infty pS_{\theta,c}(p)dp - E_{\theta,c}(P)^2 \tag{3.11}$$

Ainsi, l'approximation de la loi de Λ, lorsqu'elle est réalisable, est facile à mettre en oeuvre. La provision totale peut alors être estimée directement par :

$$E(\Lambda) = \sum_{i\in I_e} E(P_i) \tag{3.12}$$

avec comme variance :

$$V(\Lambda) = \sum_{i\in I_e} V(P_i) \tag{3.13}$$

Lorsque l'approximation normale est valide, les paramètres de la loi limite peuvent se déduire directement de façon numérique.

Dans le second cas, le Théorème Central Limite montre que toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire Gaussienne. Dans notre cas, les variables cumulées, correspondant aux provisions individuelles, sont bien indépendantes mais non identiquement distribuées. Cette dernière hypothèse n'est pas nécessaire si certaines conditions, vérifiant qu'aucune variable n'exerce une influence significativement plus importante que les autres, existent.

Ce sont les conditions de Lindeberg et de Lyapounov qui permettent la généralisation du théorème ne nécessitant plus que les variables soient distribuées selon une même loi. Dans le cadre de ce mémoire, les provisions individuelles ne sont pas identiquement distribuées. Il est donc primordial que ces deux dernières conditions soient satisfaites afin d'être en mesure d'appliquer le Théorème Central Limite.

Condition de Lyapounov et de Lindeberg :

- La condition de Lyapounov :

L'objectif de cet énoncé est de pouvoir connaitre la loi de la somme des provisions individuelles.

Soit $(P_1, ..., P_2)$ la suite de variables correspondant aux provisions individuelles, μ_i l'espérance finie de P_i et σ_i son écart-type fini, il est alors possible de définir :

$$s_n^2 = \sum_{i=1}^n \sigma_i^2 \tag{3.14}$$

6. Ces deux conditions sont énoncées ci-après.

De la même façon, en supposant que les moments d'ordre 3 sont finis pour tout n, il est noté :

$$r_n^3 = \sum_{i=0}^{n} E\left(|X_i - \mu_i|^3\right) \tag{3.15}$$

La condition de Lyapounov qui doit être vérifiée est la suivante :

$$\lim_{n\to\infty} \frac{r_n}{s_n} = 0 \tag{3.16}$$

A partir du moment où cette condition est vérifiée, il est légitime d'appliquer le Théorème Central Limite. Soit $S_n = X_1 + X_2 + ... + X_n$ la somme de variables aléatoires de variance s_n et d'espérance $m_n = \sum_{i=0}^{n} \mu_i$, on a alors :

$$\frac{S_n - m_n}{s_n} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1) \tag{3.17}$$

- La condition de Lindeberg :

La condition de Lindeberg est plus faible que celle de Lyapounov. En utilisant les mêmes notations et définitions que celles utilisées pour la condition de Lyapounov, la condition de Lindeberg s'écrit de la manière suivante :

$$\lim_{n\to\infty} \sum_{i=1}^{n} E\left(\left\{ \frac{(X_i - \mu_i)^2}{s_n^2} \;\middle|\; |X_i - \mu_i| > \epsilon s_n \right\}\right) \forall\, \epsilon > 0 \tag{3.18}$$

A partir du moment où cette condition est vérifiée, le Théorème Central Limite peut être appliqué et la convergence suivante est obtenue :

$$\frac{S_n - m_n}{s_n} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1) \tag{3.19}$$

Pour conclure, lorsque l'une des deux conditions énoncées *supra* est vérifiée, le Théorème Central Limite peut s'appliquer et de ce fait, la loi de la somme totale des provisions est connue.

La loi statistique utilisée dans l'application de ce mémoire : la loi de Weibull :

Il est supposé dans cette partie que la loi suivie par les coûts des sinistres X est une loi de Weibull de paramètres (α, λ).

La loi de Weibull est souvent caractérisée comme une alternative intéressante aux lois de coûts classiques comme les lois Log-Normale ou Gamma. De plus, la loi de Weibull est souvent utilisée dans le cadre de l'application des modèles de censure.

- Présentation de la loi de Weibull :

Si X est une variable aléatoire suivant une loi de Weibull de paramètre (α, λ) alors X admet comme fonction de répartition la fonction suivante :

$$F_{\alpha,\lambda}(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right) \text{ avec :} \tag{3.20}$$

$\alpha > 0$, le paramètre de forme et $\lambda > 0$, le paramètre d'échelle de la distribution. Lorsque $\alpha = 1$, la loi de Weibull est en fait une loi Exponentielle.

La fonction de densité de X est définie par :

$$f_{\alpha,\lambda}(x) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right) \tag{3.21}$$

Le graphique ci-dessous permet d'illustrer le comportement de la fonction de répartition en faisant varier les paramètres de forme et d'échelle :

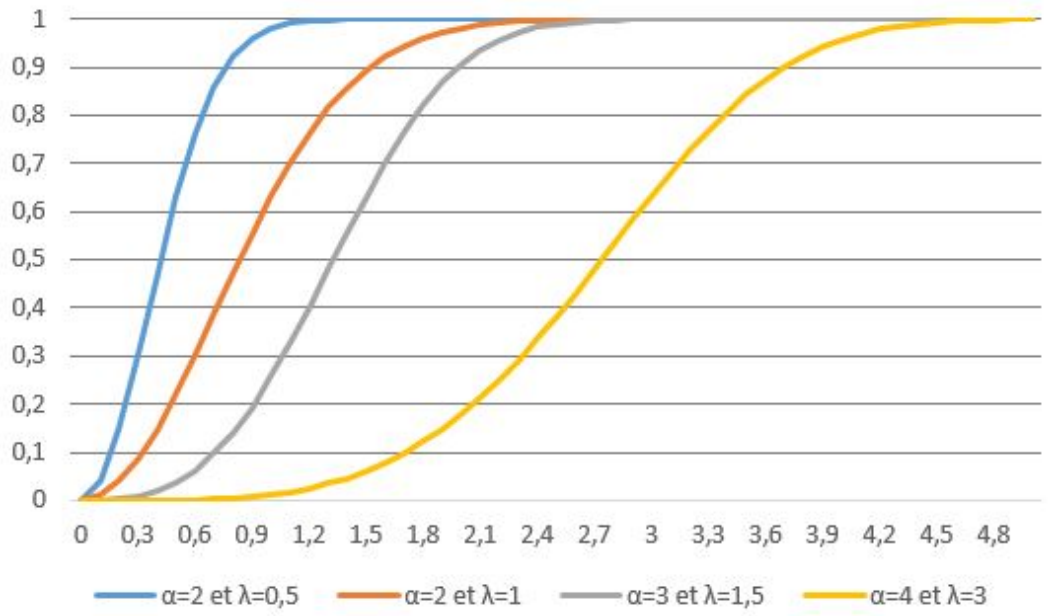


FIGURE 3.4 – Représentation de différentes fonctions de répartition de loi de Weibull

La fonction de répartition croit plus rapidement pour des paramètres de faibles valeurs. Graphiquement, on remarque que le paramètre α détermine la courbure de la fonction de répartition alors que le paramètre λ engendre une croissance vers 1 plus ou moins forte (plus λ diminue et plus la croissance vers 1 est forte).

Enfin, l'espérance de X si X suit une loi de Weibull est égale à :

$$\mathbb{E}(x) = \lambda\Gamma\left(1+\frac{1}{\alpha}\right) \tag{3.22}$$

- La loi de Weibull appliquée aux coûts des sinistres :

Soit X, la variable aléatoire représentant le montant cumulé relatif à un sinistre. Dans le cadre de notre étude, les fonctions de survie et de hasard permettent de caractériser entièrement la loi de la variable aléatoire étudiée, X admet ici la fonction de survie suivante :

$$S_{\alpha,\lambda}(x) = \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right) \tag{3.23}$$

et la fonction de hasard :

$$h_{\alpha,\lambda}(x) = -\frac{d}{dx}\ln\left(S_{\alpha,\lambda}(x)\right) = \alpha\frac{x^{\alpha-1}}{l^{\alpha}} \tag{3.24}$$

A partir de ces deux fonctions et de la méthode du maximum de vraisemblance, il est possible d'estimer les paramètres relatifs à ces fonctions tout en étant en présence de

données censurées, ce qui permet de définir complètement la loi.

- L'estimation des paramètres :

Nous avons vu précédemment, à la section 3.3.4, page 32, que la vraisemblance, dans le cadre de données censurées appliquée aux coûts des sinistres, peut s'écrire sous cette forme :

$$L(\alpha,\lambda) \propto \prod_{i=1}^{n}[f(x_i)^{d_i}S(x_i)^{1-d_i}] \tag{3.25}$$

En notant $d=\sum_{i=1}^{n} d_i$ le nombre de sorties observées (nombre de données non censurées), l'expression de la vraisemblance si X suit une loi de Weibull est égale à :

$$\begin{aligned} L(\alpha,\lambda) &\propto \left(\frac{\alpha}{\lambda^{\alpha}}\right)^{d}\prod_{i=1}^{n}x_i^{(\alpha-1)d_i}\exp\left(-d_i\left(\frac{x_i}{\lambda}\right)^{\alpha}\right)\exp\left(-(1-d_i)\left(\frac{x_i}{\lambda}\right)^{\alpha}\right) \\ &\propto \left(\frac{\alpha}{\lambda^{\alpha}}\right)^{d}\exp\left(-\lambda^{-\alpha}\sum_{i=1}^{n}x_i^{\alpha}\right)\exp\left((\alpha-1)\sum_{i=1}^{n}d_i\ln x_i\right) \end{aligned} \tag{3.26}$$

A partir de cette dernière expression, on obtient facilement la fonction de log-vraisemblance appliquée au modèle de Weibull en considérant la présence de censures :

$$\ln L(\alpha,\lambda) = \ln k + d(\ln\alpha - \alpha\ln\lambda) - \lambda^{-\alpha}\sum_{i=1}^{n}x_i^{\alpha} + (\alpha-1)\sum_{i=1}^{n}d_i\ln x_i \tag{3.27}$$

On en déduit les équations des dérivés partielles par rapport aux deux paramètres α et λ :

$$\begin{cases} \dfrac{\partial}{\partial\lambda}\ln L(\alpha,\lambda) &= -\dfrac{d}{\lambda} + \alpha\lambda^{-\alpha-1}\sum_{i=1}^{n}x_i^{\alpha} \\ \dfrac{\partial}{\partial\alpha}\ln L(\alpha,\lambda) &= d\left(\dfrac{1}{\alpha}-\ln\lambda\right) + \lambda^{-\tau}\left(\ln\lambda\sum_{i=1}^{n}x_i^{\alpha} - \sum_{i=1}^{n}x_i^{\alpha}\ln x_i\right) + \sum_{i=1}^{n}d_i\ln x_i \end{cases}$$

Afin de déterminer les expressions des paramètres, ces dérivées partielles sont supposées égales à zéro, la résolution de ces équations donne les expressions suivantes pour les deux paramètres recherchés :

$$\begin{cases} \lambda &= \left(\dfrac{1}{d}\sum_{i=1}^{n}x_i^{\alpha}\right)^{\frac{1}{\alpha}} \\ \dfrac{1}{\alpha} &= \dfrac{\sum_{i=1}^{n}x_i^{\alpha}\ln x_i}{\sum_{i=1}^{n}x_i^{\alpha}} - \dfrac{1}{d}\sum_{i=1}^{n}d_i\ln x_i \end{cases}$$

On ne peut pas résoudre ces équations à la main, il est nécessaire de passer par une méthode de résolution numérique comme celle d'écrite par l'algorithme de Newton-Raphson[7].

Pour la suite de l'étude, les estimateurs $\hat{\alpha}$ et $\hat{\lambda}$ sont supposés connus. De plus, si la loi de Weibull permet un ajustement de qualité des coûts de sinistres relatifs à une branche de sinistres donnée, alors la loi suivie par ces coûts est par conséquent complètement caractérisée.

7. Cette méthode est détaillée sur le blog d'Arthur Charpentier [12].

- Calcul de la provision individuelle :

Dans la section précédente, l'expression de la fonction de survie de la provision individuelle a été déterminée en fonction de la fonction de survie relative au coût du sinistre et du montant x_i déjà réglé à la date d'observation pour le sinistre i.

Comme défini précédemment, p est la réalisation de la variable relative à la provision individuelle. Pour chaque sinistre i, la provision est caractérisée de la façon suivante : $P = X - c|X \geq c$, avec $S_{\alpha,\lambda,c}(p)$ la fonction de survie associée :

$$\begin{aligned} S_{\alpha,\lambda,c}(p) &= \frac{S_{\alpha,\lambda}(c+p)}{S_{\alpha,\lambda}(c)} \\ &= \exp\left(-\left(\frac{c+p}{\lambda}\right)^{\alpha} + \left(\frac{c}{\lambda}\right)^{\alpha}\right) \end{aligned} \tag{3.28}$$

A partir cette expression, la variable P est complètement caractérisée et le montant de provision individuelle de chaque sinistre i peut être déterminé. Il s'agit de l'espérance de la variable aléatoire P qui peut s'écrire de cette manière à partir de la fonction de survie :

$$\begin{aligned} \mathbb{E}_{\alpha,\lambda,c}(P) &= \int_0^{\infty} S_{\alpha,\lambda,c}(p)dp \\ &= \int_0^{\infty} \exp\left(-\left(\frac{c+p}{\lambda}\right)^{\alpha} + \left(\frac{c}{\lambda}\right)^{\alpha}\right) dp \end{aligned} \tag{3.29}$$

Il est possible de résoudre cette intégrale à l'aide de la fonction Gamma d'Euler notée $\Gamma(\theta)$ et de la fonction Gamma incomplète notée $\Gamma(\theta, t)$:

$$\Gamma(\theta) = \int_0^{\infty} x^{\theta-1}\exp(-x)dx \tag{3.30}$$

$$\Gamma(\theta, t) = \int_0^{t} x^{\theta-1}\exp(-x)dx \tag{3.31}$$

On effectue un premier changement de variable : $u = \dfrac{c+p}{\lambda}$ et donc $p = lu - c$; $du = \dfrac{dp}{\lambda}$ permettant d'obtenir cette expression :

$$\mathbb{E}_{\alpha,\lambda,c}(P) = l\exp\left(\left(\frac{c}{\lambda}\right)^{\alpha}\right)\int_{\frac{c}{\lambda}}^{\infty} \exp(-u^{\alpha})du \tag{3.32}$$

Puis, un second changement de variable : $v = u^{\alpha}$ et donc $u = v^{\frac{1}{\alpha}}$; $dv = \alpha u^{\alpha-1}du$ et donc $du = \dfrac{dv}{\alpha v^{\frac{\alpha-1}{\alpha}}}$ permet d'écrire :

$$\begin{aligned} \mathbb{E}_{\alpha,\lambda,c}(P) &= \frac{\lambda}{\alpha}\exp\left(\left(\frac{c}{\lambda}\right)^{\alpha}\right)\int_{\left(\frac{c}{\lambda}\right)^{\alpha}}^{\infty} \exp(-v)\, v^{\frac{1}{\alpha}-1}dv \\ &= \frac{\lambda}{\alpha}\exp\left(\left(\frac{c}{\lambda}\right)^{\alpha}\right)\left[\int_{\left(\frac{c}{\lambda}\right)^{\alpha}}^{0} \exp(-v)\, v^{\frac{1}{\alpha}-1}dv + \int_0^{\infty} \exp(-v)\, v^{\frac{1}{\alpha}-1}dv\right] \\ &= \frac{\lambda}{\alpha}\exp\left(\left(\frac{c}{\lambda}\right)^{\alpha}\right)\left[\Gamma\left(\frac{1}{\alpha}\right) - \int_0^{\left(\frac{c}{\lambda}\right)^{\alpha}} \exp(-v)\, v^{\frac{1}{\alpha}-1}dv\right] \\ &= \frac{\lambda}{\alpha}\exp\left(\left(\frac{c}{\lambda}\right)^{\alpha}\right)\left[\Gamma\left(\frac{1}{\alpha}\right) - \Gamma\left(\frac{1}{\alpha}, \left(\frac{c}{\lambda}\right)^{\alpha}\right)\right] \end{aligned} \tag{3.33}$$

En pratique, lors de l'application détaillée au chapitre 6, page 73, l'estimation des paramètres a été réalisée avec le logiciel R. Il est ensuite possible d'implémenter et de résoudre cette équation en soumettant à cette dernière l'ensemble des coûts c_i relatifs à chaque sinistre i encore en cours à la date d'observation.

- Calcul de la provision totale :

L'espérance de la Provision pour Sinistres A Payer (PSAP) globale correspond à la somme des provisions individuelles et est définie la façon suivante :

$$\mathbb{E}(\Lambda) = \sum_{i \in I_e} \mathbb{E}(P_i) \tag{3.34}$$

Chapitre 4

La théorie relative aux copules

Une copule est une fonction qui permet de lier les lois marginales de variables aléatoires afin d'en générer la loi jointe [1]. L'idée est en fait de séparer la structure de dépendance entre les variables induites par la copule, des lois marginales.

La théorie des copules est apparue au cours des années 1950 avec les travaux de M. Fréchet dédiés aux marges de fonctions de répartition multi-variées. Quelques années plus tard, en 1959, le théorème de Sklar en découle. Ce théorème est le point central de la théorie des copules puisqu'il permet de séparer les marginales de la structure de dépendance.

Le but de cette partie théorique est d'apporter aux lecteurs les connaissances suffisantes et nécessaires afin d'appréhender l'application qui en découle au chapitre 7, page 94. Cette théorie a d'ores et déjà été utilisée dans de nombreux mémoires, dont celui de Valérie Lagenebre [2009] [13], de Kamal Armel [2010] [14], de Belguise Olivier [2001] [15], de Esterina Masiello [2010] [16] et de Gwladys Marylou Noutong [2013] [17].

De nombreux supports de cours ont également contribués à la rédaction de ce chapitre dont ceux de Frédéric Planchet [2010-2011] [2], Stéphane Loisel [2013-2014] [3] et Esterina Masiello [2013-2014] [4].

4.1 L'intérêt de modéliser la dépendance entre branches de risque

4.1.1 Le modèle Gaussien et ses limites

La distribution Normale multi-variée ou modèle Gaussien a, pendant longtemps, été considéré comme une référence pour modéliser la dépendance entre branches de risques. La dépendance entre deux variables peut être quantifiée par le coefficient de corrélation linéaire de Pearson. L'expression de la distribution Normale multi-variée étant connue, la mise en place de ce modèle est donc relativement aisée.

La principale limite de ce modèle est que chaque branche de risque est supposée suivre une loi Normale et cela est la conséquence de la Normale multi-variée. Par exemple, en assurance automobile, les fréquences des sinistres suivent généralement une loi de Poisson tandis que les coûts suivent une loi de Weibull, Gamma ou encore Log-Normale. Ainsi, dans

1. Le mot copule vient du mot latin *copula* qui signifie "lien".
2. Dépendance stochastique, introduction à la théorie des copules.
3. Mesures de corrélation et dépendance stochastique en assurance-finance.
4. Compléments du cours d'assurance non vie : Inférence statistique des copules.

ce mémoire et en règle générale, il sera nécessaire et intéressant d'évoluer vers l'utilisation de nouveaux modèles afin de ne plus se restreindre à l'utilisation de la loi Normale pour modéliser les coûts des sinistres.

4.1.2 L'intérêt de l'utilisation de nouveaux modèles

De récentes études démontrent clairement la nécessité de prendre en compte la dépendance entre branches de risques dans les domaines de l'assurance et de la finance. En effet, il a été démontré que supposer l'indépendance induisait des erreurs non négligeables. Nous pouvons citer quelques exemples les plus parlants :

- En assurance vie : lors de la tarification d'un contrat de rentes viagères sur un individu de sexe masculin avec réversion sur la conjointe, il est supposé que les deux individus sont indépendants. Néanmoins, Jagger et al. [1991] et Frees et al. [1996] [5] ont démontré que supposer l'indépendance dans ce cas de figure conduit à des erreurs lors de la tarification. Ces erreurs sont dues à l'effet du syndrome du "coeur brisé" qui conduit à diminuer l'espérance de vie résiduelle de la conjointe dans le cas où le conjoint décède.
- En assurance non vie : Olivier Belguise [2001] [15] a illustré dans son mémoire le fait qu'il existe de la dépendance entre les branches automobile et incendie lors de la survenance de tempêtes de forte intensité, ce qui n'est pas le cas sur les tempêtes de faible intensité.
- En finance : Olivier Junior Karusisi [2007] [19] a démontré que négliger la dépendance entre les défauts dans le calcul du capital économique d'une entité donnée engendre des erreurs significatives.

De nombreux chercheurs ont développé de nouveaux modèles en parcourant les travaux théoriques liés à la théorie des probabilités. En effet, la théorie des copules est née et a été développée ces dernières années en se basant sur les travaux de M. Fréchet, publiés initialement au cours des années 1950. Les points théoriques, en lien avec l'étude de la dépendance entre branches de risques développée dans ce mémoire, seront détaillés dans la suite de ce chapitre.

4.2 Notions fondamentales relatives aux fonctions de répartition

Seules les principales notions seront abordées dans cette section. Néanmoins, le lecteur intéressé pourra se référer à Jean Jacod [2001-2002] [6]. Il est supposé par la suite que l'on se situe dans un espace probabilisé (Ω, A, P) sur $\mathfrak{R}$.

4.2.1 Notions sur les fonctions de répartition uni-variées

- **Définitions**

Soit Y une variable aléatoire réelle. La fonction $F : \mathfrak{R} \to [0,1]$ est la fonction de répartition de Y, notée F_Y et définie par $F(y) = \mathbb{P}(Y \leq y), \forall y \in \mathfrak{R}$. F est continue à droite

5. Cf. [18].
6. Cours de "Théorie de l'intégration". http://www.proba.jussieu.fr/cours/Integr01.pdf.

avec des limites à gauche.

Soit F^{-1}, l'inverse généralisée de F, une fonction croissante continue à gauche et définie sur]0,1] par :

$$F^{-1}(\alpha) = \inf\{y \in \mathfrak{R}, F(y) \geq \alpha\} \tag{4.1}$$

L'inverse généralisée et l'inverse sont des notions équivalentes seulement si la fonction F est bijective de $\mathfrak{R}$ dans $]0,1[$.

C'est l'inverse généralisée de la fonction de répartition qui permet d'introduire la notion de quantile. Pour $\alpha \in]0,1]$, le quantile ou fractile d'ordre α de la loi de Y est en fait la quantité $F^{-1}(\alpha)$.

- **Propriétés**

F est croissante, continue à droite et admet les limites 0 en $-\infty$ et 1 en $+\infty$. Voici les principales propriétés vérifiées par F :

- L'inverse généralisée F^{-1} est croissante, continue à gauche et on a : $F(y) \geq \alpha \Leftrightarrow y \geq F^{-1}(\alpha)\, \forall y \in \mathfrak{R}$ et $\forall \alpha \in]0,1]$.

- $F(F^{-1}(\alpha)) \geq \alpha, \forall \alpha \in]0,1]$ et $F(F^{-1}(\alpha)) = \alpha, \forall \alpha \in]0,1]$ si $F^{-1}(\alpha) > -\infty$ et si F est continue en $F^{-1}(\alpha)$.

- F est la fonction de répartition de la variable aléatoire $F^{-1}(U)$ avec U est une variable aléatoire distribuée uniformément sur [0,1].

- Si F est continue, alors $F(Y)$ est distribuée uniformément sur [0,1] avec Y une variable aléatoire à valeurs dans $\mathfrak{R}$ de fonction de répartition F.

4.2.2 Notions sur les fonctions de répartition multi-variées

La fonction de répartition multidimensionnelle d'une variable aléatoire vectorielle permet de caractériser sa loi. Le raisonnement suivant sera utilisé dans suite de la définition des fonctions de répartition multi-variées : pour $d \geq 1, a = (a_1, ..., a_d)$ et $b = (b_1, ..., b_d)$ dans $\mathfrak{R}^d$, on a donc $a \leq b$ si $a_i \leq b_i$, $\forall\ 1 \leq i \leq d$.

Soient $d \geq 1$ et X une variable aléatoire réelle à valeurs dans $\mathfrak{R}^d$. La fonction de répartition de X est la fonction $F : \mathfrak{R}^d \to [0,1]$, que l'on peut noter F_X et définie par $F(x) = \mathbb{P}(X \leq x)$, $\forall\ x \in \mathfrak{R}^d$.

De plus, sa densité, si elle existe, est définie comme suit :

$$f = \frac{\partial^d F}{\partial x_1 ... \partial x_d} \tag{4.2}$$

En partant de la fonction de répartition jointe, il est possible de déterminer facilement les lois marginales des variables aléatoires ainsi que la structure de dépendance. *A contrario*, il n'est pas évident de connaître la fonction de répartition jointe dans les modèles pour lesquels seules les lois marginales sont explicitées. C'est dans ce cadre que la fonction copule représente un réel intérêt puisqu'elle permet de caractériser la fonction de répartition

jointe en partant des lois marginales, tout en ayant un regard sur la structure de dépendance.

4.3 La notion et les types de dépendance

Dans cette section, nous nous plaçons dans le cas où $d = 2$, c'est à dire deux risques ou deux variables aléatoires réelles afin de simplifier les formules.

Les sinistralités pour deux branches de risques sont considérées et représentées par les variables aléatoires réelles X et Y disposant chacune de n observations $(x_1, ..., x_n)$ et $(y_1, ..., y_n)$.

4.3.1 Définition de la dépendance entre deux variables aléatoires réelles

$$X \text{ et } Y \text{ sont indépendantes} \Leftrightarrow F_{X,Y}(x,y) = F_X(x)F_Y(y),\ \forall (x,y) \in \mathfrak{R}^2 \tag{4.3}$$

$$X \text{ et } Y \text{ sont dépendantes} \Leftrightarrow \exists (x,y) \in \mathfrak{R}^2 \text{ tel que } F_{X,Y}(x,y) \neq F_X(x)F_Y(y) \tag{4.4}$$

D'après cette définition, on peut dire que l'unique cas où les lois marginales seules permettent de caractériser la loi jointe du couple est le cas pour lequel il y a indépendance entre les deux variables aléatoires réelles. En d'autres termes, les lois marginales de X et de Y ne permettent pas de caractériser à elles seules la loi jointe du couple (X, Y) dès lors qu'il existe une quelconque dépendance entre les deux variables aléatoires réelles X et Y.

Effectivement, la loi jointe du couple de variables, en présence de dépendance, contient également la structure de dépendance entre X et Y, qui représente une information sur le type de dépendance qui existe entre les deux variables. Afin de quantifier cette structure de dépendance, il serait naturel de chercher à obtenir l'intensité et le sens de la dépendance à l'aide d'une mesure de dépendance. Cette dernière permettrait de caractériser le type de dépendance liant les branches de risque entre elles.

Il existe plusieurs types de dépendances, on distingue la notion de dépendance en quadrant positif, DPQ (dépendance faible), la notion d'anti-monotonie et celle de co-monotonie (dépendance forte).

4.3.2 La dépendance en quadrant positif

- **Notations et définitions**

Soient deux variables aléatoires réelles X et Y de fonction de répartition jointe $F_{X,Y}$ et F_X, F_Y les fonctions de répartition marginales de X et de Y respectivement.

En guise de rappel, il convient de définir la fonction de survie $S_{X,Y}$ du couple (X,Y) :

$$S_{X,Y}(x,y) = \mathbb{P}(X \geq x, Y \geq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \tag{4.5}$$

Les variables aléatoires réelles X et Y sont dépendantes en quadrant positif si :

$$\begin{aligned} &\mathbb{P}(X \geq x, Y \geq y) \geq \mathbb{P}(X \geq x)_\times \mathbb{P}(Y \geq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \\ &\Leftrightarrow S_{X,Y}(x,y) \geq S_X(x)_\times S_Y(y),\ \forall (x,y) \in \mathfrak{R}_+^2 \end{aligned} \tag{4.6}$$

- **Interprétation**

A partir de l'équation (4.6), on peut écrire que les variables aléatoires réelles X et Y sont dépendantes en quadrant positif si :

$$\begin{aligned} &\frac{\mathbb{P}(X \geq x, Y \geq y)}{\mathbb{P}(Y \geq y)} \geq \mathbb{P}(X \geq x),\ \forall (x,y) \in \mathfrak{R}_+^2 \\ \Leftrightarrow &\frac{\mathbb{P}(X \geq x, Y \geq y)}{\mathbb{P}(X \geq x)} \geq \mathbb{P}(Y \geq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \end{aligned} \tag{4.7}$$

D'après la formule de Bayes[7], il vient que :

$$\begin{aligned} &\mathbb{P}(X \geq x | Y \geq y) \geq \mathbb{P}(X \geq x),\ \forall (x,y) \in \mathfrak{R}_+^2 \\ \Leftrightarrow &\mathbb{P}(Y \geq y | X \geq x) \geq \mathbb{P}(Y \geq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \end{aligned} \tag{4.8}$$

Donc nous pouvons dire que, d'après l'équation (4.8), deux variables aléatoires réelles sont dépendantes en quadrant positif si la probabilité d'avoir une valeur élévée pour l'une d'entre elles est importante, sachant que l'autre variable a d'ores et déjà atteint une forte valeur.

4.3.3 La co-monotonie et l'anti-monotonie

- **Définitions**

On dit que deux variables aléatoires réelles X et Y sont co-monotones s'il existe deux fonctions non-décroissantes f et g et une variable aléatoire réelle Z telles que :

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(f(Z) \leq x, g(Z) \leq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \tag{4.9}$$

On dit que deux variables aléatoires réelles X et Y sont anti-monotones s'il existe une fonction f non-décroissante, une fonction g non-croissante et une variable aléatoire réelle Z telles que :

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(f(Z) \leq x, g(Z) \leq y),\ \forall (x,y) \in \mathfrak{R}_+^2 \tag{4.10}$$

- **Interprétation**

La co-monotonie et l'anti-monotonie s'interprètent donc de la même façon. Pour deux variables aléatoires réelles X et Y co-monotones ou anti-monotones, on peut dire que les deux branches de risques sous-jacentes à ces deux variables peuvent être expliquées par une troisième branche de risque sous-jacente à une troisième variable aléatoire réelle Z.

7. Soit (A_n) un système complet d'événements, tous de probabilité non nulle. Alors, quel que soit l'événement B, on a : $\mathbb{P}(B) = \sum_{n\geq 1} \mathbb{P}(B|A_n) \times \mathbb{P}(A_n)$.

4.4 Les mesures de dépendance

Appelons δ une mesure de dépendance entre deux risques, $\delta(X,Y)$ représente la valeur prise par δ sur le couple (X,Y).

Tout le problème réside dans le fait de trouver une mesure de dépendance de façon à ce que l'information délivrée par le triplet (F_X, F_Y, δ) et celle donnée par $F_{X,Y}$ soient équivalentes. En d'autres termes, nous souhaitons trouver une mesure de dépendance δ de sorte que les lois marginales et δ caractérisent entièrement la loi jointe de (X,Y).

δ doit vérifier les propriétés suivantes :

- $\delta(X,Y) = \delta(Y,X), \forall (X,Y)$: δ est **symétrique** ;
- $|\delta(X,Y)| \leq 1, \forall (X,Y)$: δ est **normée** ;
- $\delta(X,Y) = 1 \Leftrightarrow (X,Y)$ est **co-monotone** ;
- $\delta(X,Y) = -1 \Leftrightarrow (X,Y)$ est **anti-monotone** ;
- $\delta(g(X), g(Y)) = \delta(X,Y), \forall (X,Y)$, avec g une fonction réelle strictement croissante ;
- $\delta(X,Y) = 0 \Leftrightarrow$ les variables X et Y sont indépendantes.

Une mesure classique de la dépendance peut s'obtenir à l'aide du coefficient de corrélation linéaire de Pearson. Ce coefficient a longtemps été le seul outil permettant de mesurer la dépendance entre deux risques. En effet, la distribution Normale multi-variée a longtemps été utilisée pour modéliser la dépendance entre variables aléatoires. De plus, ce coefficient joue un rôle central en finance moderne (par exemple en gestion de portefeuille *via* la théorie de Markovitz, etc.).

Nous verrons que ce coefficient est fiable uniquement en présence d'une relation de dépendance linéaire et en considérant l'univers Gaussien, ce qui est regrettable puisque ce cadre d'étude n'est que très rarement retrouvé en pratique dans les milieux de l'assurance.

Afin de palier à cela, les praticiens utilisent d'autres mesures de dépendance. On peut citer le Rho de Spearman et le Tau de Kendall qui sont des coefficients de corrélation non linéaires. Ces derniers se basent sur les rangs et les concordances des observations d'un échantillon donné, à la différence du coefficient de corrélation linéaire de Pearson qui établit la corrélation entre les valeurs des observations elles-mêmes.

Ces deux dernières mesures de la dépendance ont également l'avantage de pouvoir s'exprimer de manière simple en fonction de la copule associée au couple de variables aléatoires réelles étudié. Cette idée est issue et approfondie dans les travaux de Cadoux et al. [2006] [20], Embrechts et al. [2002] [21], Demarta et McNeil [2004] [22] et Lindskog et al. [2003] [23].

4.4.1 Le coefficient de corrélation linéaire de Pearson

Le coefficient de corrélation linéaire de Pearson entre deux variables aléatoires réelles X et Y est défini comme suit :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{4.11}$$

$$\text{avec : } \begin{cases} Cov(X,Y) = \mathbb{E}\left\{[X - \mathbb{E}[X]]\,[Y - \mathbb{E}[Y]]\right\} = \sum_{i=1}^{n}(X_i - \bar{X})_{\times}(Y_i - \bar{Y}),\ \forall (X,Y) \\ \bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i,\ \forall X \\ Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = Cov(X,X),\ \forall X \end{cases} \tag{4.12}$$

On déduit de cette équation les deux points suivant :

- $\rho(X,Y) = 0 \Leftrightarrow$ les variables aléatoires réelles X et Y sont non corrélées linéairement ;
- $\rho(X,Y) \neq 0 \Leftrightarrow$ les variables aléatoires réelles X et Y sont corrélées linéairement.

Le coefficient de corrélation linéaire de Pearson permet de capter uniquement les dépendances linéaires. On parle de dépendance linéaire lorsque $\rho = 1$ (respectivement -1) $\Leftrightarrow Y = aX + b$, avec $a > 0$ (respectivement $a < 0$) et b $\in \Re$.

Enfin, il faut faire attention au fait que la notion de non-corrélation et d'indépendance ne sont pas équivalentes. On peut seulement affirmer que l'indépendance implique la non corrélation : X et Y indépendantes $\Rightarrow$ X et Y non corrélées.

Comme énoncé *supra*, le coefficient de Pearson peut être utilisé comme structure de dépendance dès lors que la loi jointe est supposée Normale multi-variée, ce qui n'est pas toujours le cas en pratique. Si on se limite au cas bi-varié, (X,Y) est distribué selon une loi $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ et sa fonction de répartition jointe s'écrit comme suit :

$$F_{X,Y}(x,y) = \iint_{\Re^2} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} dxdy \tag{4.13}$$

D'après cette équation, seules les lois marginales et ρ rentrent en jeu dans la caractérisation de la loi jointe. De plus, ρ représente bien la structure de dépendance de (X,Y) [8].

Nous pouvons conclure que le coefficient de corrélation linéaire de Pearson n'est pertinent qu'en présence de distribution elliptiques [9] ou de dépendance linéaire. Il est donc à utiliser en vérifiant préalablement ces points. En cas de non-respect de ces hypothèses, d'autres mesures de dépendances peuvent être utilisées. Ces dernières vont être définies *infra*, elles se basent sur la notion de concordance et de rang des observations [10].

4.4.2 Le Rho de Spearman

On peut définir le Rho de Spearman comme étant la corrélation entre les fonctions de répartition marginales des deux variables aléatoires qui forment le couple (X_1, X_2). Il est possible de définir le Rho de Spearman en fonction de ρ, le coefficient de corrélation linéaire de Pearson et nous obtenons :

8. Lorsque $d \geq 2$, la structure de dépendance sera traduite par une matrice de corrélation.
9. Notion définie à la section 4.7.1, page 56.
10. Ces notions sont définies à la section 4.6.1, page 53, et dans le mémoire de Valérie Lagenebre [13] (à la section 1.4.2).

$$\rho_s(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) \tag{4.14}$$

avec F_1 et F_2 les fonctions de répartition marginales des variables aléatoires du couple (X_1, X_2). Soit C la copule relative au couple (X_1, X_2), on obtient l'expression du Rho de Spearman suivante :

$$\rho_s(X_1, X_2) = 12 \iint_{[0,1]^2} u_1 u_2 dC(u_1, u_2) - 3 = 12 \iint_{[0,1]^2} C(u_1, u_2) du_1 du_2 - 3 \tag{4.15}$$

Il est possible de construire un estimateur empirique du Rho de Spearman à partir d'un échantillon de taille T de (X_1, X_2), $(x_1^t, x_2^t)_{1 \leq t \leq T}$. Ce dernier est défini comme suit :

$$\hat{\rho}_s(X_1, X_2) = \frac{\sum_{i=1}^T (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^T (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^T (S_i - \bar{S})^2}} \tag{4.16}$$

avec R_i le rang de x_1^i, S_i celui de x_2^i, $\forall i \in [1, T]$ et $\bar{Z} = \frac{1}{T} \sum_{i=1}^T Z_i$.

4.4.3 Le Tau de Kendall

Soit (X_1, X_2) un couple de vecteurs aléatoires et (X_1', X_2') un couple de vecteurs aléatoires identique à (X_1, X_2). En d'autres termes, (X_1', X_2') est une copie en tout point identique à (X_1, X_2). On peut définir le Tau de Kendall comme suit :

$$\tau(X_1, X_2) = \mathbb{P}\left\{(X_1 - X_1')(X_2 - X_2') > 0\right\} - \mathbb{P}\left\{(X_1 - X_1')(X_2 - X_2') < 0\right\} \tag{4.17}$$

Cette mesure de dépendance correspond à la différence entre la probabilité de concordance et celle de discordance. Soit C la copule relative au couple (X_1, X_2), le Tau de Kendall peut alors s'écrire de cette manière :

$$\begin{aligned} \tau(X_1, X_2) &= 4 \iint_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1 = 1 - 4 \iint_{[0,1]^2} \partial_{u_1} C(u_1, u_2) \partial_{u_2} C(u_1, u_2) du_1 du_2 \\ &= 4\mathbb{E}\left(C(U_1, U_2)\right) - 1 \quad \text{avec} \quad U_1, U_2 \sim U[0, 1] \end{aligned} \tag{4.18}$$

Il est possible de construire un estimateur empirique du Tau de Kendall à partir d'un échantillon de taille T de (X_1, X_2), $(x_1^t, x_2^t)_{1 \leq t \leq T}$. Ce dernier est défini comme suit :

$$\hat{\tau}(X_1, X_2) = \frac{2}{T(T-1)} \sum_{j=2}^{T} \sum_{i=1}^{j-1} sign\left\{(x_1^j - x_1^i)(x_2^j - x_2^i)\right\} \quad \text{avec } sign(z) = \begin{cases} 1 \text{ si } z \geq 0 \\ -1 \text{ si } z < 0 \end{cases} \tag{4.19}$$

Les trois mesures de dépendance que nous venons d'évoquer sont des mesures de dépendance globales. En fonction du type de données, étudier la dépendance sur les queues de distributions peut être bénéfique. Il peut être également intéressant de voir si pour un sinistre donné, un montant de remboursement élevé pour une garantie donnée implique un montant de remboursement également élevé pour une seconde garantie donnée.

Certaines copules, comme celle de Student, permettent de caractériser la dépendance de queues. Il existe également des copules qui ne permettent pas de caractériser la dépendance au niveau des queues de distribution. C'est le cas, par exemple, de la copule de Frank et de la copule Gaussienne.

4.5 Les méthodes de détection de la dépendance

4.5.1 Les méthodes de détection graphiques

L'avantage de ces méthodes graphiques est de pouvoir évaluer visuellement si nous sommes en présence de dépendance et, si possible, de déterminer le type de dépendance. L'inconvénient majeur de ces méthodes est qu'elles ne permettent pas de tester formellement l'hypothèse d'indépendance comme c'est le cas avec un test statistique.

Pour mémoire, nous sommes en présence de deux variables aléatoires réelles X et Y qui représentent chacune les montants des sinistres relatifs à une branche de risque donnée.

- **Représentation graphique des observations : le diagramme de dispersion**

Il s'agit tout simplement de représenter graphiquement le nuage des points $(x_i, y_i), i = 1, ..., n$. Il est légitime de considérer X et Y indépendantes dans le cas où les points sont uniformément répartis dans le plan.

Au contraire, si les points sont alignés sur une diagonale ascendante (respectivement descendante) alors on a de forte chance d'être en présence d'une corrélation linéaire positive (respectivement négative).

L'avantage d'un tel graphique est donc de pouvoir détecter une corrélation linéaire de manière simple et rapide.

- **Représentation graphique des rangs normalisés des observations : le dépendogramme**

Le dépendogramme est en fait un nuage de points relatif aux marginales uniformes issues de l'échantillon. Il permet de mettre en évidence la structure de dépendance d'un couple (X, Y) de variables aléatoires réelles. Il est aisé d'obtenir les marginales uniformes en transformant les observations de notre échantillon en utilisant les fonctions de répartition empiriques marginales des variables aléatoire réelles X et Y.

La fonction de répartition empirique F_n de X est déterminée à partir des n observations $(x_1, ..., x_n)$ constituant notre échantillon et est définie comme suit :

$$F_n(x_i) = \frac{Rang(x_i)}{n} \in [0, 1], \forall i = 1, ..., n \tag{4.20}$$

Soient F_n et G_n, les fonctions de répartition empiriques de X et Y respectivement. Le dépendogramme de ces deux variables aléatoires réelles est alors le nuage des points $(F_n(x_i), G_n(y_i)), \forall i = 1, ..., n$.

Ce type de représentation graphique permet de détecter une éventuelle concentration des réalisations issues de l'échantillon étudié. Une concentration qui peut notamment être présente dans les queues de distribution, ce qui peut conduire à tester l'adéquation de nos données avec une copule prenant en compte la dépendance de queues comme la copule de Student par exemple.

- **Le Khi-Plot**

Fisher et Switzer [1985, 2001] sont à l'origine de cette méthode graphique qui permet de détecter la présence d'associations entre deux variables aléatoires continues.

Cette méthode s'appuie de manière exclusive sur les couples de rangs des observations et est ainsi non paramétrique. La justification de cette méthode repose sur le fait que la structure de dépendance entre deux variables continues est caractérisée par la copule sous-jacente au modèle et que les paires de rangs sont des statistiques maximalement invariantes par rapport à toute transformation monotone croissante des marginales.

La construction de ce graphique s'effectue en plusieurs étapes. Dans un premier temps, il s'agit de calculer les trois statistiques suivantes, pour chaque couple (X_i, Y_i), avec $1 \leq i \leq n$:

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

et

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

Ensuite, il s'agit de représenter les paires (λ_i, χ_i) sur un graphique, avec :

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

et

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

$\lambda_i \in [-1, 1]$ mesure la distance entre les couples (X_i, Y_i) et le centre du nuage de points. Afin de supprimer les valeurs aberrantes, sont enlevées les paires telles que :

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

- **Le Kendall-Plot**

Le Kendall-Plot ou K-Plot a été proposé par Genest et Boies [2003] [24] et permet également de visualiser la dépendance.

La courbure, que ce graphique produit en présence de dépendance, est associée de près à la copule sous-jacente à la loi des observations. Il est plus facile d'interpréter le K-Plot que le Khi-Plot.

Le K-Plot est, comme le Khi-Plot, fonction des rangs des observations. Toutefois, le Khi-Plot s'inspire de la statistique du Khi-deux d'indépendance et est apparenté à la notion de "carte de contrôle", tandis que le K-Plot est issu de la transformation intégrale de probabilité multi-variée et découle, sur le plan graphique, du principe de la droite de Henry ou "Q-Q-Plot".

De ce fait, l'absence d'indépendance se manifeste dans le K-Plot par la présence d'une courbure caractéristique de la copule sous-jacente au modèle.

La méthode de construction du K-Plot est, comme pour celle du Khi-Plot, définie en plusieurs étapes :

1. On calcule $H_i = \frac{1}{n-1}\#\{j \neq i \ : X_j \leq X_i, Y_j \leq Y_i\}$,
2. On ordonne les $H_1, ..., H_n$ pour obtenir $H_{(1)} \leq ... \leq H_{(n)}$,
3. On trace les paires $(W_{i:n}, H_{(i)})$, $1 \leq i \leq n$, où $W_{i:n}$ est l'espérance de la i^{eme} statistique d'ordre d'un échantillon de la taille n de K_0. $K_0(w) = w - w\log(w)$, pour $w \in [0,1]$. Genest et Boies [2003] [24] montrent que K_0 est la loi asymptotique de H_i sous l'hypothèse d'indépendance.

Seules les principales propriétés et interprétations des Kendall-Plot seront détaillées ci-dessous. Néanmoins, le lecteur intéressé pourra se référer à Genest et Boies [2003] [24] pour davantage de détails. Les deux principales propriétés sont :

- il est possible d'exprimer le Tau de Kendall à partir de la fonction K :

$$\tau(X,Y) = 3 - 4\int_0^1 K(w)dw, \tag{4.21}$$

- pour n suffisamment grand, les points de coordonnées $(W_{i:n}, H_{(i)})$ se retrouvent concentrés sur la courbe $p \to (K_0^{-1}(p), K^{-1}(p))$. Le K-Plot ressemblera alors au graphique associé à la courbe $w \to K^{-1}\{K_0(w)\}$.

Ces propriétés permettent d'interpréter le K-Plot :

- de la linéarité sera observée lorsque $K = K_0$, c'est à dire, sous l'hypothèse d'indépendance entre X et Y. Ainsi, plus les points se rapprochent de la bissectrice, plus la dépendance est faible,

- $\tau(X,Y) = 1$ si X et Y sont co-monotones, soit $K(p) = p$ et donc $K^{-1}(p) = p$, $\forall p \in [0,1]$. Le K-Plot est alors aligné sur la courbe d'équation $w \to K_0(w) = w - w\log(w)$,

- $\tau(X,Y) = -1$ si X et Y sont anti-monotones, soit $K^{-1}(p) = 0, \forall p \in [0,1]$. Le K-Plot est alors aligné sur l'axe horizontal.

Les observations effectuées ci-dessus permettent de résumer les allures des Kendall-Plot obtenues selon la dépendance en présence :

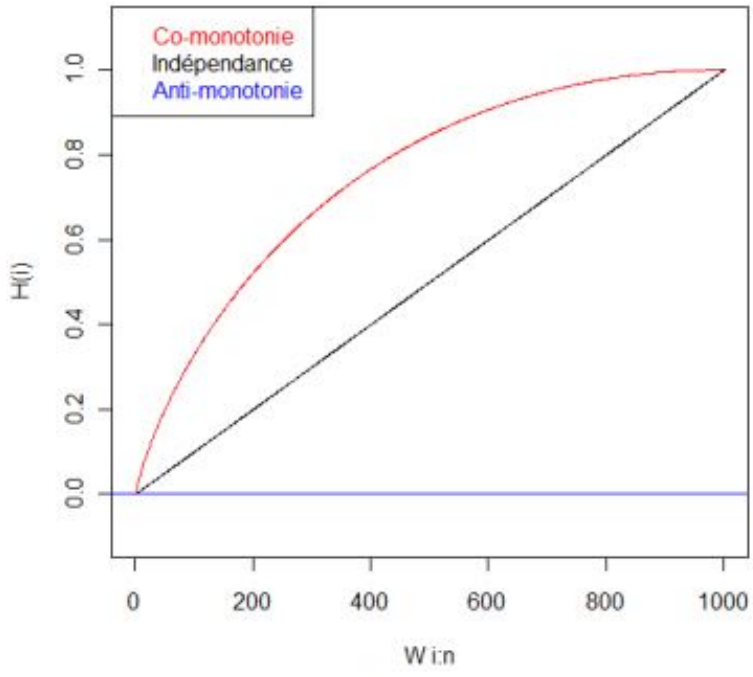


FIGURE 4.1 – Les interprétations des Kendall-Plot

4.5.2 Les tests statistiques d'indépendance

Pour les trois tests statistiques suivants, on définit X et Y comme étant deux variables aléatoires réelles.

- **Le test de corrélation de Pearson**

Soit $\rho = \rho(X, Y)$ le coefficient de corrélation linéaire de Pearson entre X et Y.

Les hypothèses suivantes sont testées :

$$\begin{cases} H_0 & : \rho = 0 \\ H_1 & : \rho > 0 \end{cases} \tag{4.22}$$

Soit T, la statistique de test définie comme suit :

$$T = \sqrt{n-2}\frac{\rho}{\sqrt{1-\rho^2}} \approx \text{Student}(n-2) \tag{4.23}$$

avec n, la taille de l'échantillon et $(n-2)$, le nombre de degrés de liberté (ddl).

En ce qui concerne la région critique, il s'agit de calculer $T_n(\omega)$ sur la base des observations et on rejette H_0 avec un risque α si $T_n(\omega) > t_\alpha$ où t_α est le quantile d'ordre α de la loi de Student à $(n-2)$ ddl.

- **Le test de corrélation de Spearman**

Soit $\rho_s = \rho_s(X, Y)$ le Rho de Spearman.

Les hypothèses suivantes sont testées :

$$\begin{cases} H_0 & : \rho_s = 0 \\ H_1 & : \rho_s \neq 0 \end{cases} \tag{4.24}$$

La distribution de ρ_s, sous H_0, est proche de $N\left(0, \frac{1}{n-1}\right)$.

En ce qui concerne la région critique, en considérant un test bilatéral au seuil de risque de 5 %, on ne rejettera pas H_0 si $\sqrt{n-1}\rho_s \in [-1,96, +1,96]$.

- **Le test de corrélation de Kendall**

Soit $\tau = \tau(X, Y)$ le Tau de Kendall.

Les hypothèses suivantes sont testées :

$$\begin{cases} H_0 & : \tau = 0 \\ H_1 & : \tau \neq 0 \end{cases} \tag{4.25}$$

La distribution de τ, sous H_0, est proche de $N\left(0, \frac{2(2n+5)}{9n(n-1)}\right)$.

En ce qui concerne la région critique, en considérant un test bilatéral au seuil de risque de 5 %, on ne rejettera pas H_0 si $\sqrt{\frac{9n(n-1)}{2(2n+5)}}\tau \in [-1,96, +1,96]$.

4.6 Définitions et propriétés relatives aux copules

L'objectif de cette section n'est pas de présenter de manière exhaustive l'ensemble des définitions et propriétés relatives aux copules mais plutôt de restituer les plus importantes, celles nécessaires à la compréhension de ce mémoire. Néanmoins, s'il souhaite obtenir des compléments, le lecteur intéressé pourra se référer à l'ouvrage de Nelsen [1999] [25] et à celui de Joe [1997] [26].

4.6.1 Définition d'une fonction copule

On appelle copule de dimension d, une fonction $C : [0,1]^d \to [0,1]$, ou tout simplement copule, si C est (la restriction à $[0,1]^d$ de) la fonction de répartition d'une variable aléatoire $U = (U_1, \ldots, U_d)$ à valeurs dans $\mathfrak{R}^d$, avec $U_1, ..., U_d$, des variables aléatoires de loi uniformes[11] sur $[0,1]$: $C(u) = \mathbb{P}(U \leq u)\ \forall u \in [0,1]^d$.

On définit $X = (X_1, ..., X_d)$ une variable aléatoire à valeur dans $\mathfrak{R}^d$ de fonction de répartition F et F_k la fonction de répartition de X_k pour $1 \leq k \leq d$. Une copule pour la variable aléatoire X est une copule qui vérifie la relation suivante : $F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))\ \forall x$ tel que $x = (x_1, ..., x_d) \in \mathfrak{R}^d$.

En guise d'illustration, la copule $C(u_1, ..., u_d) = \prod_{k=1}^{d} u_k$ représente la fonction de répartition des variables aléatoires indépendantes $U_1, ..., U_d$ de lois uniformes sur [0,1].

11. Soit U une variable aléatoire de loi uniforme sur $[0,1]$, alors U a pour fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases} \tag{4.26}$$

Si on se limite à l'étude des copules bi-variées, on peut définir la copule bi-variée C, fonction de $[0,1]^2 \rightarrow [0,1]$, par les points suivants :

- $C(u,0) = C(0,v) = 0, \forall (u,v) \in [0,1]^2$;
- $C(u,1) = C(1,u) = u, \forall u \in [0,1]$: les marges des distributions marginales sont des marges uniformes ;
- C est 2-increasing, alors on a :
  $C(v_1,v_2) - C(u_1,v_2) - C(v_1,u_2) + C(u_1,u_2) \geq 0, \forall (u_1,u_2) \in [0,1]^2$ et $\forall (v_1,v_2) \in [0,1]^2$ avec $0 \leq u_1 \leq v_1 \leq 1$ et $0 \leq u_2 \leq v_2 \leq 1$.

La définition suivante est vérifiée : $C(u_1,u_2) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2), \forall (u_1,u_2) \in [0,1]^2$, dès lors que U_1 et U_2 sont deux variables aléatoires de lois uniformes sur $[0,1]$. Selon cette définition des copules à partir des probabilités, le second point de la définition d'une copule bi-variée est vérifié : la copule est caractérisée comme étant une distribution de probabilité avec des marges uniformes.

En pratique, nous allons travailler directement avec les uniformes empiriques et non plus avec les réalisations des variables aléatoires réelles. En d'autres termes, il s'agit de transformer les réalisations $(x_1, ..., x_n)$ de la variable aléatoire X en uniformes empiriques $(u_1, ..., u_n)$ avec $u_i = \dfrac{Rang(x_i)}{n+1}, \forall i \in \{1, ..., n\}$ et n étant la taille de l'échantillon.

Il convient de définir la fonction $Rang()$, qui retourne le rang (la position) de l'observation x_i dans l'échantillon de taille n. La fonction $rank()$ permet d'effectuer ce travail sous le logiciel R [12].

Par exemple, $rank(c(1,1,2,1,3))$ renvoit les rangs suivants : 2 ;2 ;4 ;2 ;5 alors que $rank(c(1,1,2,1,1,3))$ renvoit 2,5 ;2,5 ;5,0 ;2,5 ;2,5 ;6,0. Cet exemple illustre parfaitement ce que l'on entend par rangs des observations et permet de s'apercevoir qu'en cas d'ex aequo, une moyenne est effectuée entre la position du premier ex aequo et celle du dernier. C'est cette moyenne qui va caractériser le rang de tous les ex aequo. Autrement dit, en cas d'ex aequo, ces derniers auront le même rang.

La copule est donc la fonction qui permet de lier les rangs des observations constituant l'échantillon de taille n.

4.6.2 Densité d'une fonction copule

Une copule est, par définition, une fonction de répartition d'un vecteur de variables aléatoires uniformes sur $[0,1]$. Ainsi sa densité, lorsqu'elle existe, est la différentiation de cette copule. L'intérêt de travailler avec la densité de la copule est donc de permettre de séparer la structure de dépendance des densités marginales. C'est ce qui va être explicité *infra*.

Pour rappel, la densité f d'une fonction de distribution F est définie, si elle existe, par :

$$f(x_1, ..., x_d) = \frac{\partial F(x_1, ..., x_d)}{\partial x_1 ... \partial x_d} \tag{4.27}$$

En supposant que la copule C ainsi que les distributions marginales $F_1, ..., F_d$ sont différentiables, on peut dans ce cas écrire la densité jointe de la variable aléatoire $X = (X_1, ..., X_d)$ comme suit :

12. Les options de cette fonction sont détaillées à partir de ce lien bibliographique [27].

$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \prod_{k=1}^{d} f_k(x_k) \tag{4.28}$$

avec f_k, pour $1 \leq k \leq d$, la densité de probabilité définie *supra*, f la densité jointe issue de F et c la densité de la copule C définie comme suit :

$$c = \frac{\partial^d C}{\partial u_1 ... u_d} \tag{4.29}$$

A partir de cette définition, on constate que l'on peut éclater la densité jointe pour former deux parties distinctes. La première partie : $c(F_1(x_1), ..., F_d(x_d))$, exprime l'information relative à la structure de dépendance des variables aléatoires $X_1, ..., X_d$ et la seconde est tout simplement le produit des densités marginales. Ceci illustre parfaitement le fait que les copules permettent d'isoler la structure de dépendance de la distribution jointe et ainsi de l'étudier indépendamment des comportements marginaux.

La densité jointe et les densités marginales permettent donc de caractériser entièrement la densité d'une copule. Soit F une fonction de répartition qui admet la densité continue f. De ce fait, les lois marginales $F_1, ..., F_d$ sont continues et admettent des densités notées respectivement $f_1, ..., f_d$. Ainsi, la copule de F admet la densité c définie comme suit :

$$c(u_1, ..., u_d) = \frac{f(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) ... f_d(F_d^{-1}(u_d))} \tag{4.30}$$

C'est selon cette définition que l'estimation de la densité d'une copule donnée est réalisable à partir de l'estimation des lois marginales et de la loi jointe.

Enfin, dans le cas simple d'une copule C bi-variée, la densité c est définie par :

$$c(u, v) = \frac{\partial^2 C}{\partial u \partial v}(u, v), \forall (u, v) \in [0, 1]^2 \tag{4.31}$$

4.6.3 Enoncé du théorème de Sklar

La notion de copule a été introduite par Sklar en 1959. C'est à partir de ses travaux qu'il a énoncé le théorème central qui a permis par la suite le développement de la théorie des copules. L'intérêt de ce théorème est de pouvoir préciser le lien entre les fonctions de répartition marginales uni-variées $F_1, ..., F_d$ et la distribution complète multi-variée F, sachant que ce lien est défini par la copule C.

Pour la démonstration du théorème suivant, le lecteur intéressé pourra se référer à Nelsen [1999] [25] ou Sklar [1959] [28].

Théorème de Sklar :

- Soit $F_1, ..., F_d$ des fonctions de répartition de variables aléatoires à valeurs dans $\Re$ et C une copule. La fonction F définie par $F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$ est la fonction de répartition d'une variable aléatoire à valeurs dans $\Re^d$.
- Soit $X = (X_1, ..., X_d)$ une variable aléatoire à valeurs dans $\Re^d$. Il existe donc une copule pour X. Si en plus, les fonctions de répartition des variables aléatoires $X_1, ..., X_d$ sont continues, alors la copule C est unique.

Ce théorème traduit le fait qu'une copule permet d'exprimer une fonction de répartition multi-variée en fonction des fonctions de distribution marginales. De plus, cette copule

contient toute l'information sur la structure de dépendance.

4.6.4 Enoncé du théorème d'invariance

L'invariance par transformations strictement croissantes est également un théorème important pour la bonne compréhension de la théorie des copules.

Théorème d'invariance :
Soient X_1 et X_2, deux variables aléatoires continues de marges F_1 et F_2 et de copule $C(X_1, X_2)$ associée à la distribution F du vecteur aléatoire (X_1, X_2). Soient h et g deux fonctions strictement croissantes sur $Im(X_1)$ et $Im(X_2)$ [13] respectivement. On peut donc établir l'égalité suivante :

$$C(h(X_1), g(X_2)) = C(X_1, X_2) \tag{4.32}$$

D'après ce théorème, on peut dire que la copule est invariante par transformations strictement croissantes des variables aléatoires, c'est à dire, en présence de variables aléatoires co-monotoniques (Cf. section 4.3.3, page 45).

Ce théorème traduit simplement le fait que le dépendogramme [14] entre deux variables aléatoires réelles X_1 et X_2 est identique à celui de $g(X_1)$ et $h(X_2)$, g et h étant deux fonctions strictement croissantes. Ainsi, le dépendogramme reste le même si une copule est invariante par transformations croissantes des marginales.

4.7 Les deux grandes familles de copules paramétriques

L'objectif d'une étude portant sur l'analyse de la dépendance entre les branches de risque est de déterminer la copule paramétrique la plus adaptée à la structure de dépendance d'un triplet de branches de risque à savoir la Responsabilité Civile corporelle, la Responsabilité Civile matérielle et le Dommage.

Dans l'application réalisée au chapitre 7, page 94, et afin de rechercher la copule qui caractérise le mieux la structure de dépendance, il a été préalablement décidé de sélectionner six copules : la copule Gaussienne, de Student, de Clayton, de Gumbel, de Frank et enfin la copule de Joe. Ces copules vont être détaillées *infra* et ont été retenues selon des critères de flexibilité, de simplicité analytique et également du fait de la diversité des formes de dépendance qu'elles comportent.

Enfin, ces copules peuvent être classées selon deux grandes familles de copules : les copules elliptiques et les copules archimédiennes.

4.7.1 Les copules elliptiques

$M_d(\Re)$ est défini comme étant l'ensemble des matrices carrées réelles de taille d^2. Une copule est dite elliptique si elle est la copule d'une loi elliptique. Une loi continue est dite elliptique de paramètre de position $\mu = (\mu_1, ..., \mu_d) \in \Re^2$ et de matrice de forme symétrique définie positive $\Sigma \in M_d(\Re)$ si sa densité f peut s'écrire comme suit :

13. $Im(X_1) = \{x_2 / \exists \omega \in \Omega \ : x_2 = X_1(\omega)\}$.
14. Un dépendogramme donne la structure de dépendance que la copule induit.

$$f(x) = (\det\Sigma)^{-\frac{1}{2}} g\left((x-\mu)\Sigma^{-1}(x-\mu)'\right) \forall x = (x_1, ..., x_d) \in \mathfrak{R}^d \tag{4.33}$$

avec z' la transposée de z et g une fonction à valeurs positives qui vérifie $\int_{\mathfrak{R}^d} g(xx')dx = 1$.

Soit $\epsilon(\mu, \Sigma, g)$ cette famille de lois dites elliptiques (car les courbes de niveaux de la densité sont des ellipses). La loi est dite sphérique si $\Sigma = kI_d$ où $k > 0$ et I_d est la matrice unité de $M_d(\mathfrak{R})$

Les lois elliptiques relatives à la même fonction g font partie de la même famille elliptique, dans laquelle on distingue le représentant standard (centré réduit) pour lequel $\mu = 0$ et $\Sigma = I_d$. On parlera, pour la suite, de Σ comme étant une matrice de corrélation.

Comme exemple classique de lois elliptiques, on peut citer la loi d'un vecteur Gaussien, associé au choix de $g(y) = (2\pi)^{-d/2}\exp(-y/2)$. Comme les vecteurs Gaussiens, ces lois vérifient des propriétés algébriques utiles. La propriété la plus intéressante réside dans le fait que les lois elliptiques forment une classe stable par transformation affine. Le lecteur intéressé par des détails sur ces lois pourra se référer à Fang et al. [1990] [29].

Les lois Normale et de Student sont des lois elliptiques. Deux exemples classiques de copules elliptiques, la copule Gaussienne et la copule de Student, sont présentés en annexe .6, page 144.

4.7.2 Les copules archimédiennes

La famille des copules archimédiennes a été définie par Genest et Mackay [1986] [30]. Contrairement aux copules relatives à la famille des copules elliptiques, les copules archimédiennes peuvent caractériser des structures de dépendance assez variées. Par exemple, on peut citer les dépendances asymétriques, où les coefficients relatifs à la dépendance de queue (inférieure et supérieure) peuvent prendre des valeurs différentes, ce qui permet un meilleur ajustement et une meilleure approche de la modélisation de la structure de dépendance.

Soit $\phi : [0,1] \to \mathfrak{R}_+$, une fonction continue, convexe et strictement décroissante telle que : $\phi(1) = 0$ et $\phi(0) = +\infty$. La fonction $C(u_1, ..., u_d) = \phi^{-1}\left(\phi(u_1) + ... + \phi(u_d)\right), \forall (u_1, ..., u_d) \in [0,1]^d$ est une copule si ϕ est d fois dérivable sur $]0,1[$ et si $\phi^i > 0$ pour i paire et $\phi^i < 0$ sinon, $\forall i$ avec $1 \leq i \leq d$. Les fonctions de ce type définissent les copules archimédiennes de générateur ϕ.

A partir de la définition ci-dessus, l'écriture de la densité d'une copule archimédienne d-variée se déduit aisément :

$$c(u_1, ..., u_d) = (\phi^{-1})^d \left(\phi(u_1) + ... + \phi(u_d)\right) \prod_{i=1}^{d} \phi'(u_i) \tag{4.34}$$

Les copules archimédiennes présentent deux grands intérêts :
- elles permettent de construire une grande variété de familles de copules. Il est ainsi possible de caractériser de nombreuses structures de dépendance.

- les copules générées présentent l'avantage d'avoir des formes analytiques fermées et il est assez aisé de les simuler.

Afin d'obtenir davantage d'informations sur les copules archimédiennes, le lecteur pourra se reporter à Nelsen [1999] [25].

Les copules testées dans ce mémoire lors de l'application à nos données (Cf. chapitre 7, page 94) sont les copules de Clayton, Gumbel, Frank et Joe. Elles sont définies en annexe .7, page 146, à l'exception des copules retenues pour modéliser au mieux la dépendance entre nos branches de risque, à savoir, les copules de Frank et de Gumbel, qui sont définies dans les deux paragraphes suivants.

- **La copule de Gumbel**

La copule de Gumbel, appelée aussi copule de Gumbel Hougaard, permet de modéliser les dépendances extrêmes. Elle permet ainsi de pallier aux insuffisances de la copule Gaussienne.

La copule de Gumbel permet de prendre en compte les dépendances positives et a l'avantage de représenter des risques dont la structure de dépendance est plus marquée sur la queue supérieure. Pour cette raison, cette copule est utilisée en assurance pour étudier l'impact de la survenance d'événements de forte intensité (en termes de coût de sinistre) sur la dépendance entre branches de risque.

Le générateur de cette copule est défini de la façon suivante :

$$\phi(u) = (-\ln(u))^{\alpha} \text{ avec } \alpha > 1 \text{ et } u \in]0,1]. \tag{4.35}$$

On déduit ainsi l'écriture de la copule de Gumbel :

$$C(u_1, ..., u_d) = \exp\left(-\left[\sum_{i=1}^{d}(-\ln(u_i))^{\alpha}\right]^{1/\alpha}\right) \tag{4.36}$$

- **La copule de Frank**

La copule de Frank permet la modélisation des dépendances aussi bien négatives que positives. Cependant, cette dernière ne présente pas de dépendance sur les queues de distribution. Le générateur associé à la copule de Frank est défini comme suit :

$$\phi(u) = -\ln\left(\frac{e^{-\alpha u} - 1}{e^{-\alpha} - 1}\right) \text{ avec } \alpha \neq 0 \text{ et } u \in [0,1]. \tag{4.37}$$

De plus, la copule de Frank d-variée peut s'écrire de cette façon :

$$C(u_1, ..., u_d) = -\frac{1}{\alpha}\ln\left(1 + \frac{1}{(e^{-\alpha} - 1)^{d-1}}\prod_{i=1}^{d}(e^{-\alpha u_i} - 1)\right) \tag{4.38}$$

La densité de cette copule pour le cas bi-varié s'écrit comme suit :

$$c(u_1, u_2) = \frac{\alpha(1 - e^{-\alpha})e^{-\alpha(u_1+u_2)}}{((1 - e^{-\alpha}) - (1 - e^{-\alpha u_1})(1 - e^{-\alpha u_2}))^2} \tag{4.39}$$

4.8 Les copules empiriques

Cette section a pour objectif de présenter deux méthodes de construction de la copule empirique d'un ensemble de variables aléatoires à partir de leurs observations respectives. Tandis que la première méthode, la copule de Deheuvels, se base sur la statistique de rang,

la seconde repose sur l'estimation empirique des densités de probabilité marginales et de la densité jointe.

4.8.1 La copule de Deheuvels

Deheuvels [1979] [31] est à l'origine de la notion de copule empirique. Cette dernière repose sur le rang des observations dans le but d'extraire la structure de dépendance.

Soit $(X_1, ..., X_d)$ une suite de variables aléatoires et $(x_1^n, ..., x_d^n)_{1 \leq n \leq N}$ un échantillon d'observations de taille N. De plus, $(r_1^n, ..., r_d^n)_{1 \leq n \leq N}$ est la statistique de rang associée à cet échantillon multi-varié.

Pour rappel, la statistique de rang est tout simplement le rang d'une observation, ou encore sa position dans l'échantillon préalablement ordonné. En conséquence, on a que pour tout $1 \leq i \leq d$, r_i^n est le rang de x_i^n dans $(x_i^n)_{1 \leq n \leq N}$ et donc la copule empirique introduite par Deheuvels est définie sur l'ensemble $\left\{ \left(\frac{n_1}{N}, ..., \frac{n_d}{N} \right), 1 \leq i \leq d, n_i = 0, ..., N \right\}$ par l'équation suivante :

$$\hat{C}_N \left(\frac{n_1}{N}, ..., \frac{n_d}{N} \right) = \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{d} 1_{\{r_i^n \leq n_i\}} \tag{4.40}$$

Cette copule empirique satisfait de nombreuses propriétés dont la convergence asymptotique vers C. En revanche, elle présente certains inconvénients comme le fait de représenter plusieurs points de discontinuité, ce qui limite son utilisation, notamment dans les situations où sa densité est requise.

La relation suivante permet tout de même d'estimer la fonction de densité de la copule empirique de Deheuvels à partir de la copule $\hat{C}_N$:

$$\hat{C}_N \left(\frac{n_1}{N}, ..., \frac{n_d}{N} \right) = \sum_{j_1=1}^{2} ... \sum_{j_d=1}^{2} (-1)^{j_1 + ... + j_d} \hat{C}_N \left(\frac{t_1 - j_1 + 1}{N}, ..., \frac{t_d - j_d + 1}{N} \right) \tag{4.41}$$

4.8.2 Estimation empirique par les densités de probabilité

Le principe de cette seconde méthode est d'estimer empiriquement les densités marginales $f_1, ..., f_d$ et la densité jointe f d'un vecteur de variables aléatoires $X = (X_1, ..., X_d)$. Ce dernier travail permet ensuite d'estimer la densité de sa copule à partir de l'équation suivante :

$$c(u_1, ..., u_d) = \frac{f\left(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)\right)}{f_1(F_1^{-1}(u_1)) ... f_d(F_d^{-1}(u_d))} \tag{4.42}$$

Cette dernière formule suppose que tous les termes au dénominateur sont différents de zéro.

4.9 Estimation d'une copule

Afin de faciliter les notations et sans perte de généralité, cette section traitera seulement le cas bivarié (d=2). Soit (X, Y) un vecteur aléatoire et on suppose que les variables aléatoires X et Y sont réelles avec $F(x)$ et $F(y)$ leurs fonctions de répartition respectives.

Pour mémoire, la fonction de distribution jointe du couple (X, Y) peut s'exprimer comme suit :

$$H(x, y) = C(F(x), G(x)), \tag{4.43}$$

avec C une fonction de distribution uniforme dont les marginales sont des uniformes sur $[0, 1]$. Lorsque F et G sont continues, C est unique et coïncide avec la fonction de distribution du couple $(U, V) = (F(X), G(Y))$. En pratique, H est inconnu. Il est possible de déterminer une copule pour le couple (X, Y) en supposant que F, G et C appartiennent à des classes de distribution spécifiques. Comme énoncé précédemment, l'avantage de cette approche est que la copule C, caractérisant la dépendance entre X et Y, peut-être choisie indépendamment des caractéristiques relatives aux marginales.

L'objectif de cette section est de fournir une description détaillée des procédures d'estimation existantes pour les fonctions copules. Evidemment, en fonction des propriétés des copules étudiées, ces procédures vont nous conduire à choisir des méthodes paramétriques, semi-paramétrique ou encore non-paramétrique pour réaliser les estimations.

Concernant l'approche paramétrique, on suppose que la copule inconnue appartient à l'une des familles paramétriques décrites à la section 4.7, page 56.

Il existe deux méthodes paramétriques. La première consiste à maximiser la fonction de vraisemblance pour l'ensemble des paramètres simultanément, ceux de la copule et des marginales. Cette méthode de calcul est très exigeante et augmente considérablement les difficultés. Par conséquent, une approche plus simple mais moins précise a été développée. Il s'agit de la méthode *IFM*, Joe [1997] [26], qui consiste à estimer dans un premier temps les paramètres des marginales pour ensuite les prendre en compte dans la vraisemblance de la copule. Enfin, cette vraisemblance totale est maximisée en respectant les paramètres de la copule. Ces deux méthodes paramétriques sont détaillées en annexe .8, page 146.

Le succès de cette méthode repose sur le fait de pouvoir trouver au préalable un modèle paramétrique pour caractériser les marginales, ce qui n'est pas souvent le cas. Il serait donc préférable d'avoir une méthode qui n'impose rien quant à la caractérisation des marginales. De nombreux chercheurs, Demarta, McNeil [2004] [22] et Romano [2002] ont proposé une approche semi-paramétrique, la méthode *CML (Canonical Maximum Likelihood)* qui n'impose aucune hypothèse concernant la caractérisation des marginales. Les fonctions de répartition empiriques univariées sont intégrées dans la vraisemblance. Cette méthode a été retenue et est utilisée en pratique dans l'application, elle est donc détaillée *infra*, dans le corps de cette section.

Une alternative consiste à appliquer la méthode classique des moments en utilisant les relations existantes entre les paramètres de la copule et des mesures d'association. Cette méthode est illustrée en annexe .9, page 148.

Enfin, une approche entièrement non-paramétrique peut être considérée, ainsi la copule peut être estimée à partir de la copule empirique ou en adoptant des méthodes basées sur le noyau. Cette approche est mise en pratique dans l'application (section 7.3.1, page 108).

- **Une approche semi-paramétrique : La méthode *CML***

Les méthodes paramétriques sont assez faciles à mettre en place lorsque nous sommes en présence d'un nombre de variables faible. Cependant, dès que le nombre de variables augmente, il convient d'utiliser une autre méthode, les méthodes semi-paramétriques

peuvent être la solution.

De plus, les méthodes paramétriques sont sensibles à une erreur de spécification des marginales puisque celles-ci interviennent directement dans le calcul de la log-vraisemblance. Pour cette raison, une troisième méthode est utlilisée en pratique, pour laquelle il n'est pas indispensable de procéder à une estimation des marginales.

Il est important d'obtenir un estimateur robuste α de la copule. Pour ce faire, la méthode *CML* permet de réaliser ce travail sans imposer de choisir préalablement une loi paramétrique connue pour approcher les distributions marginales. En effet, les marginales sont estimées par les fonctions de répartition empiriques suivantes :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} = \frac{R_i}{n} \qquad G_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \leq y\}} = \frac{S_i}{n} \tag{4.44}$$

où R_i est le rang de X_i parmi $(X_1, ..., X_n)$ et S_i est le rang de Y_i parmi $(Y_1, ..., Y_n)$.

Le paramètre α est estimé en maximisant la pseudo-vraisemblance, basée sur les pseudo-observations et nous obtenons :

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} log\left\{c_\alpha(F_n(X_i), G_n(Y_i))\right\} \tag{4.45}$$

Comme on peut le voir ci-dessus, l'estimateur relatif à la méthode *CML* présente l'avantage de ne pas reposer sur une spécification des marginales. Autrement dit, cette méthode permet de déterminer le paramètre de la copule sans prendre en considération la forme paramétrique des lois marginales.

Troisième partie

Application et résultats

Chapitre 5

L'analyse de la base de données des sinistres

Il est intéressant de réaliser une analyse des sinistres clos, ces derniers constituant la base de référence de la variable coût du sinistre. Ensuite, les délais de règlement et les sinistres classés sans suite seront analysés. L'ensemble des statistiques présenté *infra* a été réalisé à partir de données *"as if"*, construites d'après la méthodologie décrite section 2.3, page 23.

5.1 Analyse de la répartition des sinistres

La répartition des sinistres clos ou en cours, de l'ensemble du portefeuille, est la suivante :

Etats sinistres	**Effectif**	**Proportion**
Sinistres en cours	504 826	2,11 %
Sinistres clos	23 452 236	97,89 %
Total	23 957 062	100 %

TABLE 5.1 – Répartition des sinistres selon leur état (ensemble du portefeuille)

La répartition des sinistres clos ou en cours, survenus en 2009, est la suivante :

Etats sinistres	**Effectif**	**Proportion**
Sinistres en cours	329 458	24,84 %
Sinistres clos	996 627	75,16 %
Total	1 326 085	100 %

TABLE 5.2 – Répartition des sinistres survenus en 2009 selon leur état

	Code de la garantie																								Total	
	A		C		D		E		F		J		K		M		S		T		V		W			
	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%
Indicateur de clôture des sinistres																										
N	199155	0.83	215579	0.90	3341	0.01	19517	0.08	22777	0.10	234	0.00	27066	0.11	162	0.00	20	0.00	738	0.00	6929	0.03	9308	0.04	504826	2.11
O	4830803	20.16	9809901	40.95	112915	0.47	1647564	6.88	535866	2.24	2704	0.01	6072323	25.35	14391	0.06	2231	0.01	88286	0.37	133438	0.56	201814	0.84	23452236	97.89
Total	5029958	21.00	10025480	41.85	116256	0.49	1667081	6.96	558643	2.33	2938	0.01	6099389	25.46	14553	0.06	2251	0.01	89024	0.37	140367	0.59	211122	0.88	23957062	100.00

FIGURE 5.1 – Proportion de sinistres clôturés par garanties

	Code de la garantie																								Total	
	A		C		D		E		F		J		K		M		S		T		V		W			
	Indicateur de clôture des sinistres																									
	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Années de survenance																										
1993	0.00	1.26	0.00	2.36	0.00	0.02	0.00	0.53	0.00	0.12	.	0.00	0.00	1.33	0.00	0.00	.	0.00	.	0.00	.	.	.	0.12	0.00	5.75
1994	0.01	1.27	0.00	2.40	0.00	0.02	0.00	0.54	0.00	0.14	.	0.00	0.00	1.49	.	0.00	.	0.00	.	0.00	.	.	.	0.09	0.01	5.96
1995	0.00	1.29	0.00	2.45	0.00	0.02	0.00	0.51	0.00	0.15	.	0.00	0.00	1.27	.	0.00	.	0.00	.	0.00	.	.	0.00	0.01	0.01	5.72
1996	0.04	1.23	0.04	2.35	0.00	0.02	0.01	0.49	0.00	0.15	0.00	0.00	0.02	1.26	.	0.00	.	0.00	0.00	0.00	.	.	0.00	0.01	0.11	5.52
1997	0.01	1.31	0.01	2.45	0.00	0.02	0.00	0.47	0.00	0.16	0.00	0.00	0.01	1.38	.	0.00	.	0.00	.	0.00	.	.	0.00	0.01	0.02	5.80
1998	0.00	1.35	0.00	2.56	0.00	0.02	0.00	0.47	0.00	0.17	0.00	0.00	0.00	1.50	.	0.00	0.00	0.00	0.00	0.01	.	.	0.00	0.03	0.01	6.12
1999	0.00	1.41	0.00	2.75	0.00	0.03	0.00	0.46	0.00	0.18	0.00	0.00	0.00	1.51	0.00	0.01	.	0.00	0.00	0.26	.	.	0.00	0.06	0.01	6.68
2000	0.00	1.37	0.00	2.73	0.00	0.03	0.00	0.46	0.00	0.17	0.00	0.00	0.00	1.48	.	0.01	.	0.00	0.00	0.01	.	.	0.00	0.07	0.01	6.32
2001	0.00	1.34	0.00	2.73	0.00	0.03	0.00	0.48	0.00	0.15	0.00	0.00	0.00	1.50	0.00	0.00	.	0.00	0.00	0.00	.	.	0.00	0.03	0.01	6.27
2002	0.01	1.25	0.00	2.60	0.00	0.03	0.00	0.45	0.00	0.14	0.00	0.00	0.00	1.57	.	0.01	.	0.00	.	0.01	.	.	0.00	0.06	0.01	6.11
2003	0.01	1.15	0.00	2.48	0.00	0.03	0.00	0.40	0.00	0.12	0.00	0.00	0.00	1.66	0.00	0.01	.	0.00	0.00	0.01	.	.	0.00	0.06	0.01	5.92
2004	0.01	1.14	0.01	2.43	0.00	0.03	0.00	0.36	0.00	0.12	0.00	0.00	0.00	1.57	.	0.00	.	0.00	0.00	0.01	0.00	0.05	0.00	0.03	0.03	5.74
2005	0.02	1.13	0.01	2.35	0.00	0.04	0.00	0.32	0.00	0.11	0.00	0.00	0.00	1.55	.	0.00	0.00	0.01	0.00	0.00	0.00	0.12	0.00	0.05	0.03	5.68
2006	0.03	1.09	0.02	2.30	0.00	0.03	0.00	0.29	0.00	0.11	0.00	0.00	0.00	1.58	.	0.00	.	0.00	0.00	0.01	0.00	0.11	0.00	0.05	0.05	5.58
2007	0.05	1.07	0.04	2.28	0.00	0.04	0.00	0.26	0.01	0.10	0.00	0.00	0.00	1.56	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.11	0.00	0.04	0.11	5.47
2008	0.16	0.92	0.11	2.13	0.00	0.03	0.01	0.22	0.02	0.08	0.00	0.00	0.01	1.58	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.10	0.00	0.03	0.32	5.10
2009	0.48	0.57	0.66	1.62	0.01	0.02	0.05	0.17	0.05	0.04	0.00	0.00	0.06	1.57	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.08	0.04	0.07	1.38	4.16
Total	0.83	20.16	0.90	40.95	0.01	0.47	0.08	6.88	0.10	2.24	0.00	0.01	0.11	25.35	0.00	0.06	0.00	0.01	0.00	0.37	0.03	0.56	0.04	0.84	2.11	97.89

FIGURE 5.2 – Proportion de sinistres clôturés par garanties et par années de survenance

	Code de la garantie																								Total		%
	A		C		D		E		F		J		K		M		S		T		V		W				
	Indicateur de clôture des sinistres																										
	N	O																							N	O	
	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	
Années de survenance																											
1993	181	302384	103	564227	1	5408	3	127745	33	29567	.	100	6	317707	1	90	.	24	.	560	.	.	.	28833	328	1376645	5.75
1994	1252	304101	190	575181	1	5445	5	130449	28	34620	.	153	9	356894	.	33	.	30	.	641	.	.	.	20518	1485	1428065	5.97
1995	1107	309923	364	586548	4	5382	88	121968	43	36390	.	211	40	305132	.	60	.	52	.	874	.	.	3	3552	1649	1370092	5.73
1996	9413	295049	9957	563204	21	5373	2120	116366	40	36340	2	256	4613	301678	.	11	.	26	2	942	.	.	7	2939	26175	1322184	5.63
1997	1578	312841	1535	586949	7	5413	369	111583	74	38253	2	267	1269	330226	.	752	.	17	.	663	.	.	1	2826	4835	1389790	5.82
1998	693	324144	462	614145	10	5923	57	112693	89	41090	5	280	43	358536	.	374	1	110	2	1279	.	.	3	8274	1365	1466848	6.13
1999	844	338502	252	659130	2	6050	24	110842	114	42596	6	273	46	362747	14	2193	.	57	11	62346	.	.	2	14983	1315	1599719	6.68
2000	1004	327918	309	654146	2	6073	31	109706	145	41346	6	245	40	354058	.	1510	.	40	1	3034	.	.	1	16123	1539	1514199	6.33
2001	1150	321827	350	653223	5	8037	47	114152	166	36488	11	230	58	359738	2	866	.	43	1	951	.	.	1	7341	1791	1502896	6.28
2002	1311	298823	535	621946	9	7180	83	107668	248	33404	18	210	809	375265	.	2450	.	24	.	1386	.	.	1	14263	3014	1462619	6.12
2003	1734	276241	676	593511	7	7040	116	96651	298	29618	19	139	519	396570	1	2498	.	25	1	2032	.	.	2	14745	3373	1419070	5.94
2004	3532	274122	2066	581415	39	6917	346	86336	462	28760	27	130	255	376841	.	99	.	18	5	2072	87	11255	97	7523	6916	1375488	5.77
2005	4252	271502	1989	562701	47	8425	214	77395	753	27444	22	91	242	370427	.	878	8	1557	2	692	86	27986	17	11925	7632	1361023	5.71
2006	6807	261305	3673	550778	51	8366	323	69836	1157	25835	39	57	326	377630	.	414	.	114	4	1912	133	26921	16	12473	12529	1335641	5.63
2007	12623	256471	8928	545790	111	8505	541	62028	1999	24778	40	34	781	373389	3	501	1	55	2	2021	274	25465	25	10592	25328	1309629	5.57
2008	37797	219529	27172	509170	411	7584	2294	51954	4537	19980	24	22	2448	379384	49	1252	2	31	38	1449	981	23589	341	7747	76094	1221691	5.42
2009	113877	136121	157018	387837	2613	5794	12856	40192	12591	9357	13	6	15562	376101	92	410	8	8	669	5432	5368	18222	8791	17157	329458	996637	5.54
Total	199155	4830803	215579	9809901	3341	112915	19517	1647564	22777	535866	234	2704	27066	6072323	162	14391	20	2231	738	88286	6929	133438	9308	201814	504826	23452236	100.00

FIGURE 5.3 – Nombre de sinistres clôturés par garanties et par années de survenance

Les garanties associées aux codes présents dans les tableaux sont les suivantes : (A) Responsabilité Civile (matérielle et corporelle) ; (C) Dommage ; (D) Incendie ; (E) Vol ; (F) Recours ; (J) Protection Juridique ; (K) Bris de glace ; (M) Catastrophe Naturelle ; (S) Attentat ; (T) Tempête ; (V) Acte Vandalisme ; (W) Climatique.

La garantie Acte Vandalisme a été mise en place en 2004. Auparavant, elle n'était pas proposée aux assurés.

Les précédents tableaux ne permettent pas de faire la distinction entre la Responsabilité Civile corporelle et matérielle. Or, la RC corporelle est une branche "longue" et la RC matérielle est une branche "courte", on peut donc s'attendre à un taux de sinistres en cours, pour la RC corporelle, supérieur à celui relatif à la RC matérielle :

Etats sinistres	**Effectif**	**Proportion**
Sinistres en cours	35 829	4,89 %
Sinistres clos	696 470	95,11 %
Total	732 299	100 %

TABLE 5.3 – Répartition des sinistres Responsabilité Civile corporelle selon leur état

Etats sinistres	**Effectif**	**Proportion**
Sinistres en cours	163 326	3,8 %
Sinistres clos	4 134 333	96,2 %
Total	4 297 659	100 %

TABLE 5.4 – Répartition des sinistres Responsabilité Civile matérielle selon leur état

Voici la répartition des sinistres survenus en 2009 pour les deux dernières branches :

Etats sinistres	**Effectif**	**Proportion**
Sinistres en cours	17 445	58,78 %
Sinistres clos	12 235	41,22 %
Total	29 680	100 %

TABLE 5.5 – Répartition des sinistres RC corporelle survenus en 2009 selon leur état

Etats sinistres	Effectif	Proportion
Sinistres en cours	96 432	43,77 %
Sinistres clos	123 886	56,23 %
Total	220 318	100 %

TABLE 5.6 – Répartition des sinistres RC matérielle survenus en 2009 selon leur état

La répartition des 23 957 062 sinistres étudiés, par année de survenance et selon qu'ils soient clos ou en cours, est la suivante :

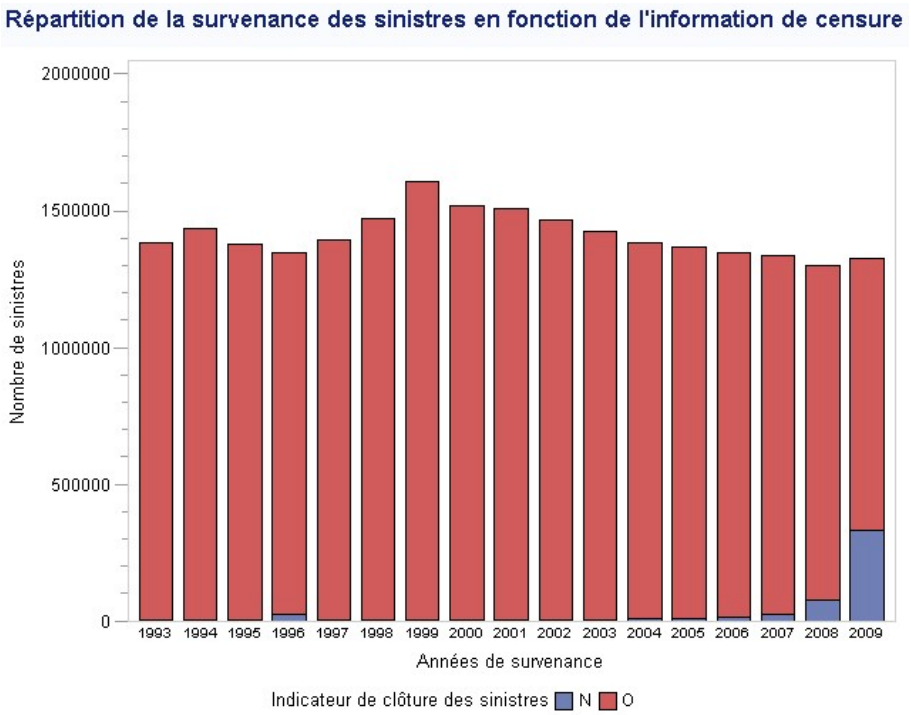


FIGURE 5.4 – Répartition de la survenance des sinistres intervenus sur l'ensemble du portefeuille en fonction de l'information de censure

On observe un nombre important de sinistres survenus en 1999. Ce nombre important est lié à la tempête de décembre 1999 qui a impacté des garanties facultatives du type Tempête.

Pour davantage de lisibilité, ce tableau restitue les proportions de sinistres clos ou en cours pour chaque année de survenance :

	Indicateur de clôture des sinistres		Total	%
	N	O		
	Effectif	Effectif	Effectif	
Années de survenance				
1993	328	1376645	1376973	5.75
1994	1485	1428065	1429550	5.97
1995	1649	1370092	1371741	5.73
1996	26175	1322184	1348359	5.63
1997	4835	1389790	1394625	5.82
1998	1365	1466848	1468213	6.13
1999	1315	1599719	1601034	6.68
2000	1539	1514199	1515738	6.33
2001	1791	1502896	1504687	6.28
2002	3014	1462619	1465633	6.12
2003	3373	1419070	1422443	5.94
2004	6916	1375488	1382404	5.77
2005	7632	1361023	1368655	5.71
2006	12529	1335641	1348170	5.63
2007	25328	1309629	1334957	5.57
2008	76094	1221691	1297785	5.42
2009	329458	996637	1326095	5.54
Total	504826	23452236	23957062	100.00

FIGURE 5.5 – Répartition de la survenance des sinistres intervenus sur l'ensemble du portefeuille en fonction de l'information de censure

Ceci illustre le fait que certains sinistres survenus en 1993 peuvent être encore en cours en 2009, notamment dans le cas de sinistres relatifs à des branches "longues". Aussi, plus l'année de survenance est proche de la date d'étude, plus le nombre de sinistres en cours est important.

5.2 Analyse des montants de sinistres clos

Le tableau ci-dessous restitue les principales statistiques descriptives relatives aux montants des sinistres clos et autres variables associées. La variable relative aux montants des frais d'épave au 31/12/2009 n'est pas renseignée avant 1997, raison pour laquelle elle présente des données manquantes.

Libellé	N	Nbre manquant	Minimum	Quartile inférieur	Médiane	Moyenne	Somme	Ecart-type	Quartile supérieur	Maximum
Montants des règlements cumulés au 31/12/2009	23452236	0	0.0	0.0	238.2	845.0	19817111504	8126.3	731.7	5770132.0
Montants des règlements commerciaux cumulés au 31/12/2009	23452236	0	0.0	0.0	0.0	3.7	86492011.8	87.0	0.0	42311.8
Montants des frais d'experts et honoraires versés au 31/12/2009	23452236	0	0.0	0.0	0.0	50.4	1182256123.9	135.2	90.2	105530.6
Montants des abandons de recours enregistrés au 31/12/2009	23452236	0	0.0	0.0	0.0	15.1	353661765.0	205.0	0.0	171160.1
Montants des frais d'épave au 31/12/2009	22712611	739625	0.0	0.0	0.0	1.6	35636102.3	33.0	0.0	71261.5
Stock de provision de sinistres vu au 31/12/2009	23452236	0	0.0	0.0	0.0	7.4	173861359.6	3730.4	0.0	7395561.5
Montants des recours encaissés au 31/12/2009	23452236	0	0.0	0.0	0.0	57.8	1355453073.7	1268.2	0.0	2868912.6
Stock de prévisions de recours au 31/12/2009	23452236	0	0.0	0.0	0.0	1.7	40285852.3	586.8	0.0	1446336.9

FIGURE 5.6 – Statistiques descriptives relatives aux montants des sinistres clos et autres variables quantitatives associées

La moyenne des coûts de sinistres est nettement supérieure à la médiane, ce qui traduit le fait que la base est constituée de sinistres caractérisés par des coûts très élevés mais qui ne représentent qu'une faible proportion.

Cette disparité dans les montants des règlements cumulés au 31/12/2009 se retrouve également dans l'étude des coûts par année de survenance. En se basant sur les données des sinistres clos, les années de survenance s'étalent de 1993 à 2009.

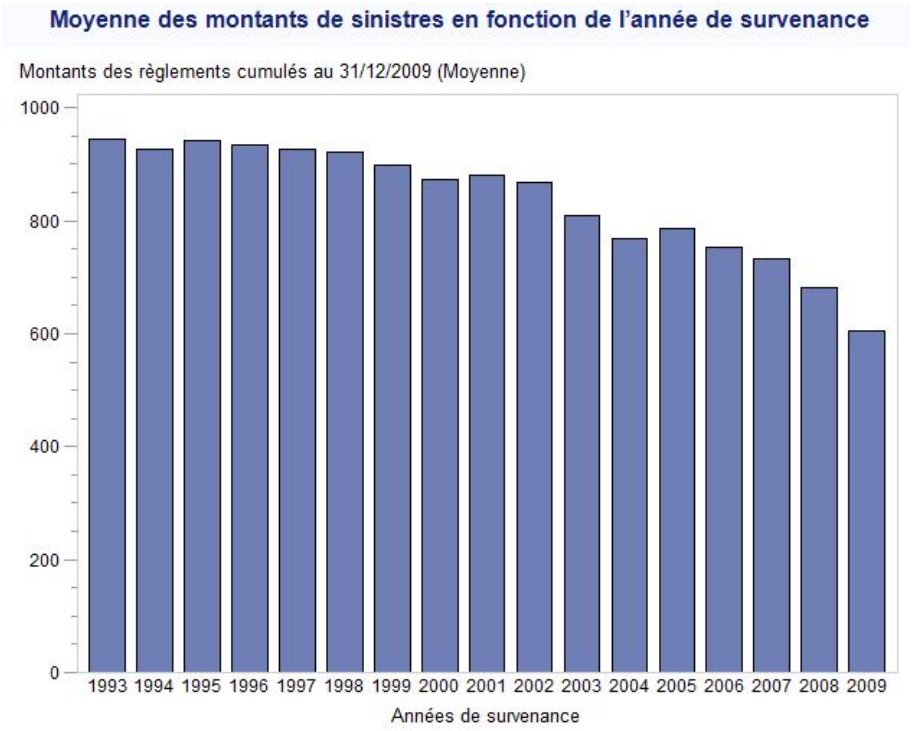


FIGURE 5.7 – Moyenne des montants de sinistres clos en fonction de l'année de survenance

En effectuant un zoom du graphique ci-dessus, le graphique suivant permet d'illustrer davantage la décroissance des montants moyens en fonction des années de survenance. Nous voyons que les sinistres qui se règlent vite ont un coût plus faible que la moyenne. En effet, les coûts des sinistres survenus en 2009 et clos en 2009 ont une moyenne nettement inférieure aux coûts des sinistres survenus en 2000 (par exemple) et clos entre les années 2000 et 2009. On peut donc déduire que le délai de règlement d'un sinistre est lié au montant de règlement cumulé relatif à ce dernier.

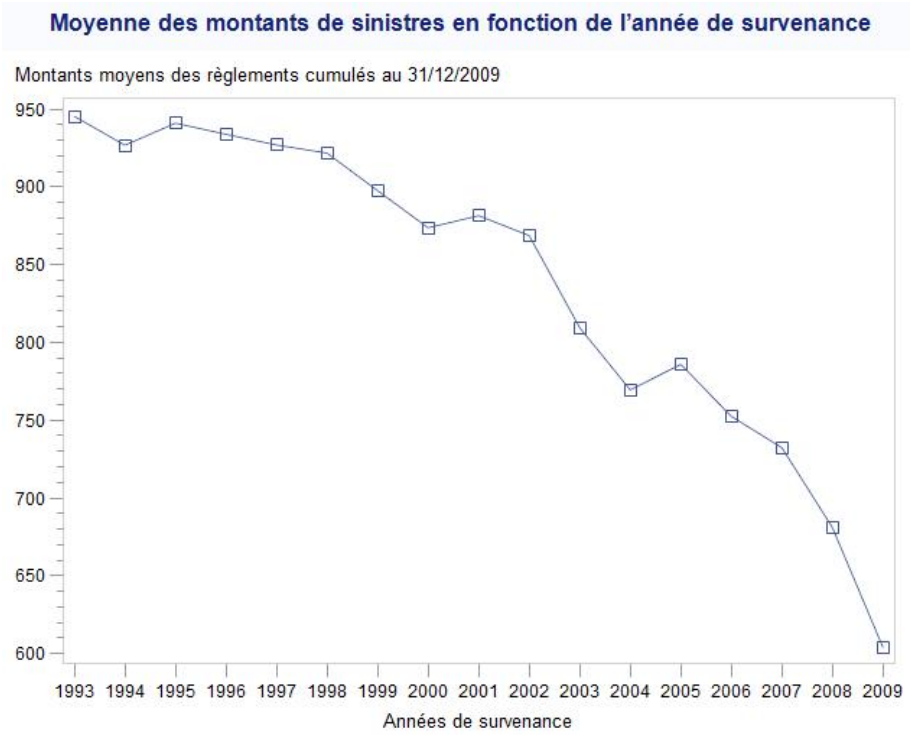


FIGURE 5.8 – Moyenne des montants de sinistres clos en fonction de l'année de survenance

5.3 Analyse des délais de règlements

Les délais de règlements des sinistres peuvent varier entre 0 et 16 ans. Un sinistre caractérisé par un délai de règlement égal à zéro signifie que le sinistre a totalement été réglé au cours de l'année pendant laquelle il est survenu.

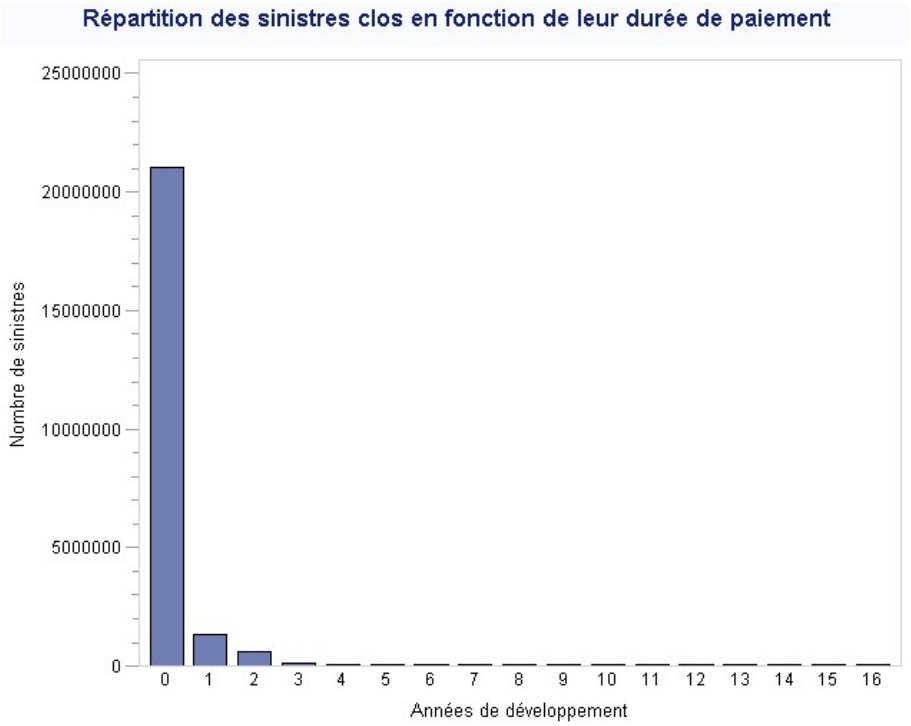


FIGURE 5.9 – Répartition des sinistres clos en fonction de leur durée de paiement

On constate que la majorité des sinistres présentent des délais de règlements inférieurs à un an. Autrement dit, la plupart des sinistres du portefeuille sont réglés et clôturés au cours de l'année de survenance. Le délai moyen de règlement est d'environ 5 mois. Il peut être intéressant de représenter le coût moyen d'un sinistre en fonction de sa durée de paiement.

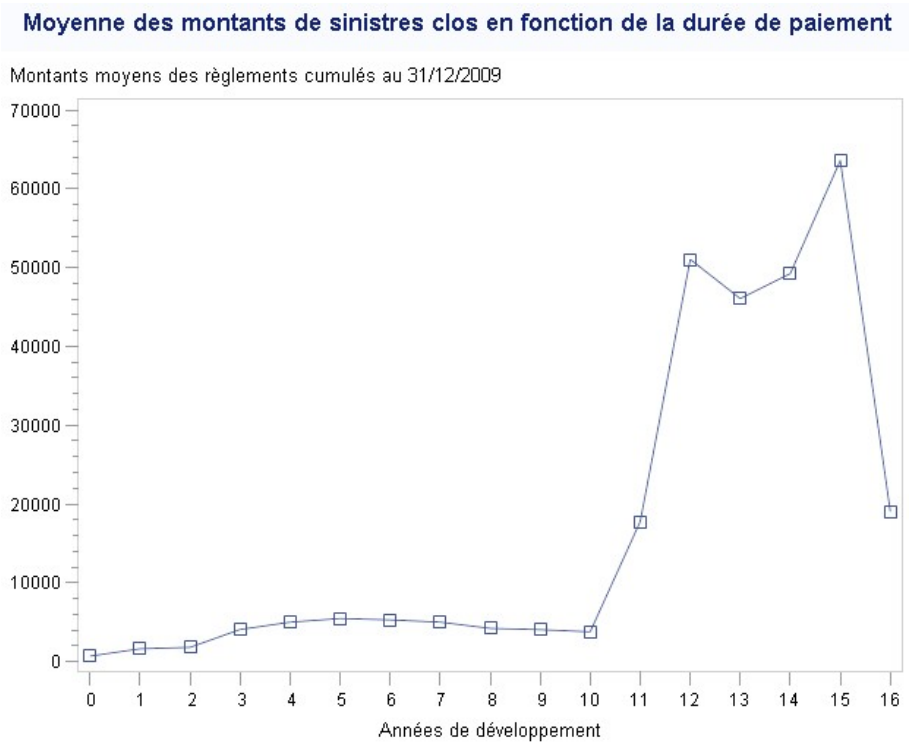


FIGURE 5.10 – Moyenne des montants de sinistres clos en fonction de la durée de paiement

D'après ce graphique, on peut dire que les sinistres pour lesquels la durée de paiement est élevée sont en moyenne caractérisés par des coûts importants, ce qui confirme l'intuition initiale.

De plus, on observe un saut important entre la moyenne des coûts des sinistres réglés au bout de 10 années et ceux réglés au bout de 12 ans. Ceci est la conséquence de la clôture des sinistres relatifs à des branches "longues", comme par exemple, ceux relatifs à de la Responsabilité Civile corporelle qui engendrent des coûts relativement élevés par rapport à la moyenne.

5.4 Analyse des sinistres classés sans suite

A travers cette section, il s'agit d'étudier la probabilité de classement sans suite pour un sinistre donné. Un sinistre classé sans suite en assurance automobile est un sinistre clos pour lequel l'application des règles en vigueur a aboutie à la décision de non prise en charge par l'assureur des frais occasionnés par un accident. La décision de l'expert fait suite au fait que le sinistre soit couvert ou non par les garanties souscrites par l'assuré. Le montant de règlement, ainsi que le montant de recours associé à un tel sinistre sont donc nuls. Cependant, des frais d'expert pourront être enregistrés.

Sinistres classés sans suite	Effectif	Proportion
Oui	2 382 179	9,94 %
Non	21 574 883	90,06 %
Total	23 957 062	100 %

TABLE 5.7 – Proportion de sinistres classés sans suite (ensemble du portefeuille)

Les proportions ainsi que les nombres de sinistres classés sans suite, par garanties et par années de survenance, sont présentés en annexe .1, page 135.

Voici les principales statistiques concernant les sinistres classés sans suite avant retraitement des valeurs aberrantes :

Variable	Libellé	N	Nbre manquant	Minimum	Quartile inférieur	Médiane	Moyenne	Somme	Ecart-type	Quartile supérieur	Maximum
reglement_sin_mt_AsIf	Montants des règlements cumulés au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	26629.4	5.4	0.0	5156.7
reglement_commercial_mt_AsIf	Montants des règlements commerciaux cumulés au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
frais_honoraires_et_presta_AsIf	Montants des frais d'experts et honoraires versés au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	4589.4	0.5	0.0	436.6
abandon_recours_mt_AsIf	Montants des abandons de recours enregistrés au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	1615.3	1.0	0.0	1615.1
dommage_frais_epave_mt_AsIf	Montants des frais d'épave au 31/12/2009	1960946	421233	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
provision_sin_mt_AsIf	Stock de provision de sinistres vu au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
recours_mt_AsIf	Montants des recours encaissés au 31/12/2009	2382179	0	0.0	0.0	0.0	0.0	1244.6	0.5	0.0	610.4
estim_recours_restants_mt_AsIf	Stock de prévisions de recours au 31/12/2009	2382179	0	0.0	0.0	0.0	0.1	134297.1	17.2	0.0	13761.0

FIGURE 5.11 – Nombre de sinistres classés sans suite par garanties et par années de survenance

Il persiste des données aberrantes car il est incompatible d'avoir, pour des sinistres sans suite, des montants non nuls pour les variables *règlement_sin_mt_AsIf*, *abandon_recours_mt_AsIf*, *recours_mt_AsIf* et *estim_recours_restants_mt_AsIf*. De ce fait, les 415 lignes associées à ces valeurs aberrantes ont été supprimées, ce qui représente peu de données relativement au volume de la base. Voici les principales statistiques concernant les sinistres classés sans suite après retraitement des valeurs aberrantes :

Variable	Libellé	N	Nbre manquant	Minimum	Quartile inférieur	Médiane	Moyenne	Somme	Ecart-type	Quartile supérieur	Maximum
reglement_sin_mt_AsIf	Montants des règlements cumulés au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
reglement_commercial_mt_AsIf	Montants des règlements commerciaux cumulés au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
frais_honoraires_et_presta_AsIf	Montants des frais d'experts et honoraires versés au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	2772.7	0.3	0.0	126.5
abandon_recours_mt_AsIf	Montants des abandons de recours enregistrés au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
dommage_frais_epave_mt_AsIf	Montants des frais d'épave au 31/12/2009	1960889	420875	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
provision_sin_mt_AsIf	Stock de provision de sinistres vu au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
recours_mt_AsIf	Montants des recours encaissés au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
estim_recours_restants_mt_AsIf	Stock de prévisions de recours au 31/12/2009	2381764	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

FIGURE 5.12 – Nombre de sinistres classés sans suite par garanties et par années de survenance (après retraitement des valeurs aberrantes)

Les applications réalisées sections 6, page 73, et 7, page 94, se basent sur les données, retraitées des valeurs aberrantes, présentées dans cette section.

Chapitre 6

Analyse des lois marginales des coûts relatifs aux garanties RC corporelle, RC matérielle et Dommage

6.1 La loi marginale des coûts relatifs à la garantie RC corporelle

6.1.1 Approche considérant le taux de sinistres encore en cours négligeable

Tout d'abord, le coût d'un contrat sinistré, sur une garantie donnée, est calculé et correspond à la somme du montant des règlements cumulés au 31/12/2009, du montant des frais d'experts et honoraires versés au 31/12/2009 et du montant des frais d'épave au 31/12/2009.

Le principe de cette approche repose sur l'ajustement d'une loi à partir des coûts relatifs aux sinistres clos uniquement. Cette approche est justifiée par le fait que la proportion de sinistres encore en cours pour la garantie RC corporelle est négligeable, de l'ordre de 4,89 % en prenant en compte toutes les années de survenance (de 1993 à 2009).

L'objectif final étant d'estimer le montant de Provision pour Sinistres A Payer (PSAP) relatif aux sinistres survenus en 2009, nous retiendrons, pour cette approche, l'hypothèse que les paramètres sont stables dans le temps, c'est à dire pour toutes les années de survenance.

Il s'agit dans un premier temps d'ajuster une loi statistique aux coûts relatifs à la garantie RC corporelle pour les années de survenance allant de 1993 à 2009 pour ensuite en déterminer les paramètres. S'agissant des lois marginales, l'ajustement est réalisé sur les coûts strictement supérieurs à zéro, soit 450 659 sinistres pour la garantie RC corporelle (contre 696 470 sinistres clos [1]).

A partir du calcul d'une espérance résiduelle que nous détaillerons par la suite, le montant de PSAP sera déterminé. Puis, disposant des triangles de paiements pour chacune des garanties, il sera possible de comparer le montant de provision déjà calculé avec celui obtenu avec la méthode de *Chain Ladder* et ainsi de justifier l'éventuel écart entre les deux montants de provision.

1. Ces chiffres sont énoncés à la section 5.1, page 63.

Cette comparaison est justifiée par le fait que la notion de coût définie en début de section est utilisée pour l'approche de provisionnement ligne à ligne et est cohérente avec l'utilisation des triangles de paiements pour déterminer le montant de PSAP avec *Chain Ladder*. De plus, la construction préalable de statistiques *"as if"* détaillée à la section 2.3, page 23, assure une cohérence permettant la comparaison des résultats relatifs à ces méthodes.

Enfin, disposer du stock de provision de sinistres vu au 31/12/2009 (provision "dossier-dossier" estimée par l'assureur) devrait nous permettre de prendre du recul quant aux résultats obtenus et ainsi de juger de la qualité des méthodes étudiées.

- Ajustement d'une loi statistique aux coûts :

Cette étude paramétrique consiste à approcher et à estimer la loi suivie par la variable représentant les coûts de sinistres à l'aide de distributions connues. Il a été décidé de tester l'adéquation des données avec plusieurs lois utilisées en pratique :

- la loi Exponentielle ;
- la loi Log-Normale ;
- la loi Normale ;
- la loi de Weibull ;
- la loi Gamma ;

et de retenir celle qui maximise la log-vraisemblance. Voici ci-dessous, les valeurs des log-vraisemblance obtenues pour les différentes lois testées :

Loi testée	**Exponentielle**	**Log-Normale**	**Normale**	**Weibull**	**Gamma**
Log-vraisemblance	-4 470 191	-4 336 861	-5 508 863	-4 271 250	-

TABLE 6.1 – Valeur de la log-vraisemblance pour chaque loi testée

L'ajustement avec la loi Gamma ne converge pas, cette loi n'est donc pas adaptée ici.

D'après ces résultats (obtenus à l'aide du logiciel R et de la fonction $fitdistr$ [2]), la log-vraisemblance la plus importante est associée à la loi de Weibull (cette loi est celle qui s'ajuste le mieux à nos données). Nous allons maintenant ajuster cette loi dans le but de mener à bien l'étude des coûts relatifs à la garantie RC corporelle.

- Estimation des paramètres de la loi retenue :

Une fois la loi choisie, il reste à déterminer les paramètres associés. Pour cela, nous avons eu recours à la méthode du maximum de vraisemblance qui est une méthode statistique courante, utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné. Voici les valeurs des paramètres obtenues pour la loi de Weibull :

2. Le lecteur intéressé pourra se référer à ce lien bibliographique [32].

Alpha (paramètre de forme)	0,67
Lambda (paramètre d'échelle)	4 628

TABLE 6.2 – Valeurs des paramètres pour la loi de Weibull ajustée

- Calcul de la Provision pour Sinistres A Payer (PSAP) :

Le choix de la loi statistique a permis de sélectionner le modèle paramétrique propre aux coûts des sinistres relatifs à la garantie étudiée. A partir de l'expression de ce modèle et des calculs développés dans le chapitre 3, page 26, l'espérance de la provision totale s'exprime ainsi :

$$E(P_i) = \frac{1}{S_{a,b}(c_i)} \times \int_{c_i}^{+\infty} S_{a,b}(t)dt \tag{6.1}$$

c_i représente les montants déjà réglés de sinistres encore en cours à la date d'observation. Cette expression nous donne la valeur de l'espérance de la provision individuelle à constituer pour chaque sinistre i. Il s'agit d'une espérance résiduelle correspondant au calcul de la charge résiduelle restant à payer au titre des sinistres encore en cours. L'ensemble des provisions individuelles est alors connu et il est possible de calculer la somme de ces espérances qui donne par linéarité l'espérance du montant total des provisions à constituer pour faire face aux sinistres en cours.

En ce qui concerne la garantie Responsabilité Civile corporelle, l'année de survenance 2009 est caractérisée par 29 680 sinistres survenus, dont 17 445 encore en cours. En appliquant la méthode décrite *supra* à ces 17 445 sinistres toujours en cours, on obtient un montant de provision totale à constituer de 126 millions d'euros.

- Comparaison avec la méthode de *Chain Ladder* :

La méthode de *Chain Ladder* est très utilisée en pratique et très simple à mettre en oeuvre dans le but de calculer des montants de PSAP. Cette méthode va servir de moyen de comparaison quant aux résultats obtenus avec les autres méthodes de provisionnement ligne à ligne mises en oeuvre dans ce mémoire.

La version déterministe de *Chain Ladder* appliquée dans ce mémoire est basée sur les triangles de règlements cumulés. Ces derniers sont obtenus suite à une agrégation des données individuelles des sinistres. Ce modèle suppose tout d'abord qu'il y ait une cadence (nommée cadence de règlements) entre les montants de paiements cumulés de deux années de développement successives. Ainsi, il suffit d'estimer l'ensemble des facteurs de développement pour être en mesure de compléter la totalité du triangle de règlements cumulés. L'estimateur naturel est défini par le ratio des sommes des montants réglés entre deux années de développement successives. A cause de la simplicité de ce modèle, les erreurs associées aux provisions calculées ne peuvent pas être quantifiées.

Les inconvénients de cette méthode peuvent être regroupés en deux catégories :

- Le premier concerne directement la prise en compte des caractéristiques du portefeuille de l'assureur. La méthode de *Chain Ladder* suppose que la composition du portefeuille n'évolue pas au cours du temps. En effet, afin d'obtenir des cadences les plus régulières

possibles, il ne faudrait pas ou peu de sinistres graves. Un portefeuille idéal pour l'application d'une telle méthode serait constitué d'un nombre important de contrats associés à une fréquence de sinistres significative. De plus, les modifications des règles de souscription ou encore de la gestion des sinistres influencent également les cadences de règlements.

- Le deuxième inconvénient se rapporte aux méthodes de calcul. Il est important de souligner que les coefficients de passage sont calculés en effectuant un ratio de sorte à ce que nous soyons en mesure de quantifier l'évolution des montants cumulés d'une année de développement à une autre. En d'autres termes, quelle que soit l'année de survenance, le montant à régler est déterminé par l'application de ces coefficients. La question qu'il faut alors se poser est la suivante : est-il légitime de considérer que les cadences futures coïncident avec celles du passé ? Ce n'est pas forcément le cas. Nous pouvons l'illustrer par différents exemples comme l'inflation qui peut augmenter le montant des règlements de l'assureur car elle peut avoir un effet "boule de neige". C'est à dire que si on considère une voiture accidentée à réparer, l'inflation peut modifier le coût de réparation (prix des pièces et coût de la main d'oeuvre, etc.).
  Un second point concerne l'utilisation de données agrégées qui fait perdre de l'information, et de surcroît, de la précision dans les calculs. Lorsque nous nous approchons de la fin des années de développement, nous avons de moins en moins de données pour calculer les coefficients de passage. Ainsi, le dernier coefficient se trouve égal au rapport entre deux valeurs uniquement et suppose être le même pour n'importe quelle année de survenance, ce qui n'est pas réaliste [3].

Afin de faire état de la pertinence de l'estimation effectuée par la méthode de *Chain Ladder*, le modèle de *Mack* a été utilisé pour construire un intervalle de confiance autour du montant de provision estimé. Cette méthode, illustrée dans le mémoire de Claire Guillaumin [2008] [33], nécessite de définir l'écart-type de la provision estimée pour l'année 2009 ($\hat{R}_i$). Ce dernier résulte du calcul du *standard error* : $se(\hat{R}_i) = \sqrt{mse(\hat{R}_i)}$.

Le modèle de Mack ne fait pas d'hypothèse sur la distribution des montants de provision estimés $\hat{R}_i$, cependant il devient nécessaire de poser une hypothèse paramétrique sur la forme de la distribution de ces derniers afin de définir les intervalles de confiance.

Pour effectuer une simplification, on utilise la loi Normale de moyenne la valeur estimée $\hat{R}_i$ et d'écart type $se(\hat{R}_i)$. L'intervalle de confiance à 95 % est alors donné par $\left[\hat{R}_i - 1,96.se(\hat{R}_i)\ ; \hat{R}_i + 1,96.se(\hat{R}_i)\right]$.

Compte tenu de l'objectif de ce mémoire, il n'a pas été jugé utile de détailler davantage les méthodes de *Chain Ladder* et de *Mack*. Cependant, le lecteur intéressé pourra se référer à l'ouvrage de Michel Denuit et de Arthur Charpentier [2004] [34], à l'ouvrage de Christian Partrat et al. [2007] [35] ainsi qu'au mémoire de Ngoc An Dinh et Gilles Chau [2012] [36].

A partir de la méthode de *Chain Ladder*, le montant de provision totale estimée pour 2009 est égal à 228 millions d'euros. Le modèle de *Mack* donne un intervalle de confiance à 95 % : [175 M€ ; 281 M€]. Le montant obtenu à partir du modèle paramétrique (126 millions d'euros) est nettement inférieur et ne rentre pas dans l'intervalle de confiance.

3. En pratique, les assureurs peuvent retraiter leur triangle à l'aide de la méthode *Tail Factor* en présence de branche "longue".

- Justification de l'écart trouvé entre les deux méthodes :

Nous avons constaté une différence significative d'environ 102 millions d'euros entre les deux méthodes présentées *supra*. Cet écart doit pouvoir s'expliquer en partie du fait que la méthode de *Chain Ladder* prend en compte les sinistres dits tardifs.

Nous disposons du triangle des nombres de sinistres pour la garantie Responsabilité Civile corporelle. A partir de ce dernier, il est possible d'estimer le nombre de sinistres *IBNR : Incurred But Not Reported* (ou sinistres dits tardifs)[4]. Ensuite, connaissant le nombre de sinistres *IBNR*, nous pouvons en déduire le coût moyen relatif à ces sinistres en multipliant l'espérance de la loi de Weibull[5] ajustée précédemment par le nombre de sinistres *IBNR*. Ainsi, nous obtenons 2 758 sinistres *IBNR* et un coût moyen déduit de l'espérance d'une loi de Weibull ajustée aux coûts des sinistres clos relatifs à la garantie Responsabilité Civile corporelle égale à 6 163 euros, soit un coût total d'environ 17 millions d'euros.

Ce résultat permet d'expliquer une petite partie seulement de l'écart constaté de 102 millions d'euros. Néanmoins, il est important de préciser qu'un retraitement préalable a été nécessaire sur le triangle des nombres de sinistres. En effet, dans certains cas, pour une année de survenance donnée, les nombres de sinistres cumulés diminuaient d'une année de règlement à une autre. Or, ces nombres sont des nombres cumulés, il est donc impossible, par définition, qu'ils diminuent d'une année de règlement à une autre. Ceci peut être la conséquence d'un nombre important de recours, ce qui provoquerait une suppression d'un certain nombre de sinistres dans la base.

Il a été décidé d'apporter une correction simple : pour une année de survenance donnée, remplacer les nombres de sinistres aberrants par le nombre de sinistres enregistrés lors de la précédente année de règlement. Cette correction relativement simple permet *a minima* d'avoir un triangle cohérent. Une approche plus prudente aurait pu être retenue afin d'assurer une cadence plus importante et donc des nombres de sinistres *IBNR* plus importants. Ceci aurait eu pour effet d'augmenter le coût total des sinistres *IBNR* et permis de justifier davantage l'écart de 102 millions constaté entre la méthode de *Chain Ladder* et celle relative à l'utilisation d'un modèle paramétrique.
Cependant, le manque d'information à ce sujet ne nous a pas permis d'aller plus loin dans cette démarche.

6.1.2 Approche prenant en compte les sinistres encore en cours à l'aide d'un modèle de censure

- Etude non paramétrique à l'aide de l'estimateur de *Kaplan-Meier*

L'objectif de cette partie est de pouvoir estimer la fonction de survie à l'aide de l'estimateur de *Kaplan-Meier* puis de comparer les fonctions de survie relatives aux trois garanties étudiées, aussi bien graphiquement qu'à partir de tests statistiques[6].

Comme énoncé précédemment dans la partie théorique à la section 3.3.4, page 32, l'estimateur de *Kaplan-Meier* est un estimateur non paramétrique de la fonction de survie notée $S(x)$. Ce dernier permet d'estimer la valeur empirique prise par le risque de sortie

4. Il s'agit des sinistres survenus mais pas encore déclarés à l'organisme assureur. Cette notion est largement développée dans le mémoire de Nicoleta Tripa [2009] [37].
5. L'espérance de la loi de Weibull est explicitée à la section 3.3.5, page 37, formule (3.22).
6. Cette comparaison sera effectuée à la section 6.4, page 91.

de l'état qui correspond, dans le cadre de ce mémoire, à la clôture du sinistre. Cette approximation ne nécessite pas la spécification d'une loi statistique. Elle est principalement utilisée dans le cas où l'on est en présence de données incomplètes telles que les sinistres censurés. Enfin, les méthodes de calcul habituellement utilisées pour calculer des durées de vie sont appliquées dans notre cas aux montants de sinistres.

Pour le calcul de l'estimateur de *Kaplan-Meier*, il est nécessaire d'avoir des données constituées des coûts des sinistres notées $X_i^* = \min(X_i, C_i)$ ainsi que de l'information de censure relative à ces sinistres. Cette dernière est notée D_i et vaut 1 si le sinistre est clos (si l'information sur celui-ci est complète), et 0 lorsque le sinistre est encore en cours au moment de l'étude (dans ce cas, il y a censure). Pour chaque sinistre i de la base, le couple (X_i^*, D_i) est une donnée connue.

L'estimation de la fonction de survie a été réalisée à l'aide du logiciel SAS et plus particulièrement à partir de la procédure $lifetest$ [7].

Voici pour les sinistres relatifs à la garantie Responsabilité Civile corporelle survenus en 2009, l'estimation de la fonction de survie à partir de *Kaplan-Meier* :

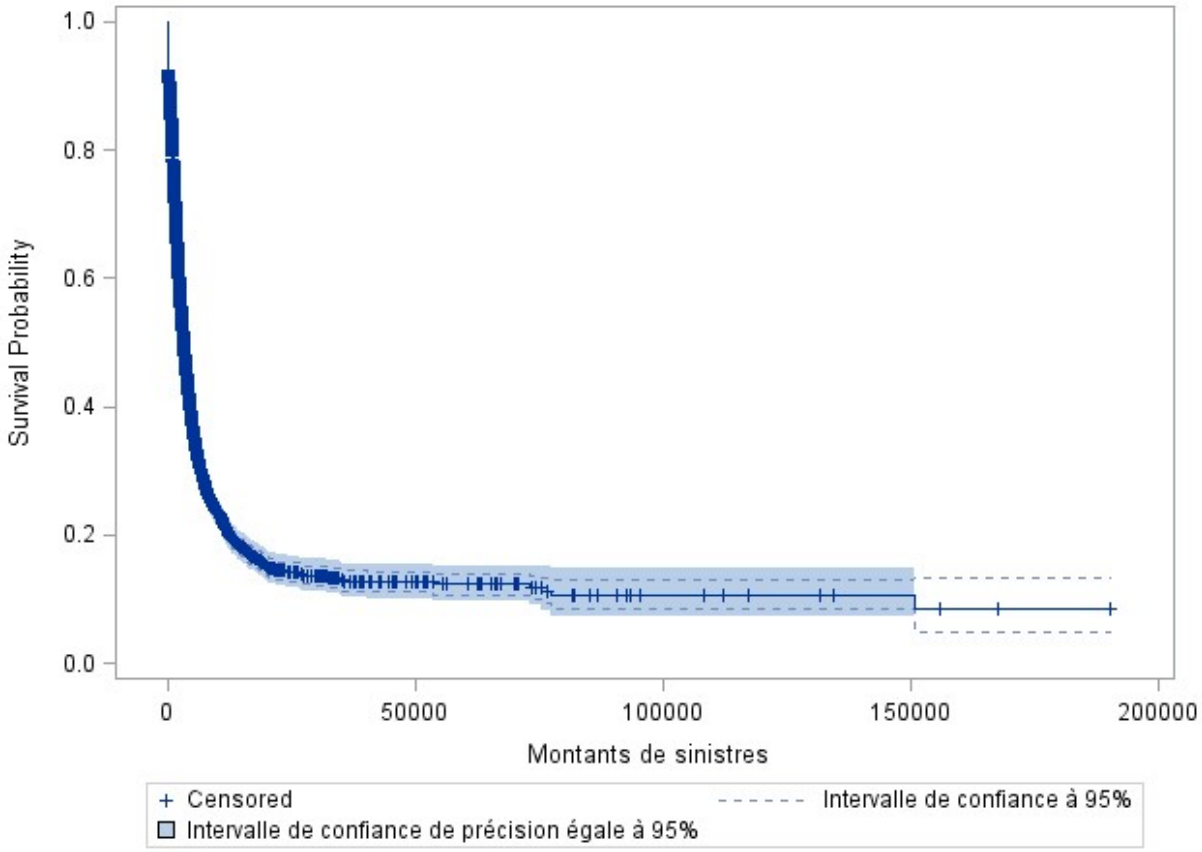


FIGURE 6.1 – Fonction de survie des sinistres survenus en 2009 pour la garantie RC corporelle

Les intervalles de confiance ponctuels et la bande de confiance à 95 %, ici de Nair [8], sont relativement proches de la fonction de survie estimée par *Kaplan-Meier*. Ceci permet d'effectuer graphiquement une première validation quant à l'estimation effectuée.

7. L'utilisation de cette procédure est détaillée à partir de ce lien bibliographique [38].
8. Ces éléments sont détaillés dans l'annexe .5, page 143.

- Etude paramétrique et calcul de la PSAP

Le pourcentage de sinistres survenus en 2009, encore en cours à la date d'observation, est égal à 58,78 % pour la garantie Responsabilité Civile corporelle. Ce taux de censure étant suffisamment élevé pour l'année de survenance 2009, il est légitime d'estimer le montant total de provision pour cette année de survenance en prenant en compte les sinistres censurés, soit 29 680 sinistres au total (sinistres clos et encore en cours).

Cette approche nécessite l'utilisation d'un modèle de censure. Sous SAS, l'utilisation de la procédure *lifereg*[9] va nous permettre d'ajuster une loi statistique à l'ensemble de nos montants de sinistres (en prenant en compte les censures) et ainsi d'estimer les paramètres par maximum de vraisemblance.

La loi qui maximise la log-vraisemblance est toujours la loi de Weibull. Les paramètres qui permettent d'ajuster au mieux cette loi aux montants des sinistres clos et encore en cours, survenus en 2009, sont les suivants :

Alpha (paramètre de forme)	0,72
Lambda (paramètre d'échelle)	7 297

TABLE 6.3 – Valeurs des paramètres pour la loi de Weibull ajustée

En appliquant la méthode relative au calcul de l'espérance de la provision totale détaillée *supra* (équation (6.1)), nous obtenons un montant de PSAP égale à 173 millions d'euros.

- Comparaison avec *C. L.* et justification de l'écart entre les deux méthodes

Pour mémoire, le montant de PSAP déterminé avec *Chain Ladder* est de 228 millions d'euros, soit un écart de 55 millions d'euros.

En considérant les *IBNR* pris en compte avec la méthode de *Chain Ladder*, mais non inclus dans la méthode ligne à ligne, nous devrions être en mesure de réduire cet écart. En effet, selon l'estimation réalisée *supra*, nous obtenons un coût moyen de 8 999 euros pour 2 758 sinistres *IBNR*, soit un coût total d'environ 25 millions d'euros.

Avec cette méthode de provisionnement ligne à ligne, le montant total de PSAP est de 198 millions d'euros, soit un écart de 30 millions d'euros avec la méthode de *Chain Ladder*.

6.1.3 Comparaison des deux approches de provisionnement ligne à ligne

Tout d'abord, ces deux approches ont la même finalité : elles permettent d'estimer le montant de PSAP *via* l'ajustement d'une loi statistique aux coûts des sinistres. Seulement, les résultats obtenus pour ces deux méthodes diffèrent et pour cause, l'ajustement de la loi statistique ne repose pas exactement sur les mêmes données et donc sur la même quantité d'informations.

Pour la première approche, il est légitime d'estimer les paramètres sur les coûts relatifs aux sinistres clos uniquement car le pourcentage de sinistres censurés est négligeable sur l'ensemble des années de survenance.

9. L'utilisation de cette procédure est détaillée à partir de ce lien bibliographique [38].

A contrario, la seconde approche montre qu'au regard d'un taux de censure significatif, il est nécessaire de prendre en compte les coûts relatifs aux sinistres censurés lors de l'ajustement de la loi statistique. Ces deux approches sont donc acceptables mais l'une est préférable en termes de qualité de l'ajustement et des résultats.

Ci-dessous un tableau de synthèse des résultats obtenus avec les deux approches ligne à ligne, *Chain Ladder*, *Mack* (pour l'intervalle de confiance) et les provisions "dossier-dossier" pour les 17 445 dossiers encore en cours en 2009 :

Méthodes utilisées	**Approche 1**	**Approche 2**	***Chain Ladder/ Mack*** ($IC_{95\%}$)	**Provision "dossier"**
PSAP	126 M€	173 M€	228 M€/ [175 M€ ; 281 M€]	176 M€
IBNR	17 M€	25 M€	-	-
Ecarts/Provision "dossier"	-50 M€	-3 M€	-	-

TABLE 6.4 – Synthèse des résultats obtenus pour la garantie RC corporelle

Pour mémoire, la provision obtenue avec *Chain Ladder* inclue les *IBNR* tandis que les provisions relatives aux deux approches ligne à ligne, ainsi que la provision "dossier-dossier", ne prennent pas en compte les *IBNR*. Elles peuvent donc être comparées hors *IBNR*. Il est délicat d'estimer un montant d'*IBNR* pour l'approche "dossier-dossier". Le montant de provision associé à cette dernière pourra donc difficilement être comparé de manière précise avec le montant relatif à la provision *Chain Ladder*.

En ce qui concerne les résultats de provision obtenus avec prise en compte des *IBNR*, l'approche 2 prenant en compte les sinistres censurés permet de se rapprocher du montant de PSAP obtenu à l'aide de la méthode de *Chain Ladder* (écart de 85 millions entre la première approche et *Chain Ladder*, réduit à 30 millions avec l'approche tenant compte des censures et *Chain Ladder*). Aussi, l'estimation des coûts des sinistres *IBNR* a permis de justifier un écart d'environ 25 millions pour l'approche 2 (montant certainement sous-estimé par de nombreux recours survenus, comme expliqué précédemment).
L'approche tenant compte des sinistres censurés paraît donc plus adaptée et plus prudente puisqu'elle aboutit à un montant de PSAP plus important.

Si l'on compare maintenant les provisions relatives aux deux approches ligne à ligne avec la provision "dossier" caractérisée comme étant notre référentiel pour cette étude, la provision de l'approche 2 sous-estime "seulement" de 3 millions la provision "dossier", contre 50 millions pour l'approche 1.
Même sans connaître le montant d'*IBNR* pour la provision "dossier", on remarque que, comparativement à l'approche 2, la méthode de *Chain Ladder* aboutit à une provision nettement plus éloignée du référentiel.

Compte tenu des résultats à proprement parler, l'approche qui considère les censures serait la plus pertinente. Considérons maintenant la quantité et la qualité d'informations utilisées pour les deux approches ligne à ligne.

L'avantage de la première approche est d'utiliser un historique important (sinistres survenus de 1993 à 2009) qui ne nécessite pas la prise en compte des censures compte tenu d'une proportion négligeable lors de l'ajustement d'une loi. L'inconvénient de cette approche est qu'il est nécessaire de supposer que les paramètres de la loi ajustée soient constants au cours du temps (pour toutes les années de survenance), ce qui n'est pas forcément le cas en pratique. Les paramètres estimés sont restitués dans le tableau suivant pour les différentes années de survenance :

Années de survenance	**1993**	**1994**	**1995**	**1996**
Alpha (paramètre de forme)	0,64	0,64	0,64	0,64
Lambda (paramètre d'échelle)	4 459	4 397	4 560	4 756

1997	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**
0,63	0,63	0,61	0,63	0,61	0,62	0,62
4 934	5 006	5 092	5 005	5 320	5 461	5 360

2004	**2005**	**2006**	**2007**	**2008**	**2009**
0,64	0,65	0,65	0,68	0,71	0,72
5 289	5 504	5 650	5 686	6 000	7 297

TABLE 6.5 – Valeurs des paramètres pour la loi de Weibull ajustée

On remarque que les paramètres de forme et d'échelle augmentent progressivement au cours du temps. Pour 2009, le paramètre d'échelle est nettement plus élevé que ceux des années précédentes. Pour mémoire, lors de l'estimation des paramètres sur l'ensemble des années de survenance (sans considérer les censures), nous avions un paramètre de forme égale à 0,67 et un paramètre d'échelle égale à 4 628.

Compte tenu de ces observations, l'hypothèse relative à l'approche 1, concernant la stabilité des paramètres au cours du temps, n'est pas valide pour la garantie Responsabilité Civile corporelle. De plus, l'évolution de ces paramètres met en évidence que le coût est instable au cours du temps (il est croissant : 6 200 euros en 1993 et 7 738 euros en 2009), ce qui explique les résultats médiocres obtenus avec l'approche 1.

Concernant la seconde approche, il est naturel de penser que d'ajuster une loi sur les coûts relatifs aux sinistres clos et encore en cours, en utilisant un modèle de censure, permet d'améliorer la qualité de l'ajustement. D'autant plus que cet ajustement est réalisé sur les coûts des sinistres survenus en 2009 et que le calcul de la PSAP se base uniquement sur les

coûts des sinistres encore en cours survenus en 2009.

Ainsi, pour estimer le montant de PSAP de la garantie Responsabilité Civile corporelle pour l'année 2009, il apparaît légitime de privilégier la seconde approche de provisionnement ligne à ligne permettant la prise en compte des censures (c'est à dire les coûts relatifs aux sinistres encore en cours à la date d'observation).

6.2 La loi marginale des coûts relatifs à la garantie RC matérielle

Tout d'abord, les mêmes démarches et hypothèses que celles effectuées pour l'analyse de la garantie RC corporelle sont posées pour l'étude de la garantie RC matérielle. Des justifications, résultats et analyses vont cependant être détaillés dans cette section.

6.2.1 Approche considérant le taux de sinistres encore en cours négligeable

De la même façon que pour la garantie RC corporelle, cette approche est justifiée par le fait que la proportion de sinistres encore en cours pour la garantie RC matérielle est négligeable, de l'ordre de 3,8 % en prenant en compte toutes les années de survenance (de 1993 à 2009).

L'ajustement de la loi statistique aux coûts relatifs à la garantie RC matérielle pour les années de survenance allant de 1993 à 2009 porte sur 3 335 912 données (coûts de sinistres strictement supérieurs à zéro) contre 4 134 333 sinistres clos [10].

- Ajustement d'une loi statistique aux coûts :

Voici les valeurs des log-vraisemblance obtenues pour les différentes lois testées :

Loi testée	**Exponentielle**	**Log-Normale**	**Normale**	**Weibull**	**Gamma**
Log-vraisemblance	-26 732 584	-26 695 608	-29 062 093	-26 493 692	-26 623 788

TABLE 6.6 – Valeur de la log-vraisemblance pour chaque loi testée

D'après ces résultats, la log-vraisemblance la plus importante est associée à la loi de Weibull. Nous allons maintenant ajuster cette loi afin de mener à bien l'étude des coûts relatifs à la garantie RC matérielle.

- Estimation des paramètres de la loi retenue :

De même que pour la garantie RC corporelle, nous avons eu recours à la méthode du maximum de vraisemblance. Voici les valeurs des paramètres obtenues pour la loi de Weibull :

10. Ces chiffres sont énoncés à la section 5.1, page 63.

Alpha (paramètre de forme)	1,11
Lambda (paramètre d'échelle)	1 163

TABLE 6.7 – Valeurs des paramètres pour la loi de Weibull ajustée

- Calcul de la Provision pour Sinistres A Payer (PSAP) :

En ce qui concerne la garantie Responsabilité Civile matérielle, l'année de survenance 2009 est caractérisée par 220 318 sinistres survenus, dont 96 432 encore en cours. En appliquant à ces 96 432 sinistres toujours en cours la méthode précédemment décrite, nous obtenons un montant de provision totale à constituer de 106 millions d'euros.

- Comparaison avec *C.L.* et justification de l'écart entre les deux méthodes :

A partir de la méthode de *Chain Ladder*, le montant de provision totale estimé est égal à 84 millions d'euros. Le modèle de *Mack* donne un intervalle de confiance à 95 % : [51 M€ ; 117 M€]. Le montant obtenu à partir du modèle paramétrique (106 millions d'euros) est supérieur mais rentre dans l'intervalle de confiance donné par le modèle de *Mack*.

Nous avons constaté une différence significative d'environ 22 Millions d'euros entre les deux méthodes présentées *supra*. Cet écart va augmenter du fait que la méthode de *Chain Ladder* tienne compte des sinistres dits tardifs.

En suivant la même logique que lors de l'étude de la garantie RC corporelle, nous obtenons 13 865 sinistres *IBNR* et un coût moyen (déduit de l'espérance d'une loi de Weibull[11] ajustée aux coûts des sinistres clos relatifs à la garantie Responsabilité Civile matérielle) égale à 1 119 euros, soit un coût total d'environ 16 millions d'euros.

Ce résultat aboutit à un écart de 38 millions d'euros entre les deux méthodes. Aussi, comme pour la garantie RC corporelle, un retraitement préalable similaire a été nécessaire sur le triangle des nombres de sinistres (la même correction a été effectuée).

6.2.2 Approche prenant en compte les sinistres encore en cours à l'aide d'un modèle de censure

- Etude non paramétrique à l'aide de l'estimateur de *Kaplan-Meier*

Voici pour les sinistres relatifs à la garantie Responsabilité Civile matérielle survenus en 2009, l'estimation de la fonction de survie à partir de *Kaplan-Meier* :

11. L'espérance de la loi de Weibull est explicitée à la section 3.3.5, page 37, formule (3.22).

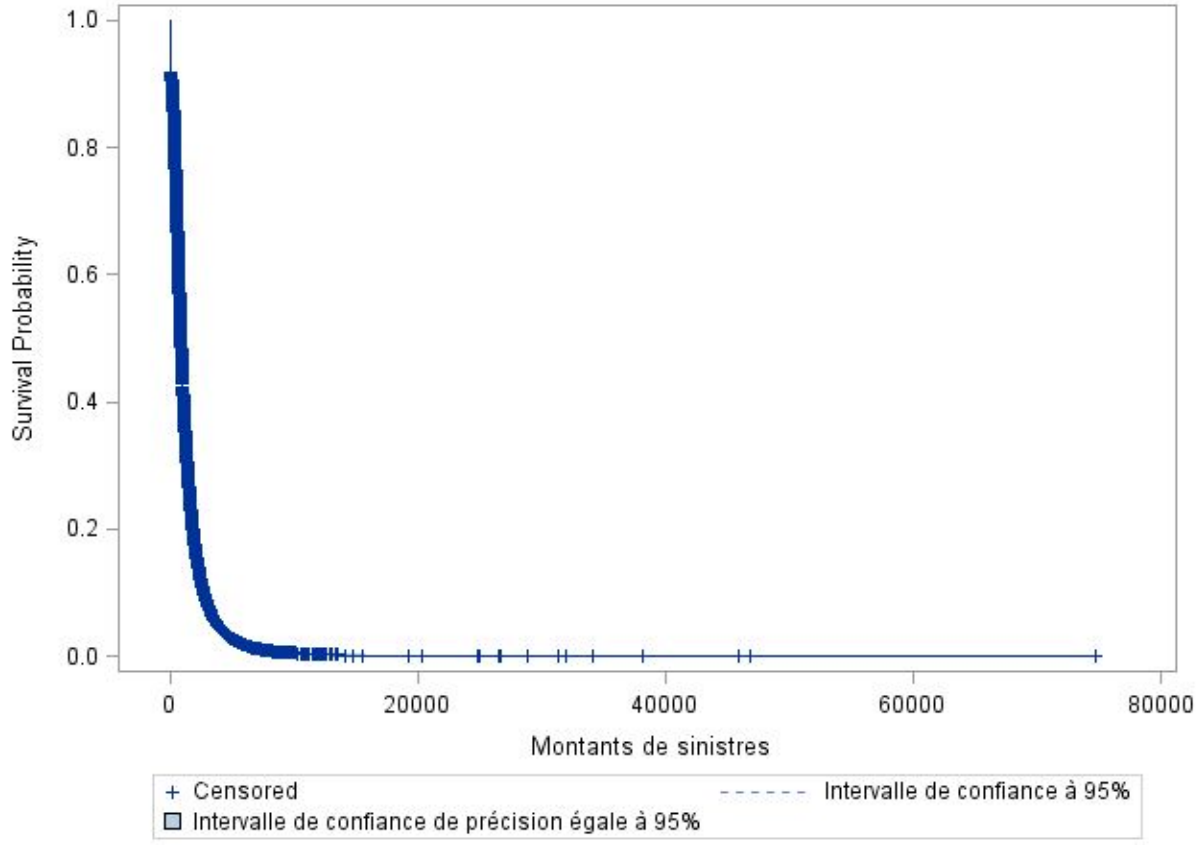


FIGURE 6.2 – Fonction de survie des sinistres survenus en 2009 pour la garantie RC matérielle

En raison d'un volume très important de données, on ne distingue pas les intervalles de confiance ponctuels et la bande de confiance à 95 %.

- Etude paramétrique et calcul de la PSAP

Le pourcentage de sinistres survenus en 2009 et encore en cours à la date d'observation est égale à 43,77 % pour la garantie Responsabilité Civile matérielle. Ce taux de censure étant suffisamment élevé pour l'année de survenance 2009, il est légitime d'estimer le montant total de provision pour cette même année en tenant compte des sinistres censurés, soit 220 318 sinistres au total (sinistres clos et encore en cours).

La loi qui maximise la log-vraisemblance est toujours la loi de Weibull. Les paramètres qui permettent d'ajuster au mieux cette loi aux montants des sinistres clos et encore en cours, survenus en 2009, sont donnés dans le tableau suivant :

Alpha (paramètre de forme)	1,17
Lambda (paramètre d'échelle)	1 257

TABLE 6.8 – Valeurs des paramètres pour la loi de Weibull ajustée

En appliquant la méthode relative au calcul de l'espérance de la provision totale détaillée *supra* (équation (6.1)), nous obtenons un montant de PSAP égale à 112 millions d'euros.

- Comparaison avec *C. L.* et justification de l'écart entre les deux méthodes

Pour mémoire, le montant de PSAP déterminé avec *Chain Ladder* est de 84 millions d'euros, soit un écart de 28 millions d'euros.

En considérant les *IBNR* pris en compte avec la méthode de *Chain Ladder*, mais non inclus dans la méthode ligne à ligne, cet écart va s'amplifier. En effet, d'après l'estimation réalisée *supra*, on obtient un coût moyen de 1 191 euros pour 13 865 sinistres *IBNR*, soit un coût total d'environ 17 millions d'euros.

Le montant total de PSAP avec la méthode ligne à ligne est donc de 129 millions d'euros, soit un écart de 45 millions d'euros avec la méthode de *Chain Ladder*.

6.2.3 Comparaison des deux approches ligne à ligne

Voici ci-dessous un tableau de synthèse des résultats obtenus avec les deux approches ligne à ligne, *Chain Ladder*, *Mack* (pour l'intervalle de confiance) et les provisions "dossier-dossier" pour les 96 432 dossiers encore en cours en 2009 :

Méthodes utilisées	**Approche 1**	**Approche 2**	***Chain Ladder/ Mack*** ($IC_{95\%}$)	**Provision "dossier"**
PSAP	106 M€	112 M€	84 M€/ [51 M€ ; 117 M€]	110 M€
IBNR	16 M€	17 M€	-	-
Ecarts/Provision "dossier"	-4 M€	+2 M€	-	-

TABLE 6.9 – Synthèse des résultats obtenus pour la garantie RC matérielle

En ce qui concerne les résultats de provision obtenus avec prise en compte des *IBNR*, l'approche 2 qui tient compte des sinistres censurés est plus éloignée du montant de PSAP obtenu à l'aide de la méthode de *Chain Ladder* que l'approche 1 (écart de 38 millions entre la première approche et *Chain Ladder* ; cet écart passe à 45 millions avec l'approche prenant en compte les censures et *Chain Ladder*).

Si l'on compare maintenant les provisions relatives aux deux approches ligne à ligne avec la provision "dossier" caractérisée comme étant notre référentiel pour cette étude, la provision de l'approche 2 surestime de 2 millions la provision "dossier", tandis que l'approche 1 la sous-estime de 4 millions. Enfin, même sans connaître le montant d'*IBNR* pour la provision "dossier", nous remarquons que, par rapport aux deux approches ligne à ligne, la méthode de *Chain Ladder* aboutit à une provision nettement plus éloignée du référentiel.

Compte tenu des résultats à proprement parler, l'approche considérant les censures serait la plus pertinente et la plus prudente. Considérons maintenant la quantité et la qualité d'informations utilisées pour les deux approches ligne à ligne.
Les paramètres estimés sont restitués dans le tableau suivant pour les différentes années de

survenance :

Années de survenance	1993	1994	1995	1996
Alpha (paramètre de forme)	1,10	1,11	1,12	1,10
Lambda (paramètre d'échelle)	1 018	1 041	1 073	1 137

1997	1998	1999	2000	2001	2002	2003
1,12	1,13	1,11	1,12	1,10	1,10	1,12
1 126	1 151	1 148	1 170	1 206	1 229	1 217

2004	2005	2006	2007	2008	2009
1,12	1,11	1,10	1,11	1,11	1,17
1 228	1 243	1 240	1 256	1 281	1 257

TABLE 6.10 – Valeurs des paramètres pour la loi de Weibull ajustée

Nous remarquons que les paramètres de forme sont constants et que les paramètres d'échelle augmentent progressivement au cours du temps. Pour mémoire, lors de l'estimation des paramètres sur l'ensemble des années de survenance (sans considérer les censures), nous avions un paramètre de forme égale à 1,11 et un paramètre d'échelle égale à 1 163.

Compte tenu de ces observations, l'hypothèse relative à l'approche 1 concernant la stabilité des paramètres au cours du temps n'est pas valide pour la garantie Responsabilité Civile matérielle (malgré que le paramètre de forme soit constant au cours du temps). De plus, l'évolution de ces paramètres met en évidence que le coût est instable au cours du temps (il est croissant : 982 euros en 1993 et 1 190 euros en 2009), ce qui explique que les résultats obtenus avec l'approche 1 sont légèrement moins satisfaisants que ceux obtenus avec l'approche 2.

Nous pouvons donc dire que pour estimer le montant de PSAP de la garantie Responsabilité Civile matérielle, pour l'année 2009, il paraît légitime de préférer la seconde approche de provisionnement ligne à ligne permettant la prise en compte des censures, c'est à dire les coûts relatifs aux sinistres encore en cours à la date d'observation. Enfin, nous remarquons que la méthode de *Chain Ladder* paraît inadaptée et conduit, pour cette garantie, à sous-estimer nettement le montant de PSAP [12].

12. En pratique, il est fréquent que les assureurs retraitent la dernière année de survenance du triangle à l'aide d'une approche par Sinistres/Primes.

6.3 La loi marginale des coûts relatifs à la garantie Dommage

Pour l'étude de la garantie Dommage, les mêmes démarches et hypothèses que celles effectuées pour l'analyse des deux garanties RC sont posées. Des justifications, résultats et analyses vont cependant être détaillés dans cette section.

6.3.1 Approche considérant le taux de sinistres encore en cours négligeable

Cette approche est justifiée du fait que la proportion de sinistres encore en cours pour la garantie Dommage est négligeable, de l'ordre de 2,15 % en prenant en compte toutes les années de survenance (de 1993 à 2009).

L'ajustement de la loi statistique aux coûts relatifs à la garantie Dommage pour les années de survenance allant de 1993 à 2009 porte sur 8 971 552 données (coûts de sinistres strictement supérieurs à zéro) contre 9 809 901 sinistres clos [13].

- Ajustement d'une loi statistique aux coûts :

Voici les valeurs des log-vraisemblance obtenues pour les différentes lois testées :

Loi testée	**Exponentielle**	**Log-Normale**	**Normale**	**Weibull**	**Gamma**
Log-vraisemblance	-68 253 999	-66 590 469	-78 547 504	-65 425 823	-

TABLE 6.11 – Valeur de la log-vraisemblance pour chaque loi testée

D'après ces résultats, la log-vraisemblance la plus importante est associée à la loi de Weibull. Nous allons maintenant ajuster cette loi pour mener à bien l'étude des coûts relatifs à la garantie Dommage.

- Estimation des paramètres de la loi retenue :

De même que pour les garanties RC, nous avons eu recours à la méthode du maximum de vraisemblance. Voici les valeurs des paramètres obtenues pour la loi de Weibull :

Alpha (paramètre de forme)	0,68
Lambda (paramètre d'échelle)	528

TABLE 6.12 – Valeurs des paramètres pour la loi de Weibull ajustée

13. Ces chiffres sont énoncés sectionsection 5.1, page 63.

- Calcul de la Provision pour Sinistres A Payer (PSAP) :

En ce qui concerne la garantie Dommage, l'année de survenance 2009 est caractérisée par 544 855 sinistres survenus, dont 157 018 encore en cours. En appliquant la méthode décrite précédemment à ces 157 018 sinistres toujours en cours, nous obtenons un montant de provision totale à constituer de 137 millions d'euros.

- Comparaison avec *C.L.* et justification de l'écart entre les deux méthodes :

A partir de la méthode de *Chain Ladder*, le montant de provision totale estimée est égal à 79 millions d'euros. Le modèle de *Mack* donne un intervalle de confiance à 95 % : [26 M€ ; 132 M€]. Le montant obtenu à partir du modèle paramétrique (137 millions d'euros) est nettement supérieur et ne rentre pas dans l'intervalle de confiance donné par le modèle de *Mack*.

La différence significative d'environ 58 millions d'euros entre les deux méthodes présentées *supra* va augmenter puisque la méthode de *Chain Ladder* prend en compte les sinistres dits tardifs.

En suivant la même logique que lors des études des garanties RC, nous obtenons 12 339 sinistres *IBNR* et un coût moyen déduit de l'espérance d'une loi de Weibull [14] ajustée aux coûts des sinistres clos relatifs à la garantie Dommage égale à 690 euros, soit un coût total d'environ 9 millions d'euros.

Ce résultat conduit à un écart de 67 millions d'euros. De plus, comme pour les garanties RC, un retraitement préalable similaire a été nécessaire sur le triangle des nombres de sinistres (la même correction a été effectuée).

6.3.2 Approche prenant en compte les sinistres encore en cours à l'aide d'un modèle de censure

- Etude non paramétrique à l'aide de l'estimateur de *Kaplan-Meier*

Voici pour les sinistres relatifs à la garantie Dommage survenus en 2009, l'estimation de la fonction de survie à partir de *Kaplan-Meier* :

14. L'espérance de la loi de Weibull est explicitée à la section 3.3.5, page 37, formule (3.22).

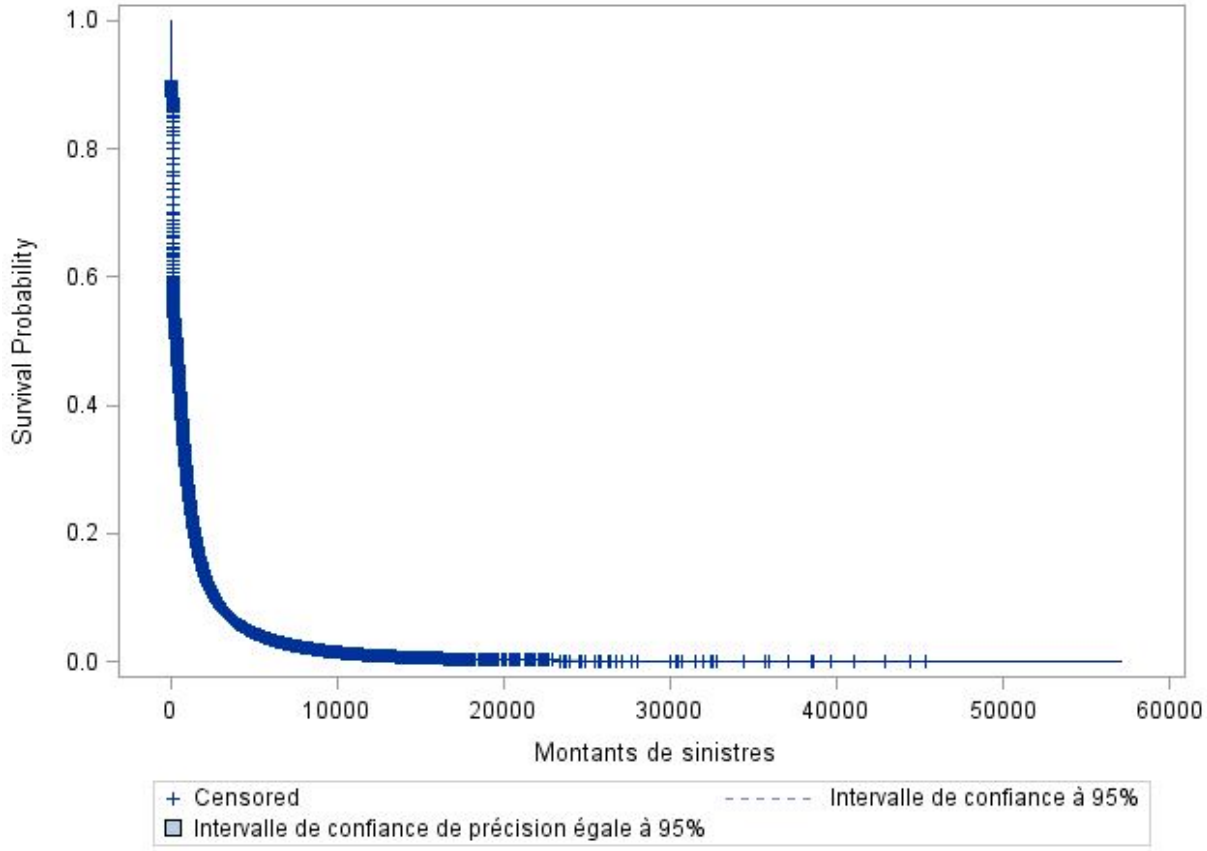


FIGURE 6.3 – Fonction de survie des sinistres survenus en 2009 pour la garantie Dommage

En raison d'un volume très important de données, nous ne distinguons pas les intervalles de confiance ponctuels et la bande de confiance à 95 %.

- Etude paramétrique et calcul de la PSAP

Le pourcentage de sinistres survenus en 2009 et encore en cours à la date d'observation est égale à 28,82 % pour la garantie Dommage. Ce taux de censure étant suffisamment élevé pour l'année de survenance 2009, il est légitime d'estimer le montant total de provision pour cette même année en tenant compte des sinistres censurés, soit 544 855 sinistres au total (sinistres clos et encore en cours).

La loi qui maximise la log-vraisemblance est toujours la loi de Weibull. Les paramètres qui permettent d'ajuster au mieux cette loi aux montants des sinistres clos et encore en cours, survenus en 2009, sont donnés dans le tableau suivant :

Alpha (paramètre de forme)	0,72
Lambda (paramètre d'échelle)	669

TABLE 6.13 – Valeurs des paramètres pour la loi de Weibull ajustée

En appliquant la méthode relative au calcul de l'espérance de la provision totale détaillée *supra* (équation (6.1)), nous obtenons un montant de PSAP égale à 153 millions d'euros.

- Comparaison avec *C. L.* et justification de l'écart entre les deux méthodes

Pour mémoire, le montant de PSAP déterminé avec *Chain Ladder* est de 79 millions d'euros, soit un écart très important de 74 millions d'euros.

En considérant les *IBNR* pris en compte avec la méthode de *Chain Ladder*, mais non inclus dans la méthode ligne à ligne, cet écart va augmenter davantage. En effet, d'après l'estimation réalisée *supra*, nous obtenons un coût moyen de 821 euros pour 12 339 sinistres *IBNR*, soit un coût total d'environ 10 millions d'euros.

Le montant total de PSAP avec la méthode ligne à ligne est donc de 163 millions d'euros, soit un écart de 84 millions d'euros avec la méthode de *Chain Ladder*.

6.3.3 Comparaison des deux approches ligne à ligne

Ci-dessous un tableau de synthèse des résultats obtenus avec les deux approches ligne à ligne, *Chain Ladder*, *Mack* (pour l'intervalle de confiance) et les provisions "dossier-dossier" pour les 157 018 dossiers encore en cours en 2009 :

Méthodes utilisées	**Approche 1**	**Approche 2**	***Chain Ladder/ Mack*** ($IC_{95\%}$)	**Provision "dossier"**
PSAP	137 M€	153 M€	79 M€/ [26 M€ ; 132 M€]	147 M€
IBNR	9 M€	10 M€	-	-
Ecarts/Provision "dossier"	-10 M€	+6 M€	-	-

TABLE 6.14 – Synthèse des résultats obtenus pour la garantie Dommage

En ce qui concerne les résultats de provision obtenus avec prise en compte des *IBNR*, l'approche 2 considérant les sinistres censurés est plus éloignée du montant de PSAP obtenu à l'aide de la méthode de *Chain Ladder* que l'approche 1 (écart de 67 millions entre la première approche et *Chain Ladder*, cet écart passe à 84 millions avec l'approche prenant en compte les censures et *Chain Ladder*).

Si l'on compare maintenant les provisions relatives aux deux approches ligne à ligne avec la provision "dossier" caractérisée comme étant notre référentiel pour cette étude, la provision de l'approche 2 surestime de 6 millions la provision "dossier", alors que l'approche 1 la sous-estime de 10 millions. Enfin, même sans connaître le montant d'*IBNR* pour la provision "dossier", on remarque que, par rapport aux deux approches ligne à ligne, la méthode de *Chain Ladder* aboutit à une provision nettement plus éloignée du référentiel.

Compte tenu des résultats à proprement parler, l'approche considérant les censures serait la plus pertinente et la plus prudente. Considérons maintenant la quantité et la qualité d'informations utilisées pour les deux approches ligne à ligne.
Les paramètres estimés sont restitués dans le tableau suivant pour les différentes années de survenance :

Années de survenance	1993	1994	1995	1996
Alpha (paramètre de forme)	0,65	0,66	0,66	0,66
Lambda (paramètre d'échelle)	450	453	470	497

1997	1998	1999	2000	2001	2002	2003
0,67	0,67	0,67	0,67	0,68	0,68	0,70
494	499	496	504	530	551	558

2004	2005	2006	2007	2008	2009
0,70	0,69	0,70	0,70	0,70	0,72
573	589	599	616	663	669

TABLE 6.15 – Valeurs des paramètres pour la loi de Weibull ajustée

Nous remarquons que les paramètres de forme et d'échelle augmentent progressivement au cours du temps. Pour mémoire, lors de l'estimation des paramètres sur l'ensemble des années de survenance (sans considérer les censures), nous avions un paramètre de forme égale à 0,68 et un paramètre d'échelle égale à 528.

Compte tenu de ces observations, l'hypothèse relative à l'approche 1 concernant la stabilité des paramètres au cours du temps n'est pas valide pour la garantie Dommage. De plus, l'évolution de ces paramètres met en évidence que le coût est instable au cours du temps (il est croissant : 615 euros en 1993 et 825 euros en 2009), ce qui explique que les résultats obtenus avec l'approche 1 sont légèrement moins satisfaisants que ceux obtenus avec l'approche 2.

Nous pouvons donc dire que pour estimer le montant de PSAP de la garantie Dommage, pour l'année 2009, il paraît légitime de préférer la seconde approche de provisionnement ligne à ligne permettant la prise en compte des censures, c'est à dire les coûts relatifs aux sinistres encore en cours à la date d'observation. Enfin, nous remarquons que la méthode de *Chain Ladder* paraît inadaptée et conduit, pour cette garantie, à sous-estimer de manière significative le montant de PSAP.

6.4 Conclusion et ouverture

Du fait de la convergence des conclusions associées aux résultats des analyses des lois marginales des coûts relatifs aux trois garanties étudiées, nous pouvons conclure que

l'approche 2 tenant compte des sinistres encore en cours à l'aide d'un modèle de censure est la plus adaptée pour provisionner les types de garanties étudiés pour cette application [15].

En effet, cette méthode de provisionnement ligne à ligne permet de maximiser la quantité d'informations utilisées par rapport à celle mise à disposition dans notre base de données et se caractérise par des résultats relativement proches des montants de provision "dossier" (entre 1,70 % et 3,92 % d'écart par rapport aux montants de provision "dossier").

Comparativement à l'approche 2 (et même aux deux approches de provisionnement ligne à ligne), les résultats associés à *Chain Ladder* sont plus éloignés des provisions "dossier" pour les trois garanties étudiées. Ces résultats conduisent à sous-estimer nettement les montants de PSAP des garanties RC matérielle et Dommage. Cependant, si on regarde de plus près les intervalles de confiance délivrés par le modèle de *Mack*, on s'aperçoit que ces derniers sont assez "larges", en particulier pour la garantie Dommage. Ceci témoigne d'une grande incertitude quant aux estimations réalisées à partir des triangles de paiements, ainsi qu'une volatilité associée relativement importante.

Dans le chapitre 1, page 19, les notions de branche "longue" et de branche "courte" ont été abordées. Elles vont être à nouveau illustrées à travers l'étude non paramétrique réalisée à partir de l'estimateur de *Kaplan-Meier*. Voici, pour les sinistres relatifs aux trois garanties étudiées (RC matérielle, RC corporelle et Dommage) survenus en 2009, l'estimation des fonctions de survie à partir de *Kaplan-Meier* :

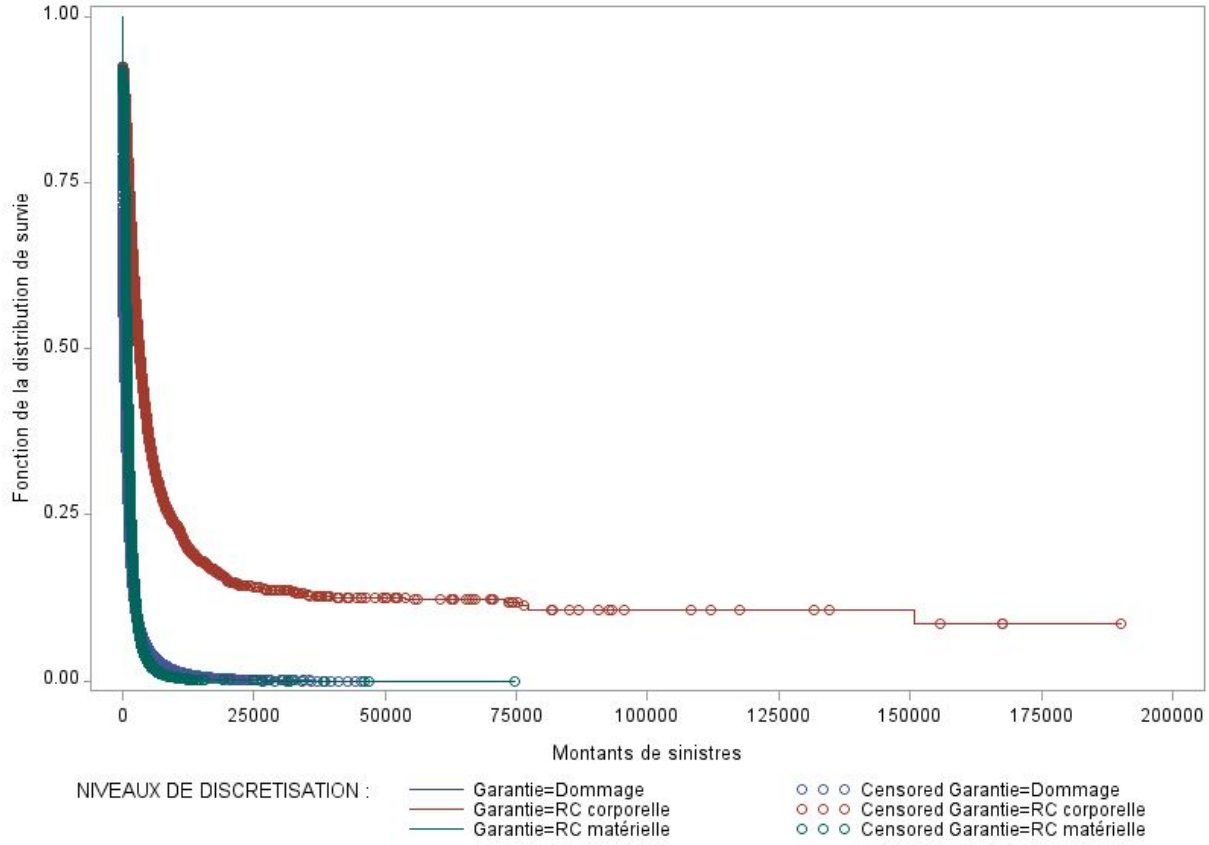


FIGURE 6.4 – Fonctions de survie des sinistres survenus en 2009 pour les trois garanties étudiées (RC matérielle, RC corporelle et Dommage)

15. En supposant que les provisions "dossier", notre référentiel pour cette étude, sont cohérentes.

Graphiquement, on remarque que la fonction de survie relative à la garantie RC corporelle (branche "longue") a une allure différente de celles des branches "courtes" (RC matérielle et Dommage). En effet, sur ce graphique, la probabilité d'obtenir un coût de sinistre supérieur à 25 000 euros est d'environ 0,15 si on considère un sinistre impactant la garantie RC corporelle alors qu'elle est proche de zéro pour un sinistre impactant l'une des deux branches "courtes".

Il est délicat de se prononcer sur l'équivalence des fonctions de survie des deux branches "courtes", des tests d'égalité des survies devraient permettre d'en dire davantage. Les résultats des tests de *logrank* et de *Wilcoxon-Gehan*[16] aboutissent aux mêmes conclusions et conduisent à rejeter très largement ($p - Values < 0,0001$ pour les trois comparaisons de couples de garanties), au seuil de risque de 5 %, l'hypothèse d'égalité des survies. L'équivalence en termes de délais de développement et de vitesse de stabilisation des paiements au cours du temps est donc rejetée.

Il est important de prendre du recul quant aux valeurs de $p - Values$ très faibles de ces tests car le volume de données important conduit certainement à rejeter plus facilement l'hypothèse d'égalité des survies. Cette remarque est particulièrement valable pour la comparaison des survies relatives aux deux branches "courtes".

16. Ces deux tests sont détaillés dans le cours de modèles de survie de Gilbert Colletaz [2012].

Chapitre 7

Analyse de la structure de dépendance entre les coûts relatifs aux garanties RC corporelle, RC matérielle et Dommage

7.1 Sélection et préparation des données à étudier

Pour mémoire, une ligne de notre base de données initiale correspond à l'état des règlements cumulés sur une garantie donnée d'un contrat sinistré. Or, un sinistre peut impacter plusieurs garanties. Pour étudier la dépendance entre les garanties, il a été nécessaire de retravailler la base de données afin d'obtenir pour chaque ligne les montants d'un sinistre associés à chaque garantie. Ce travail a pu être réalisé grâce à la variable relative aux numéros des sinistres.

Afin d'étudier la dépendance entre les branches de risque RC corporelle, RC matérielle et Dommage, nous avons créé une base regroupant pour chaque ligne, l'état des règlements cumulés d'un sinistre avec la ventilation pour chaque branche de risque.

Le modèle proposé au chapitre 3, page 26, impose de ne sélectionner que les sinistres pour lesquels les montants relatifs à la garantie Dommage sont supérieurs à zéro. Il y avait encore des montants égaux à zéro pour la garantie RC corporelle et RC matérielle et des données manquantes pour ces deux garanties (garanties non impactées par le sinistre). Les données manquantes ont été assimilées par des montants égaux à zéro (car les garanties RC sont des garanties obligatoires et donc l'assuré est forcément couvert). Ceci permet d'obtenir une base de 4 333 841 lignes, en ayant sélectionné initialement les sinistres clos uniquement.

Afin d'illustrer la spécificité des données, les dépendogrammes des branches de risque deux à deux sont représentés ci-dessous en considérant un échantillon de 10 000 données sélectionnées de manière aléatoire. Le caractère aléatoire est justifié par le fait que les 10 000 données représentent toutes les années de survenance et pas seulement les années de survenance les plus récentes ou les plus anciennes. Il s'agit donc de procéder par échantillonnage en vérifiant que la corrélation de l'échantillon reflète bien la corrélation de l'ensemble des données.

Ce travail de vérification a été effectué (représentation des dépendogrammes, coefficients de corrélation et tests associés, choix de la meilleure copule), ce qui permet d'extrapoler les résultats obtenus à partir de l'échantillon de 10 000 données. L'intérêt de travailler sur un échantillon (à condition qu'il soit représentatif) est ici de pouvoir obtenir certains graphiques

plus rapidement (par exemple, il n'a pas été possible d'obtenir de K-plot sur l'ensemble des données). Il s'agit donc de minimiser les temps de calcul et d'obtenir des dépendogrammes plus lisibles. En effet, représenter plus de 10 000 points sur un dépendogramme devient illisible et s'avère coûteux en temps de traitement.

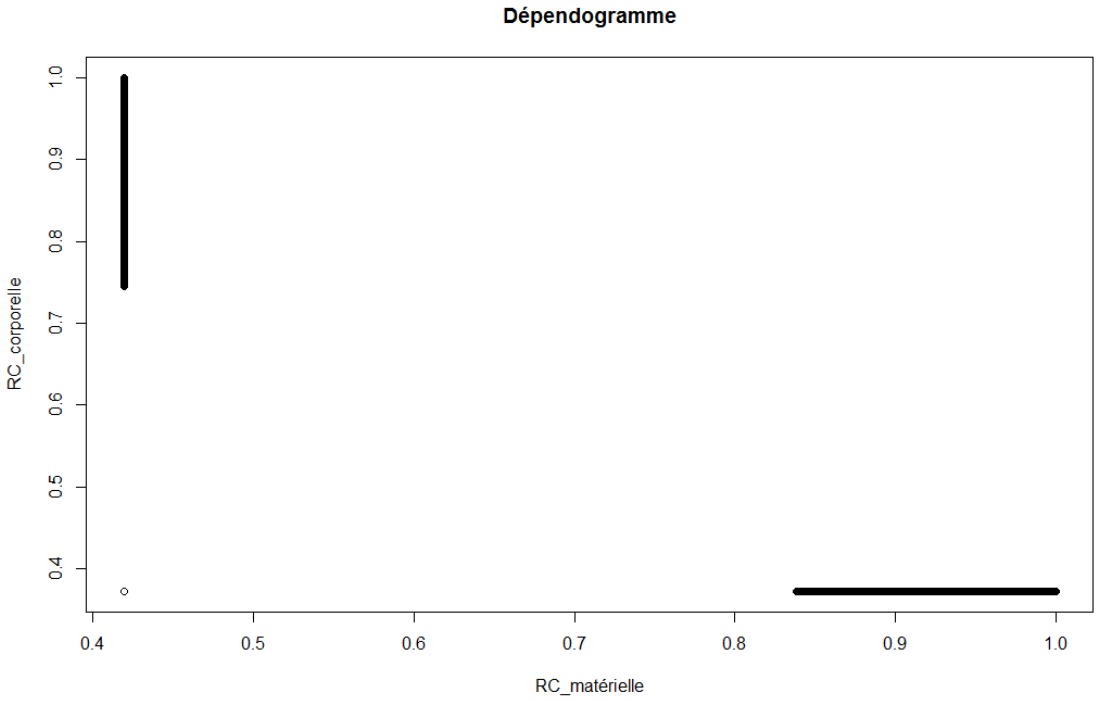


FIGURE 7.1 – Dépendogramme du couple RC corporelle / RC matérielle

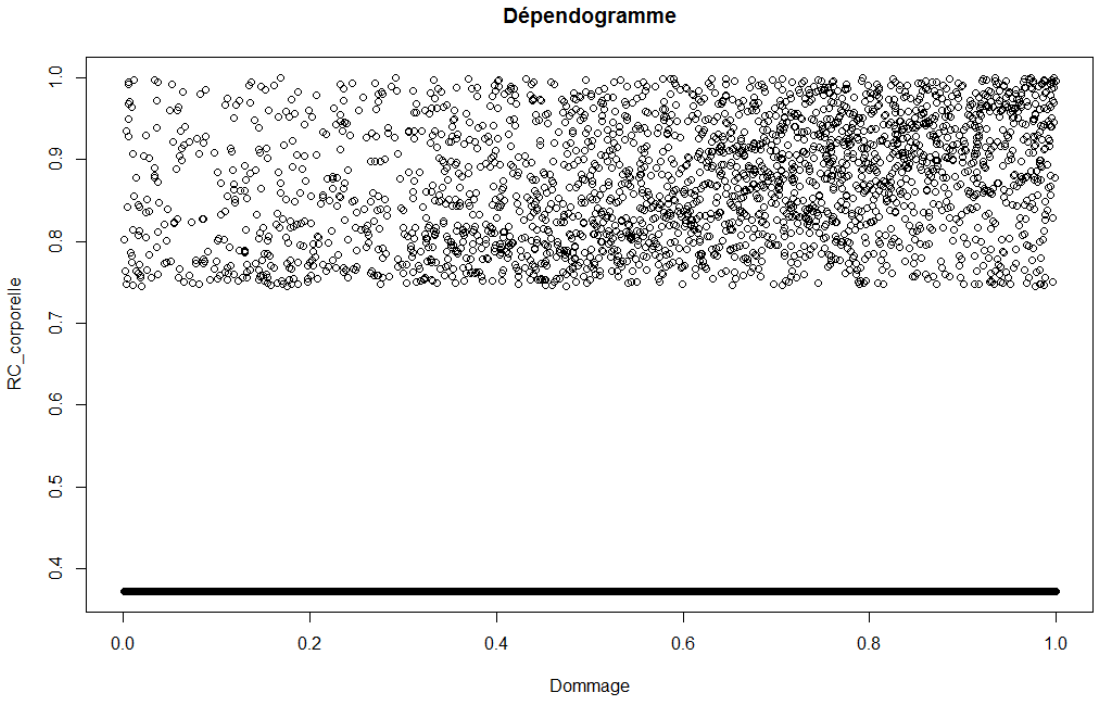


FIGURE 7.2 – Dépendogramme du couple RC corporelle / Dommage

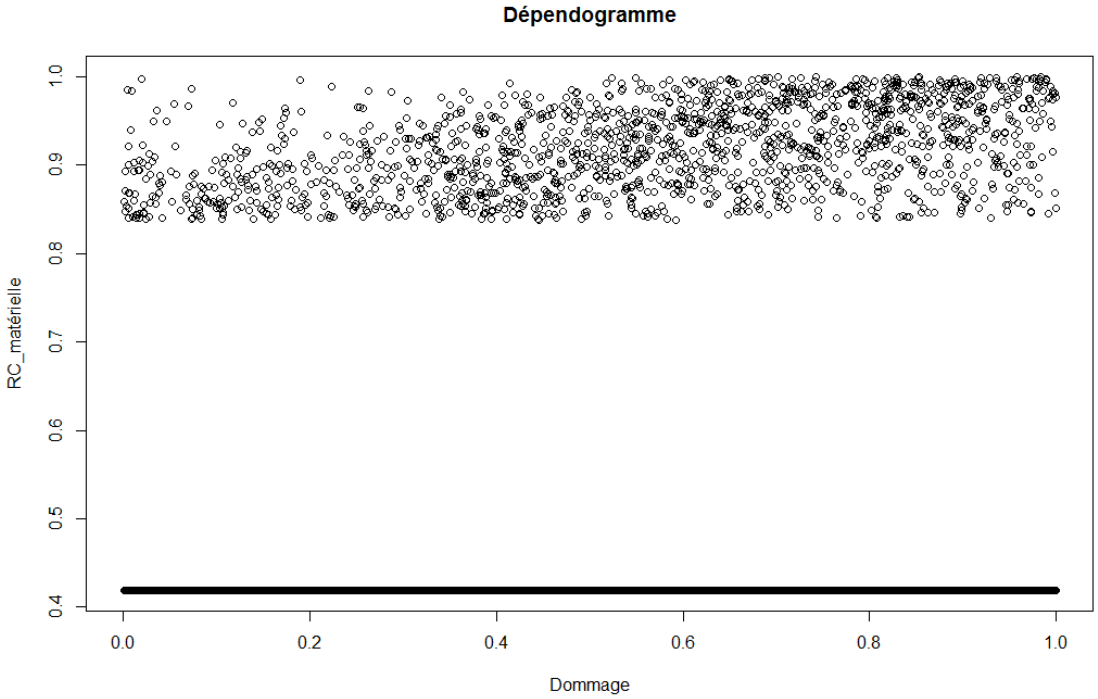


FIGURE 7.3 – Dépendogramme du couple RC matérielle / Dommage

Les dépendogrammes ci-dessus font ressortir les masses en zéro (traits noirs) présentes dans les coûts relatifs aux branches RC corporelle et RC matérielle. Sur les deux derniers dépendogrammes, l'espace sans point entre la masse en zéro et le nuage de points est la conséquence directe de la construction des rangs à l'aide de la fonction *rank*() du logiciel R qui associe le même rang à un groupe d'ex-aequo [1]. Les coûts égaux à zéro se voient attribuer le même rang, ce qui provoque le trait noir et l'espace vide entre ce trait et le nuage de points.

Nous aurions pu "randomiser" pour avoir des rangs différents pour nos ex-aequo. Cela aurait permis d'éliminer l'espace vide entre la masse en zéro et le nuage de points. Toutefois, cette méthode serait dénuée de sens car nous avons bien la valeur zéro associée à ces coûts et non des montants proches de zéro que l'on aurait au préalable arrondi. Ainsi, il n'est pas justifié d'affecter une position différente à nos ex-aequo. En revanche, si nous avions été en présence d'un calcul de jours avec des arrondis effectués préalablement, l'intérêt de "randomiser" aurait été justifié pour obtenir des rangs différents pour nos ex-aequo.

Pour la suite de l'étude, la masse de Dirac en zéro va être écartée afin de réaliser l'étude sur la partie continue uniquement. En effet, les graphiques présentés lors de l'application ne sont pas adaptés pour une masse de Dirac. Aussi, il convient de réaliser des ajustements à partir de copules continues pour ne les appliquer que sur la partie continue. Enfin, aucune copule n'est à ce jour connue pour être capable de caractériser la dépendance que reflètent les dépendogrammes ci-dessus.

Considérer seulement les montants de sinistres strictement supérieur à zéro pour les trois branches de risque réduit considérablement le volume de données :

- 61 sinistres seulement ont impacté les trois garanties simultanément
- 100 sinistres ont impacté le couple RC corporelle / RC matérielle
- 203 105 sinistres ont impacté le couple RC corporelle / Dommage

1. La fonction *rank*() est détaillée à la section 4.6.1, page 53.

- 1 058 337 sinistres ont impacté le couple RC matérielle / Dommage

Compte tenu du volume insuffisant de sinistres qui ont impacté les trois garanties simultanément, il a été décidé de se concentrer sur l'étude de la dépendance des branches de risque deux à deux. De la même manière que précédemment, il a été choisi de procéder par échantillonnage pour les couples RC corporelle/Dommage et RC matérielle/Dommage en retenant 10 000 données sélectionnées de façon à représenter au mieux nos données initiales (nombre de données identique pour chaque année de survenance, etc.).

7.2 Recherche de la dépendance

Cette section est destinée à mettre en évidence la dépendance qui réside entre nos trois branches de risques d'un point de vue statistique. Seules les dépendances par couple de branches seront analysées du fait d'un volume de données trop faible pour le triplet relatif aux trois branches de risques (61 données seulement).

Le coefficient de corrélation linéaire de Pearson, le Tau de kendall et le Rho de Spearman sont trois indicateurs qui devraient permettre de donner une idée de la dépendance qui existe entre nos trois variables. Ensuite, des tests statistiques d'indépendance permettront de confirmer ou d'infirmer la présence de dépendance.

En ce qui concerne la représentation graphique de la dépendance, un Rank-Rank Plot (ou dépendogramme) et un diagramme de dispersion devraient représenter graphiquement les grandes structures de dépendance. Enfin, un Khi-Plot, ainsi qu'un K-Plot, vont permettre de préciser et de valider les structures de dépendance préalablement trouvées à l'aide des outils précédemment énoncés.

7.2.1 Mesures et tests statistiques d'indépendance

- **Matrice de corrélation de Pearson**

	RC corporelle	RC matérielle	Dommage
RC corporelle	1	0,688	0,076
RC matérielle		1	0,341
Dommage			1

Table 7.1 – Corrélation linéaire de Pearson

- **Matrice des coefficients de Spearman**

	RC corporelle	RC matérielle	Dommage
RC corporelle	1	0,208	0,318
RC matérielle		1	0,402
Dommage			1

Table 7.2 – Rho de Spearman

- **Matrice des coefficients de Kendall**

	RC corporelle	RC matérielle	Dommage
RC corporelle	1	0,140	0,221
RC matérielle		1	0,277
Dommage			1

TABLE 7.3 – Tau de Kendall

Il faut souligner une valeur élevée pour le coefficient de corrélation linéaire de Pearson (0,688) entre les branches de risque Responsabilité Civile corporelle et matérielle. *A contrario*, il ne semble pas se dégager de corrélation linéaire entre les branches Dommage et RC corporelle.

En cohérence avec la forte valeur du coefficient de corrélation linéaire de Pearson entre les branches RC corporelle et RC matérielle, nous remarquons des valeurs nettement plus faibles pour le Rho de Spearman et le Tau de Kendall (0,208 et 0,140 respectivement).

En ce qui concerne les liaisons entre les branches Dommage et RC corporelle ainsi que Dommage et RC matérielle, on note une valeur assez élevée pour le Rho de Spearman (0,318 et 0,402 respectivement) et un peu plus faible pour le Tau de Kendall (0,221 et 0,277 respectivement).

Afin de compléter cette analyse, il est nécessaire de se référer à des tests statistiques d'indépendance [2]. Ces derniers devraient permettre de savoir si les coefficients ci-dessus sont significativement différents de zéro ou non.

- **Tests statistiques d'indépendance**

Les p-values des tests associés à un risque d'erreur de 5 % sont renseignées dans les tableaux suivants :

	RC corporelle	RC matérielle	Dommage
RC corporelle		2,665e-15	2,665e-14
RC matérielle			< 2,2e-16
Dommage			

TABLE 7.4 – P-Value du test de corrélation de Pearson

	RC corporelle	RC matérielle	Dommage
RC corporelle		0,038	< 2,2e-16
RC matérielle			< 2,2e-16
Dommage			

TABLE 7.5 – P-Value du test de "corrélation" de Spearman

2. Ces tests sont présentés à la section 4.5.2, page 52, et sont réalisés avec la fonction *cor.test* sous le logiciel R.

	RC corporelle	RC matérielle	Dommage
RC corporelle		0,038	0
RC matérielle			0
Dommage			

TABLE 7.6 – P-Value du test de "corrélation" de Kendall

Pour l'ensemble de ces tests, l'hypothèse d'indépendance H_0 est rejetée (P-Value inférieure au seuil de 5 %). Nous en concluons une dépendance entre nos trois couples étudiés.

Cependant, concernant le couple RC matérielle et RC corporelle, on rejette moins nettement l'hypothèse d'indépendance H_0 pour les tests de "corrélation" de Spearman et de Kendall. Ainsi, il faudra être vigilant quant aux résultats associés à ce couple. Ceci peut également venir du fait que nous travaillons sur un petit échantillon (100 données seulement).

7.2.2 Analyse graphique de la dépendance

L'étude statistique a apporté de l'information uniquement sur l'intensité de la dépendance, positive ou négative, des différentes corrélations. Il s'agit à présent d'effectuer une analyse graphique afin de caractériser davantage la dépendance entre ces 3 branches de risque.

- **Les diagrammes de dispersion**

Le diagramme de dispersion représente graphiquement les couples (X_i, Y_i) de deux variables quantitatives X et Y. Il permet de se rendre compte rapidement des dépendances entre les variables aléatoires étudiées et de déterminer la nature des dépendances. Il s'agit d'un outil pour une première approche car dans certaines situations, il est difficile de se rendre compte des structures de corrélation.

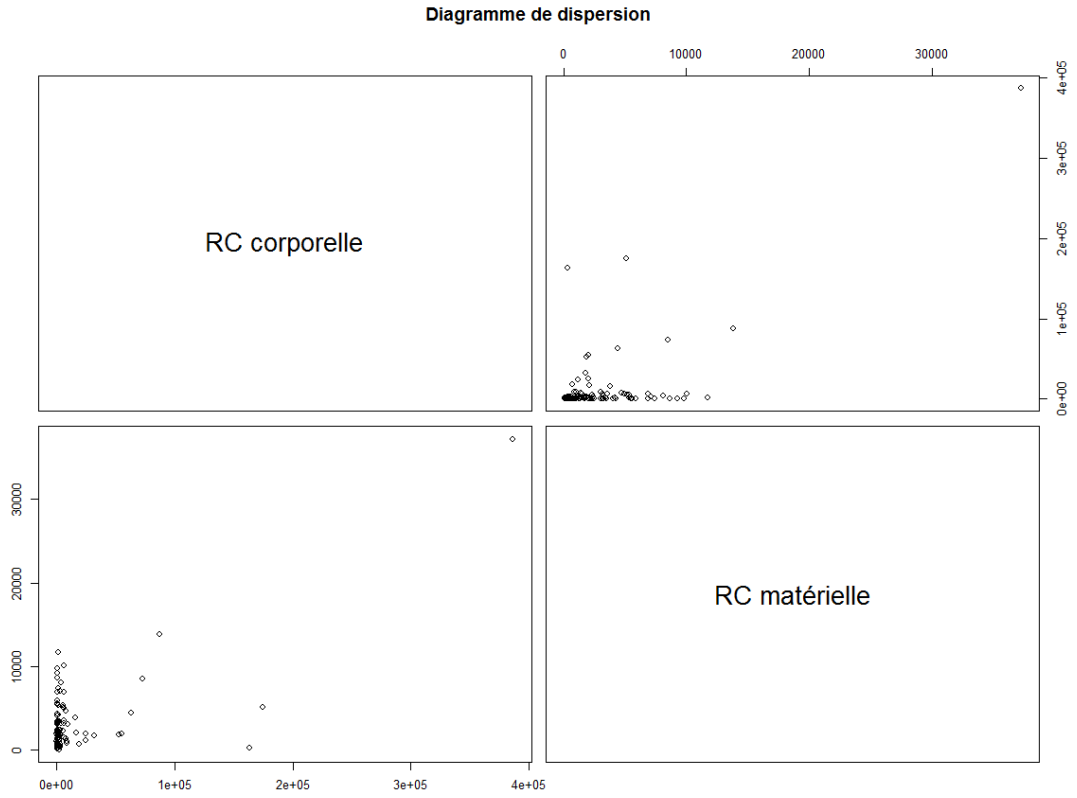


FIGURE 7.4 – Diagramme de dispersion pour le couple RC corporelle/RC matérielle

Bien que le nombre d'observations soit restreint (une centaine) pour ce couple de branches de risque étudié, une structure de dépendance se démarque. En effet, d'après ce diagramme de dispersion, une corrélation linéaire est observée, marquée par un alignement des points selon une droite.

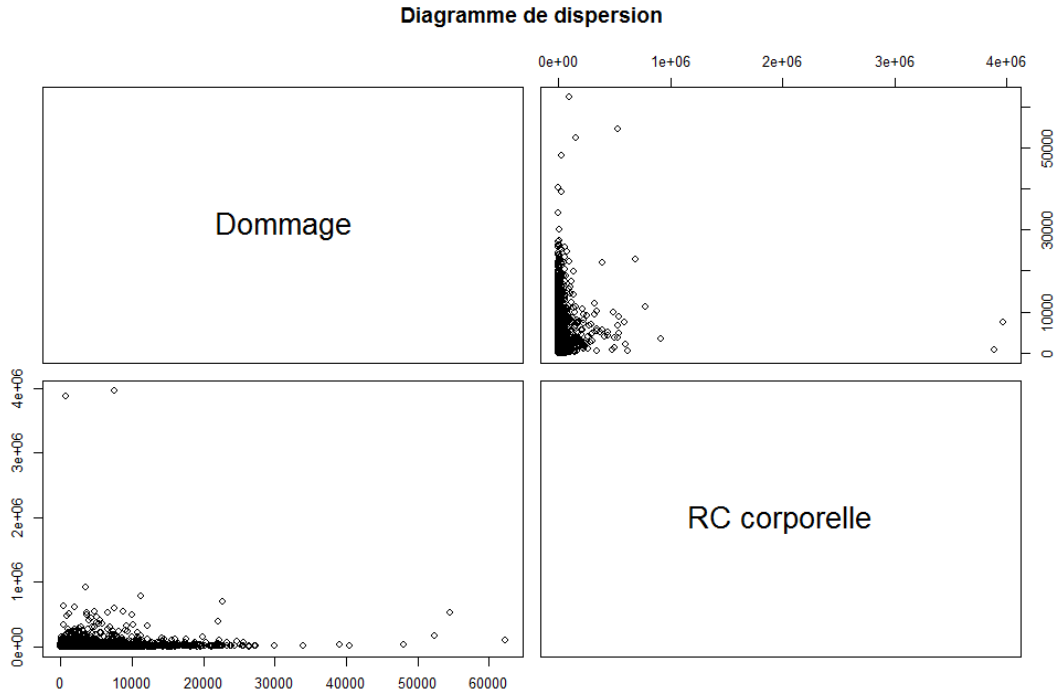


FIGURE 7.5 – Diagramme de dispersion pour le couple Dommage/RC corporelle

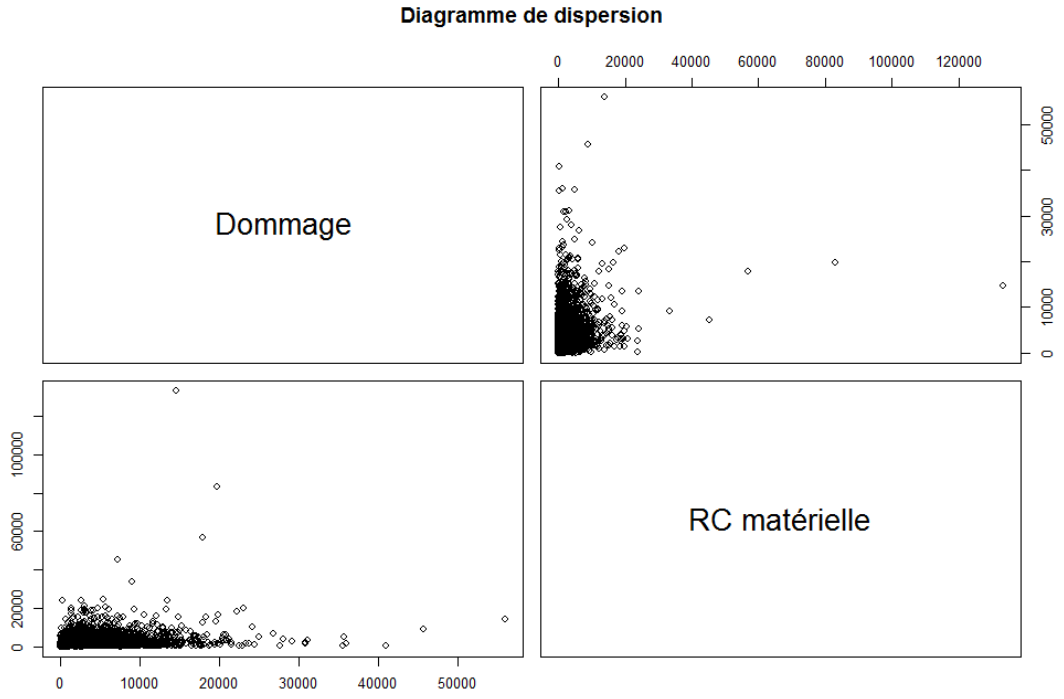


FIGURE 7.6 – Diagramme de dispersion pour le couple Dommage/RC matérielle

Les deux diagrammes ci-dessus montrent une concentration des points assez forte, ce qui laisse penser que ces deux derniers couples dégagent de la dépendance. Cependant, l'analyse des diagrammes de dispersion ne permet pas d'extraire davantage d'information. L'analyse des dépendogrammes devrait le permettre.

- **Les dépendogrammes**

Les dépendogrammes apportent de l'information sur le type de dépendance que la copule devra modéliser. Ils permettent également d'évaluer les réalisations simultanées d'importantes ou de faibles charges de sinistres entre deux branches de risque données.

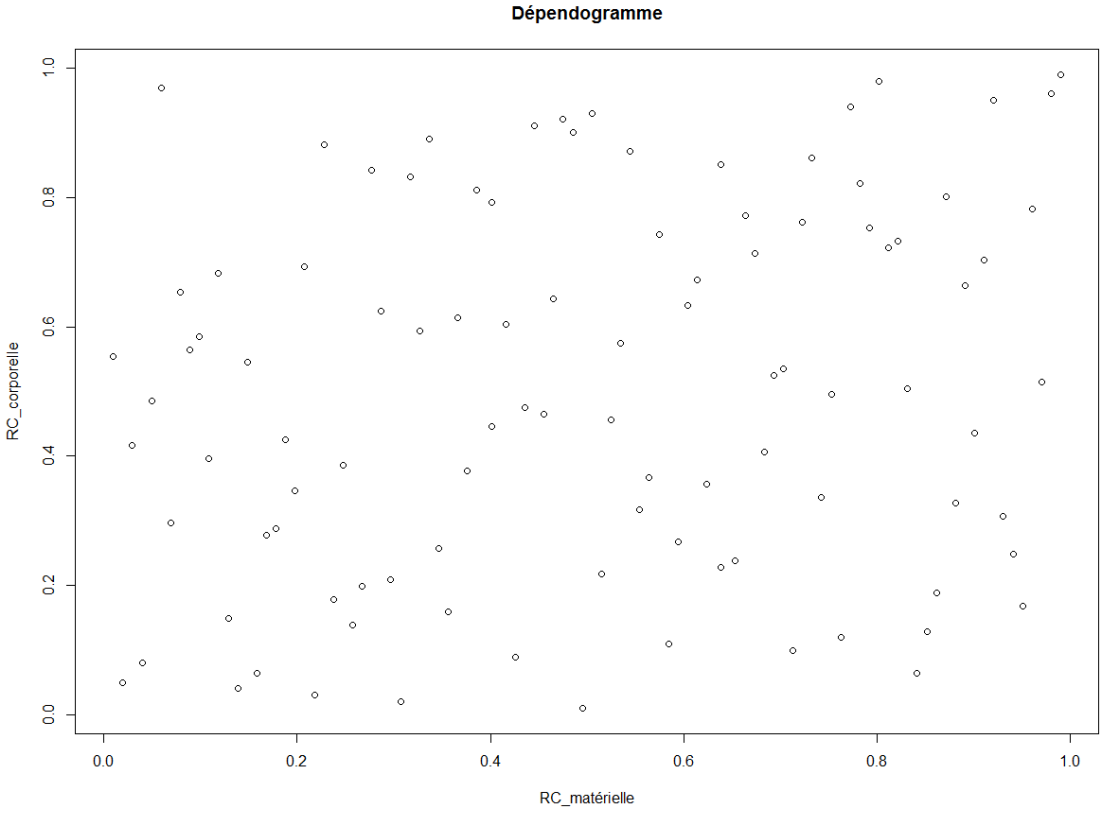


FIGURE 7.7 – Dépendogramme du couple RC corporelle/RC matérielle

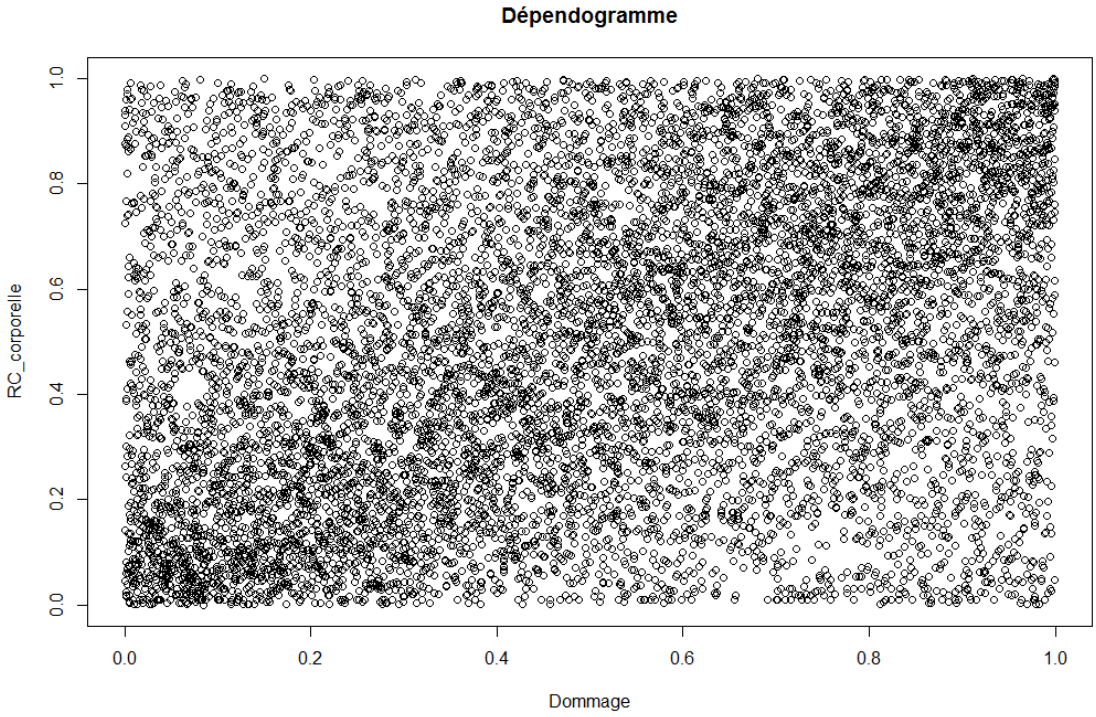


FIGURE 7.8 – Dépendogramme du couple RC corporelle/Dommage

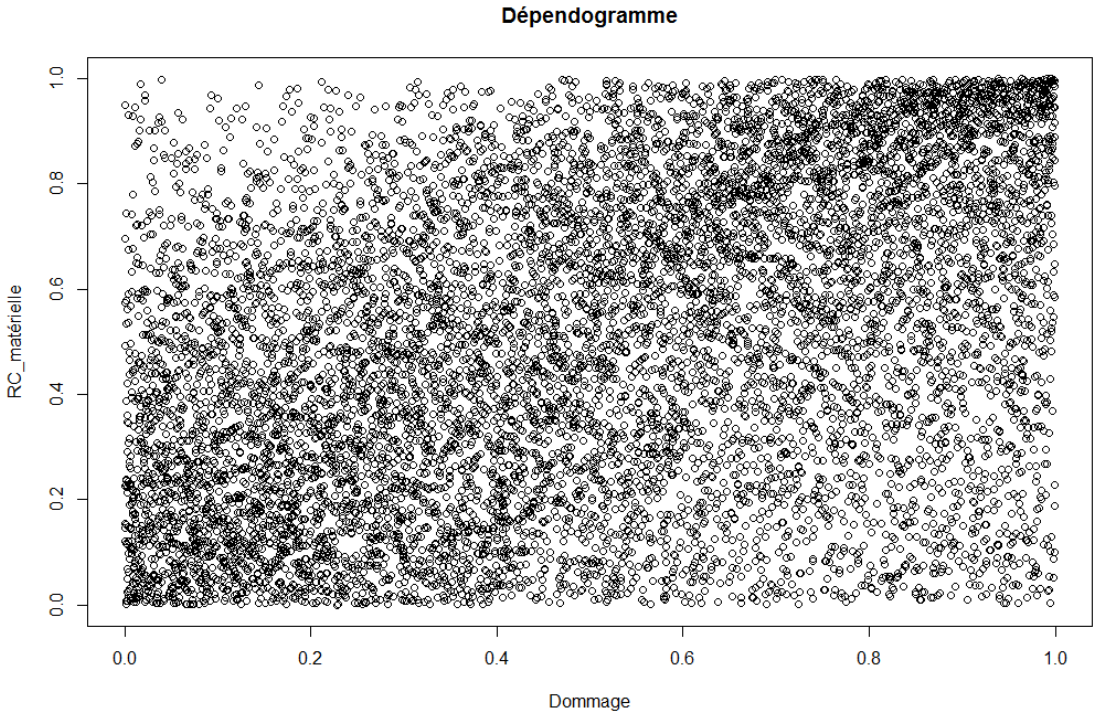


FIGURE 7.9 – Dépendogramme du couple RC matérielle/Dommage

Pour le couple RC matérielle/RC corporelle, le dépendogramme traduit une légère dépendance linéaire (légère concentration des points sur la diagonale ascendante et absence de point sur les coins supérieurs à gauche et inférieurs à droite du diagramme). Néanmoins, cela ne permet pas de tirer de conclusion quant à l'éventuelle structure de dépendance entre

ces deux garanties RC. De plus, ce graphique ne montre pas de dépendance significative. Ceci est peut-être la conséquence d'un volume de données trop faible (une centaine d'observations seulement).

Pour les deux autres couples étudiés, on observe seulement des dépendances positives caractérisées par un alignement des points sur la diagonale ascendante. De la dépendance sur les queues de distribution est également observées (concentration un peu plus marquée en bas à gauche pour le couple RC corporelle/Dommage et en haut à droite pour le couple RC matérielle/Dommage). Il y a donc une réalisation simultanée de faibles charges de sinistres pour les branches RC corporelle et Dommage et une réalisation simultanée de fortes charges de sinistres pour les branches RC matérielle et Dommage. Il s'agit donc d'une dépendance de queue plus ou moins asymétrique.

- **Le Khi-Plot**

Les graphiques ont été obtenus à l'aide du logiciel R. Une borne supérieure et une borne inférieure de l'intervalle de confiance à 95 % relatif au test d'indépendance ont été rajoutées aux graphiques.

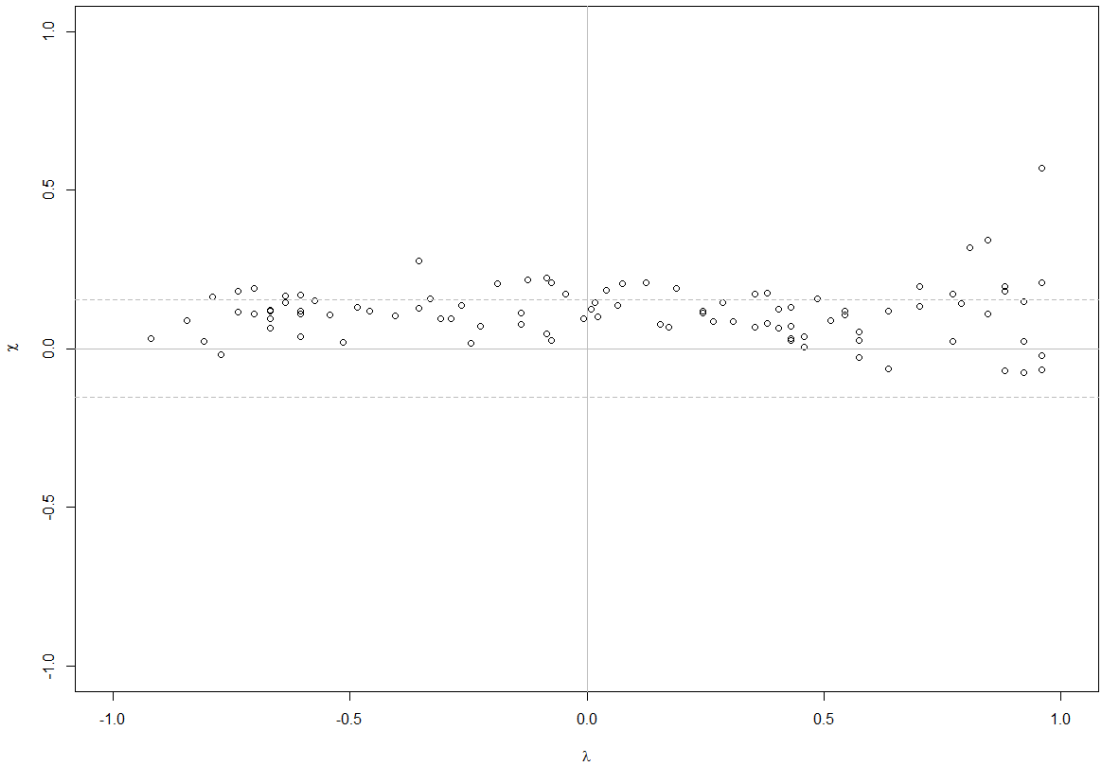


FIGURE 7.10 – Khi-Plot du couple RC corporelle/RC matérielle

Pour le Khi-Plot relatif au couple RC corporelle/RC matérielle, la majorité des points sont situés dans l'intervalle de confiance à 95 % relatif au test d'indépendance. Il s'agit des points situés à l'intérieur de la "bande d'indépendance" centrée sur l'axe des abscisses. Les autres points, en nombre minoritaire, situés en dehors de cette bande sont donc les points susceptibles d'être corrélés.

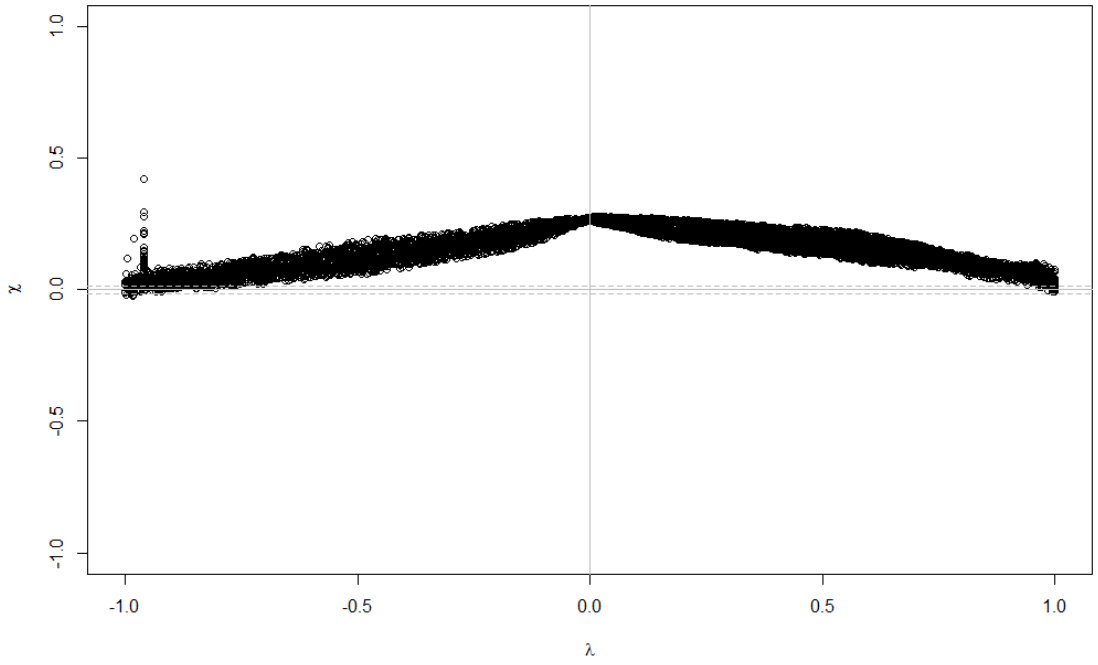


FIGURE 7.11 – Khi-Plot du couple RC corporelle/Dommage

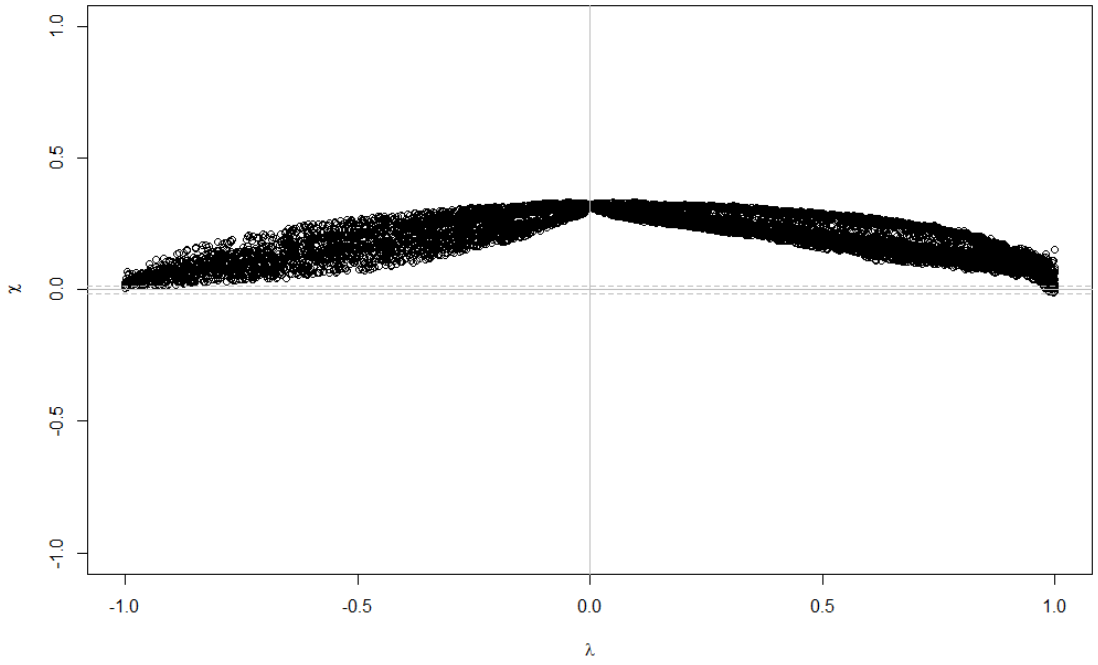


FIGURE 7.12 – Khi-Plot du couple RC matérielle/Dommage

Pour les couples RC corporelle/Dommage et RC matérielle/Dommage, la majeure partie des points est située largement en dehors de la "bande d'indépendance". Il est raisonnable de penser que ces deux couples présentent une dépendance assez forte, ce qui confirme les analyses précédentes.

L'inconvénient du Khi-Plot est qu'il ne permet pas de caractériser davantage la structure de dépendance. Le Kendall-Plot devrait permettre d'en dire davantage.

- **Le K-Plot**

L'avantage des Kendall-Plot est de pouvoir décrire la structure de dépendance de façon plus précise que le Khi-Plot [3].

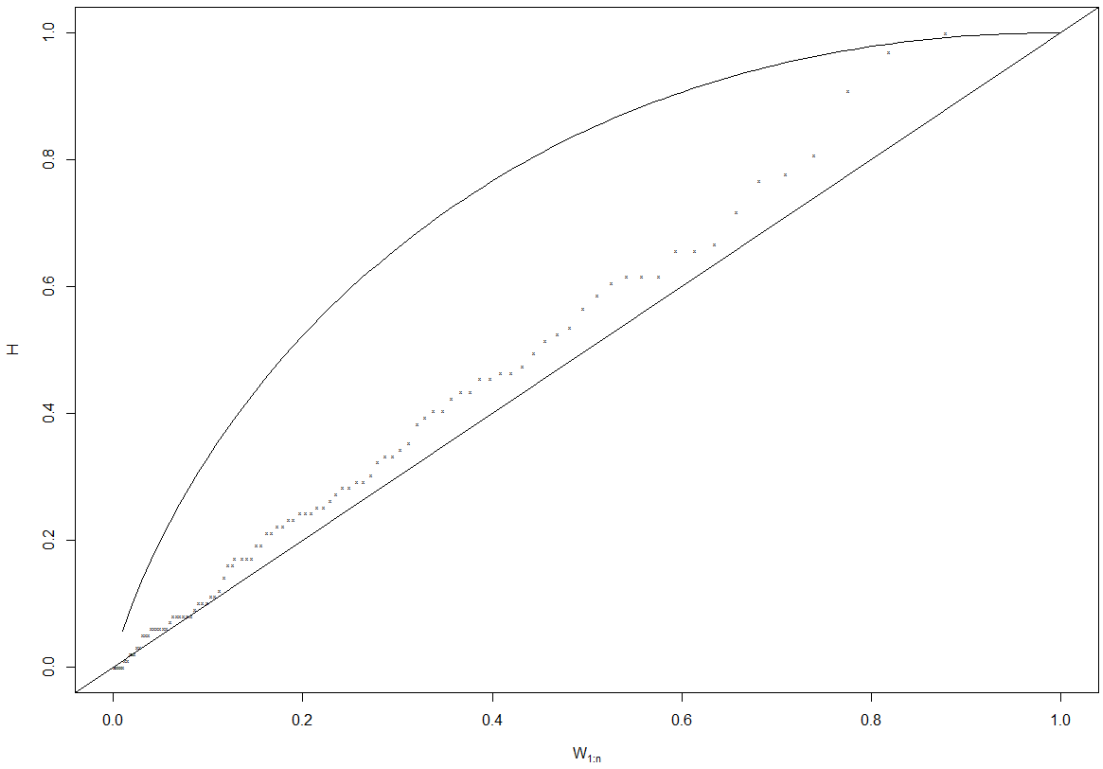


FIGURE 7.13 – K-Plot du couple RC corporelle / RC matérielle

Le graphique ci-dessus permet de confirmer une dépendance positive relativement faible entre les branches de risque RC corporelle et RC matérielle.

3. La section 4.5.1, page 49, détaille les avantages et inconvénients de ces deux graphiques ainsi que la façon de les interpréter.

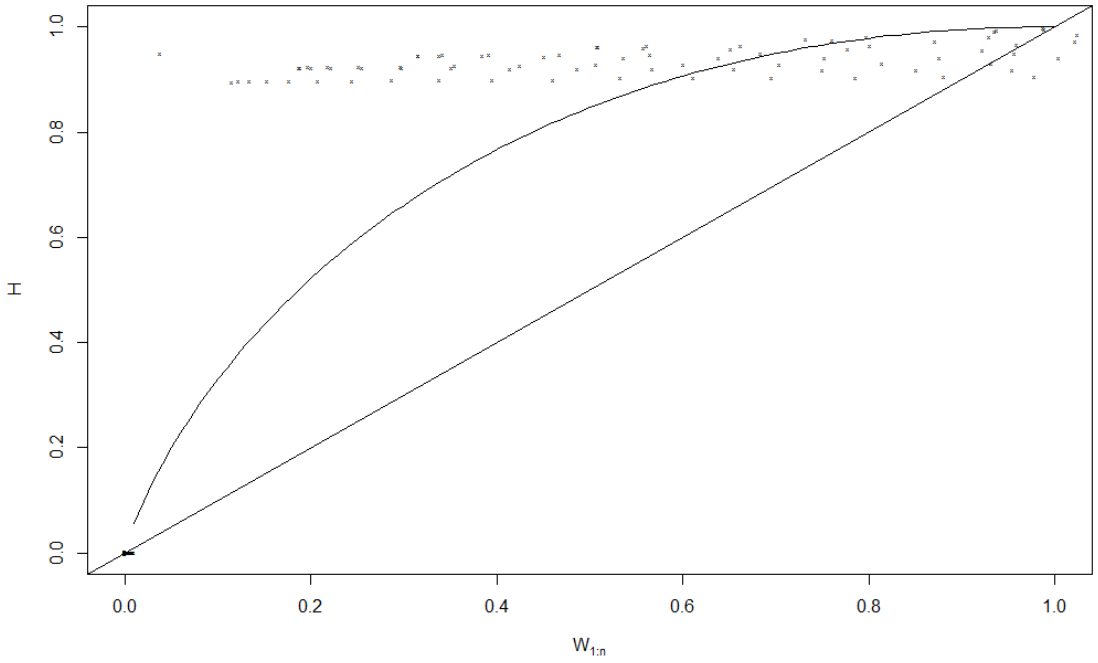


FIGURE 7.14 – K-Plot du couple RC corporelle / Dommage

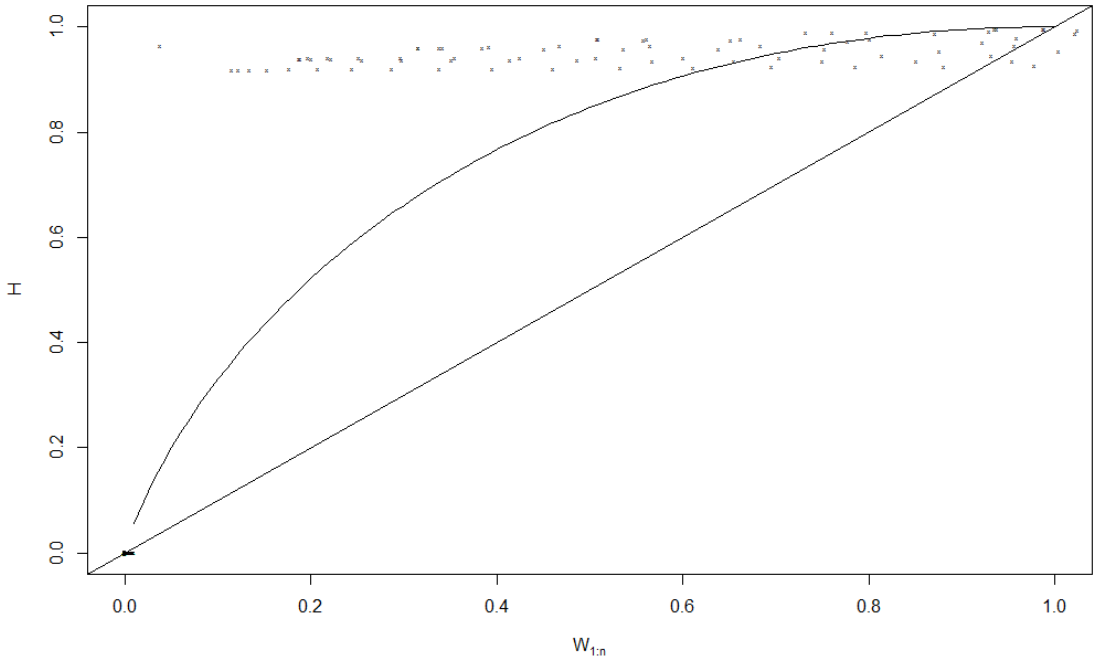


FIGURE 7.15 – K-Plot du couple RC matérielle / Dommage

D'après les deux graphiques ci-dessus, les couples RC corporelle/Dommage et RC matérielle/Dommage sont caractérisés par de la dépendance positive relativement élevée.

7.2.3 Conclusions sur les dépendances observées entre les trois branches de risque étudiées

Les mesures de dépendance, les tests d'indépendance et les différentes représentations graphiques confirment l'existence de dépendances entre les coûts de sinistres relatifs aux trois branches de risque étudiées.

Plus particulièrement, il existe potentiellement une dépendance assez faible pour le couple RC corporelle/RC matérielle. Ce couple est caractérisé par une centaine d'observations seulement, ce qui peut rendre l'analyse délicate. En effet, il est délicat de savoir si le manque d'observation n'est pas la cause d'une dépendance assez faible. Dans tous les cas, compte tenu de la situation, il sera délicat de modéliser la dépendance entre ces deux garanties.

En ce qui concerne les deux autres couples, ils sont significativement marqués par de la dépendance positive. Selon les dépendogrammes, ils présentent une concentration sur les queues de distribution mais de façon légèrement asymétrique. En effet, le couple RC corporelle/Dommage est caractérisé par une concentration des points plus marquée sur la queue inférieure que sur la queue supérieure, sachant que les deux queues présentent une concentration de points significative. La situation inverse est observée pour le couple RC matérielle/Dommage.

Le type de copule qui modélisera le mieux la structure de dépendance et que nous retiendrons, va permettre de savoir si cette asymétrie est significative, suffisamment marquée et si elle nécessite ou non l'utilisation d'une copule permettant de la modéliser.

Il serait ainsi possible d'écarter la piste des copules elliptiques lors de la recherche de la meilleure copule pour modéliser la structure de dépendance de ces deux couples puisque les copules elliptiques modélisent les dépendances symétriques. Malgré tout, nous ne les écarterons pas lors de la recherche afin de pouvoir confirmer notre intuition. De plus, l'asymétrie constatée sur les dépendogrammes n'est pas particulièrement marquée et significative.

7.3 Modélisation de la dépendance et estimation de la copule

7.3.1 Représentation de la copule

- **La densité empirique**

La densité empirique[4] est similaire à un histogramme bivarié de nos observations. En considérant les couples de branches de risque (X_i, Y_i), nous nous plaçons sur un grillage c x c de telle manière que pour le premier axe, une unité soit de longueur $\frac{max(X)-min(X)}{c}$ et pour le second, $\frac{max(Y)-min(Y)}{c}$. L'axe vertical représente la fréquence qu'un point se retrouve dans le carré unitaire correspondant.

A noter que choisir un c trop petit ou trop grand entraîne une perte d'information sur l'histogramme. Il est donc important de prendre un c adéquat pour pouvoir interpréter l'histogramme dans de bonnes conditions. Ici, une valeur égale à 8 pour c a été retenue.

Nous obtenons ainsi un histogramme en 3 dimensions, ce qui permet de visualiser la densité de la copule recherchée.

4. Cf. section 4.8.2, page 59.

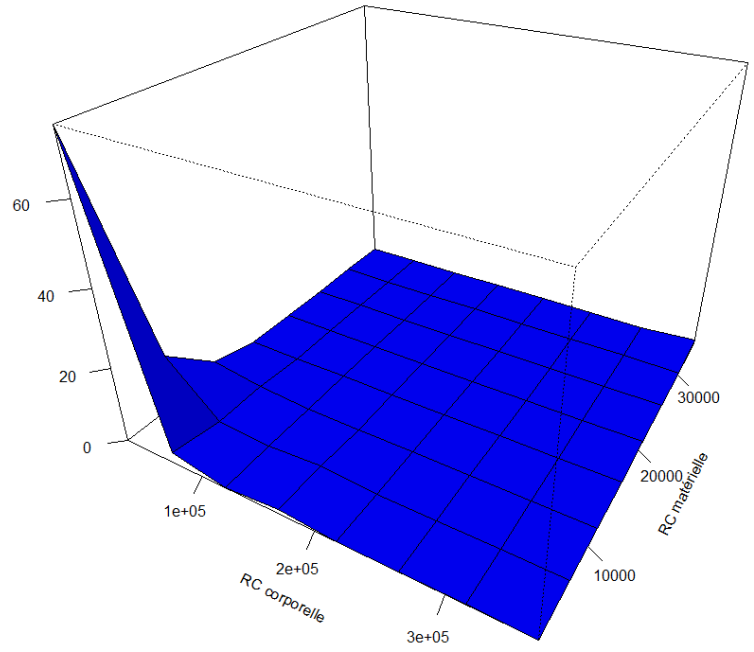


FIGURE 7.16 – Histogramme 3D de la RC corporelle et de la RC matérielle

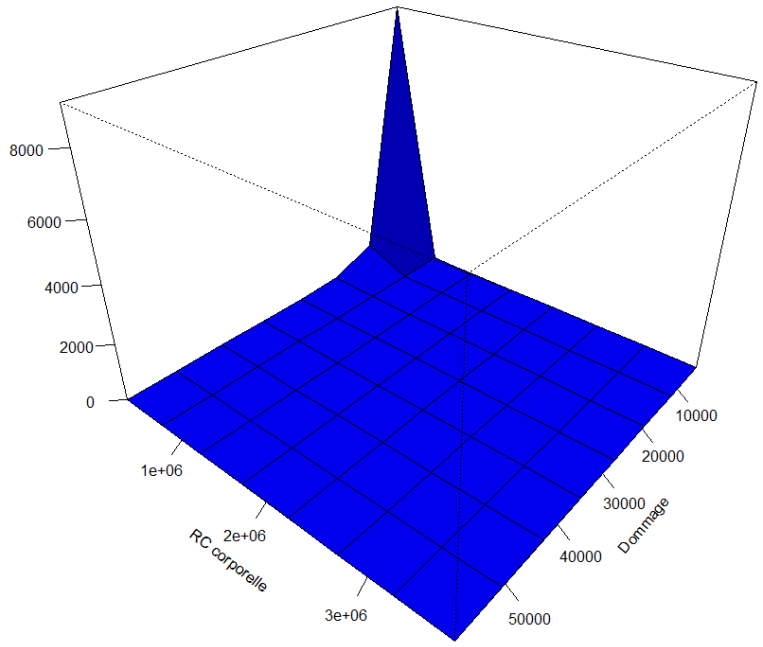


FIGURE 7.17 – Histogramme 3D de la RC corporelle et du Dommage

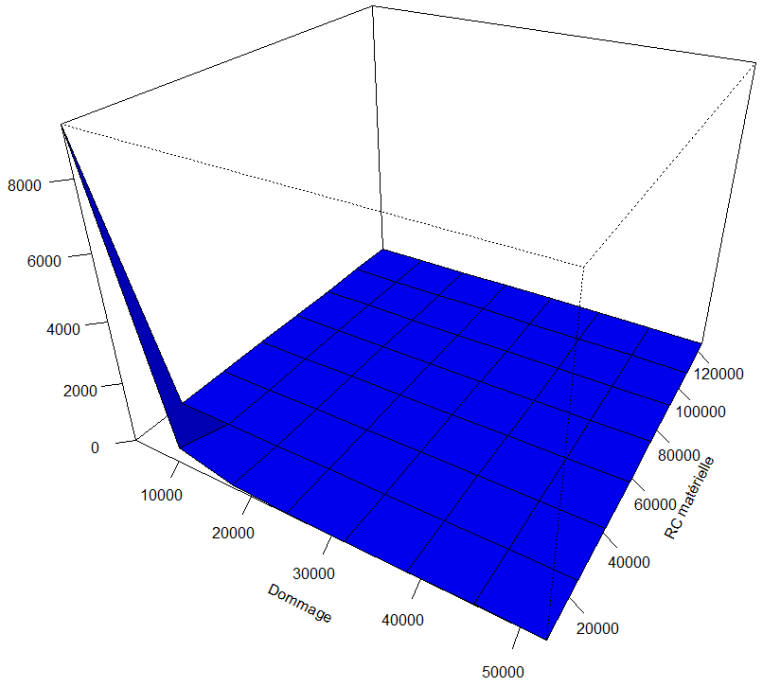


FIGURE 7.18 – Histogramme 3D de la RC matérielle et du Dommage

Ces trois graphiques sont en accord avec l'observation effectuée sur les diagrammes de dispersion, à savoir que les couples d'observations sont concentrés majoritairement pour de faibles valeurs. Autrement dit, les couples d'observations sont plus concentrés vers le minimum que vers le maximum.

- **La copule empirique**

Comme énoncé dans la section 4.8.1, page 59, une estimation non-paramétrique des copules remonte à Deheuvels [1979] [31] qui propose d'estimer les copules par la fonction de répartition empirique :

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\frac{R_i}{n}\leq u, \frac{S_i}{n}\leq v\}}, u, v \in [0,1]$$

Cette loi de probabilité place une masse $\frac{1}{n}$ en chaque paire de rangs normalisés R_i et S_i. Par ailleurs, notons que C_n est la fonction de répartition des couples $(\frac{R_i}{n}, \frac{S_i}{n})$.

Par ailleurs, Deheuvels [1981a, 1981b] a étudié la cohérence de C_n et a obtenu la loi exacte de $\sqrt{n}(C_n - C)(u,v)$ lorsque les deux marginales sont supposées indépendantes.

Après implémentation de la copule empirique sur le logiciel R à l'aide de la fonction *gplots* [5], nous obtenons les graphiques suivants pour les couples de branches de risque étudiés :

5. Les détails concernant l'utilisation de cette fonction sont accessibles à partir de ce lien bibliographique [39].

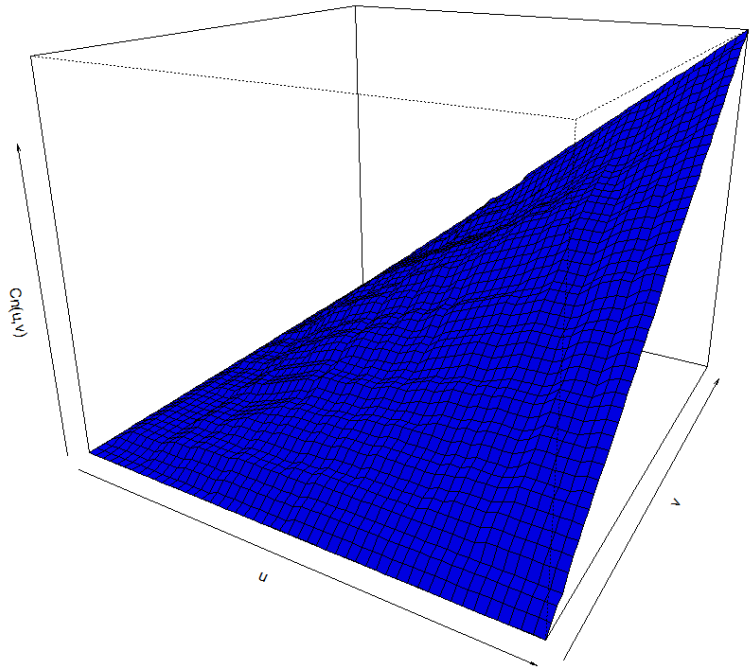


FIGURE 7.19 – Copule empirique de la RC corporelle et de la RC matérielle

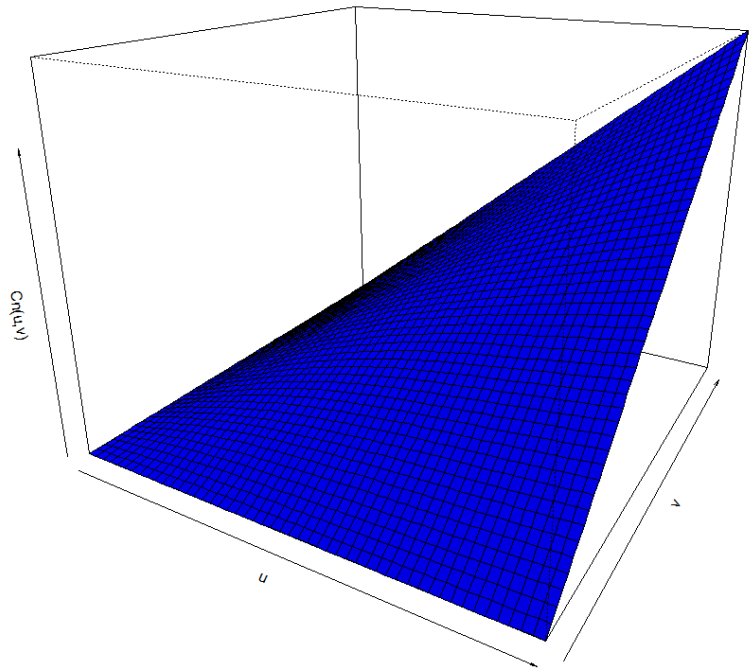


FIGURE 7.20 – Copule empirique de la RC corporelle et du Dommage

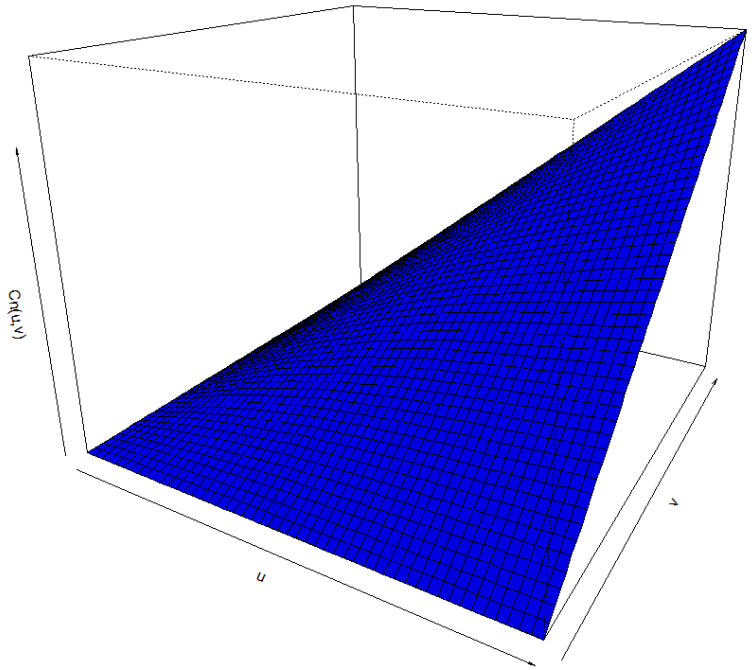


FIGURE 7.21 – Copule empirique de la RC matérielle et du Dommage

On remarque que les trois graphiques ont la même allure. La surface du premier graphique est moins lisse que celles des deux suivants du fait d'un nombre d'observations moindre.

7.3.2 Ajustement des lois marginales

L'objectif de cette section est de rapprocher les lois marginales de nos copules avec des lois classiques pour trouver la loi la plus adéquate et ainsi en estimer le paramètre à l'aide de la méthode du maximum de vraisemblance.

Nous avons tout d'abord estimé les densités de chaque branche de risque.

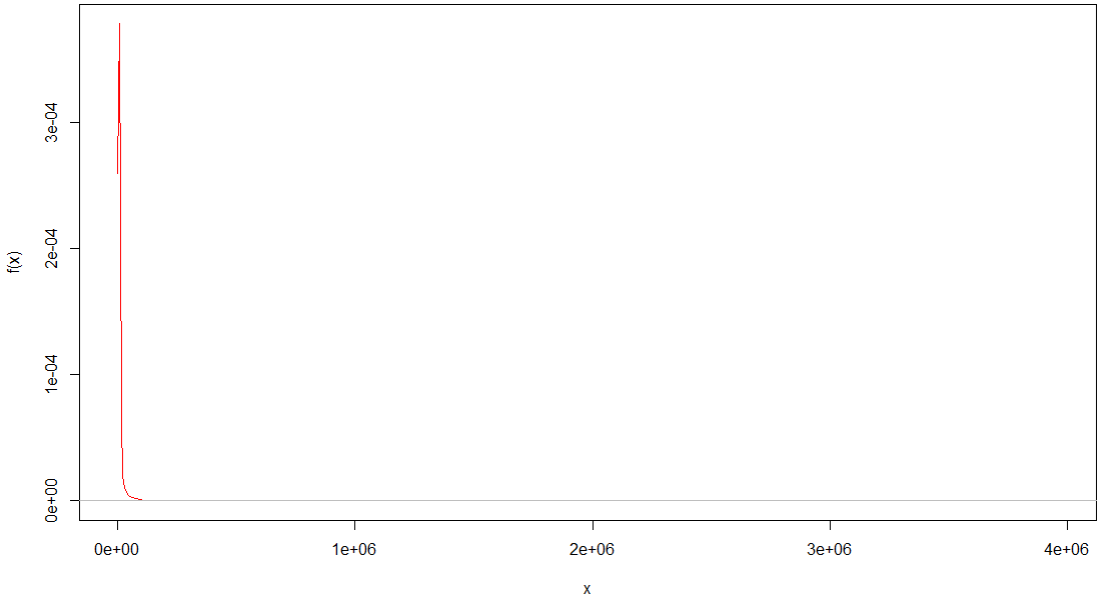


FIGURE 7.22 – Densité estimée de la branche RC corporelle

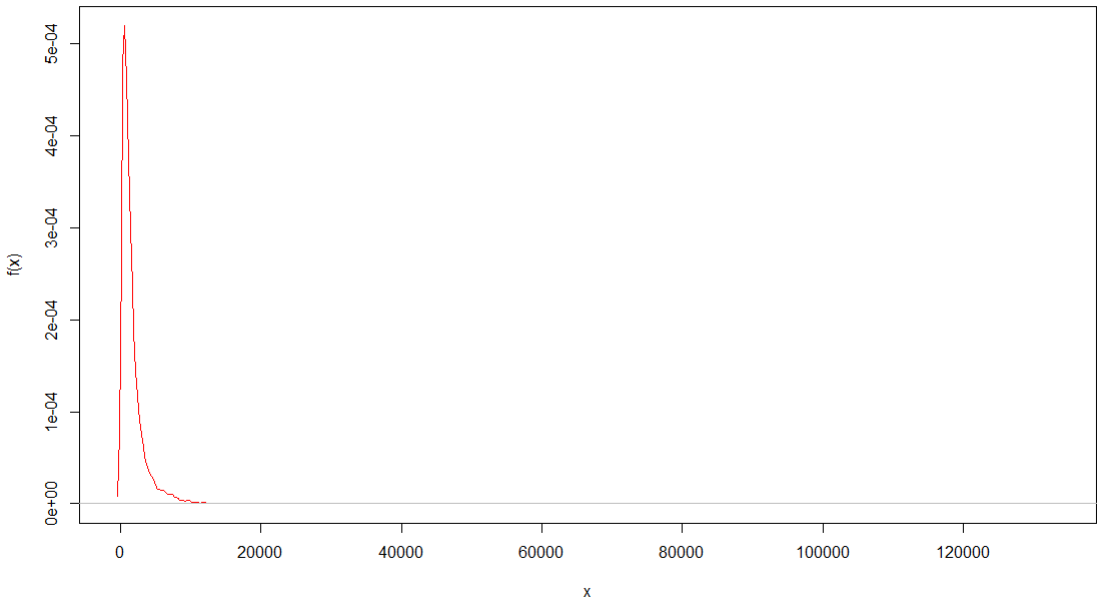


FIGURE 7.23 – Densité estimée de la branche RC matérielle

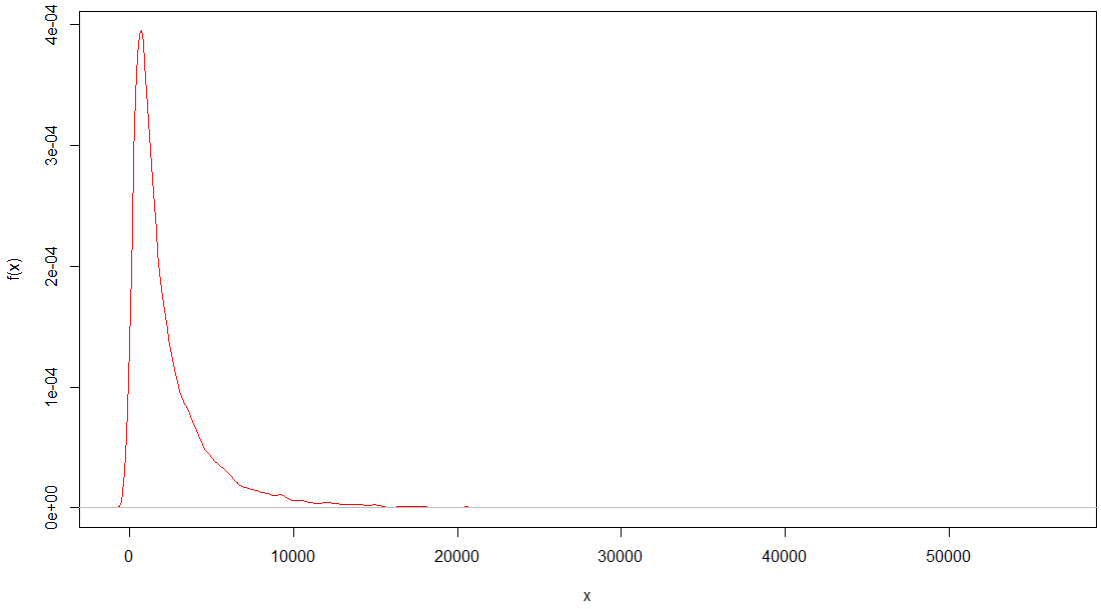


FIGURE 7.24 – Densité estimée de la branche Dommage

L'allure générale des courbes permet de restreindre nos tests aux lois :

$$F = \{Normale, Log - Normale, Gamma, Weibull, Exponentielle\}$$

Avec la méthode du maximum de vraisemblance, nous obtenons les résultats sous R :

Loi testée	RC corporelle	RC matérielle	Dommage
Normale	-124 946	-92 968	-95 784
Log-Normale	-96 258	-83 773	-90 687
Gamma	-	-	-
Weibull	-97 714	-84 175	-90 720
Exponentielle	-98 612	-84 715	-91 210

TABLE 7.7 – Valeur de lnL

L'ajustement avec la loi Gamma ne converge pas et n'est donc pas adaptée ici.

En prenant la loi dont la log-vraisemblance est la plus importante, nous trouvons la loi appropriée et ses paramètres associés :

- Les coûts liés à la garantie RC corporelle suivent une loi Log-Normale de moyenne 7,847 et d'écart-type 1,434 ;
- Les coûts liés à la garantie RC matérielle suivent une loi Log-Normale de moyenne 7,586 et d'écart-type 1,065 ;
- Les coûts liés à la garantie Dommage suivent une loi Log-Normale de moyenne 6,881 et d'écart-type 1,080.

A noter que, contrairement à l'application du chapitre 6, page 73, portant sur l'ensemble des données, les coûts suivent ici des lois Log-Normale et non des lois de Weibull. Ceci

est lié au fait que l'on réalise ici un ajustement sur un échantillon, autrement dit sur des données différentes et surtout sur un volume de données moindre.

Voici les densités ajustées pour les différentes branches de risque :

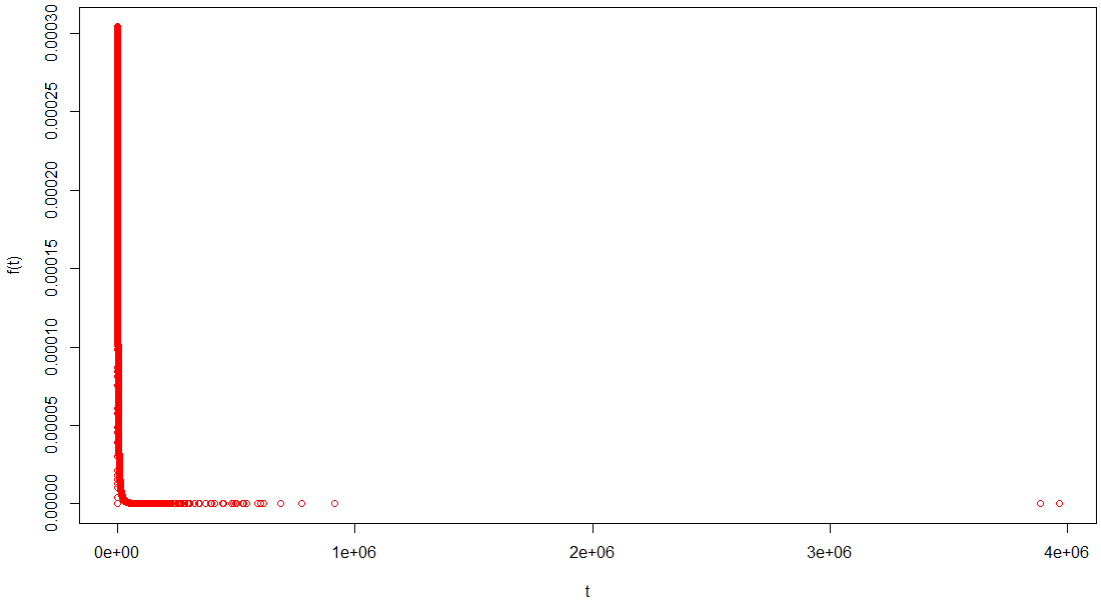


FIGURE 7.25 – Densité ajustée de la branche RC corporelle

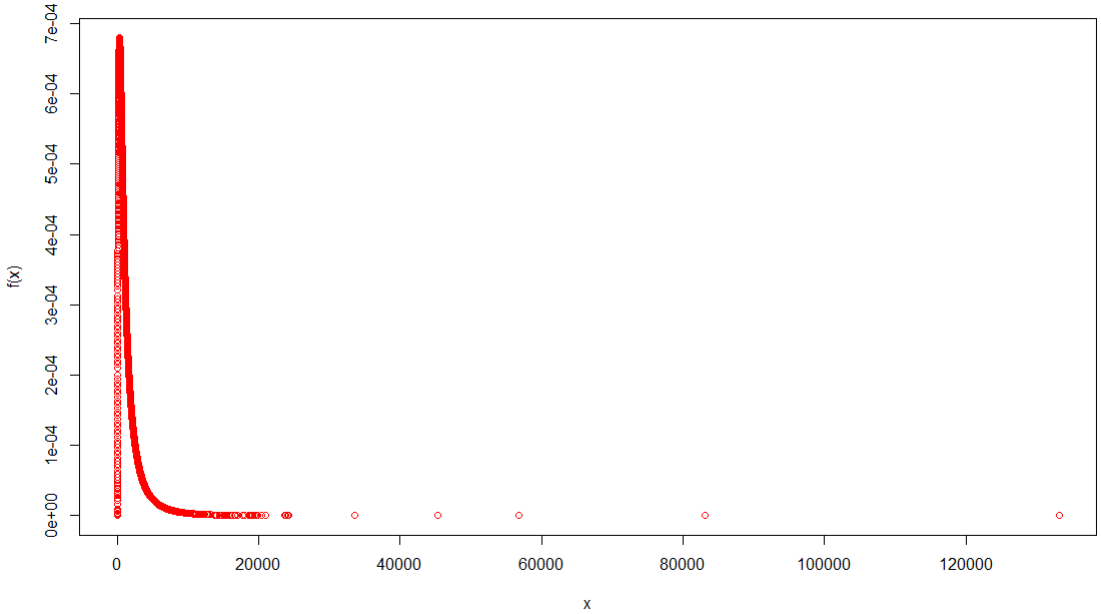


FIGURE 7.26 – Densité ajustée de la branche RC matérielle

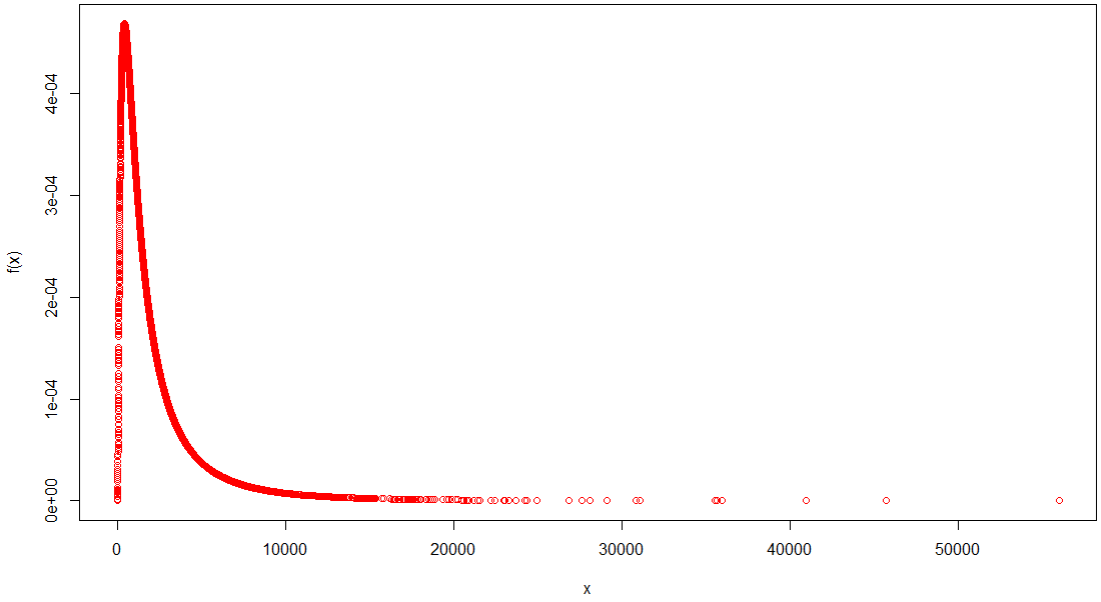


FIGURE 7.27 – Densité ajustée de la branche Dommage

En comparant ces graphiques avec les densités estimées, nous percevons une différence significative. Nous avons donc testé la qualité de l'ajustement avec le test de Kolmogorov-Smirnov avec un seuil de 5 % :

Ajustement	Statistique du test	P-Value
RC corporelle Log-Normale	0,040	3,73e-14
RC matérielle Log-Normale	0,038	1,141e-12
Dommage Log-Normale	0,035	6,015e-11

TABLE 7.8 – Résultat du test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test d'hypothèse utilisé ici pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue.

L'hypothèse nulle H_0, qui signifie que les coûts associés à nos garanties suivent bien une loi Log-Normale, est pour chacun des cas largement rejetée. Nous obtenons des P-Value largement inférieures à 1 % pour chacun des tests. La qualité d'ajustement n'est donc pas satisfaisante et c'est pour cette raison que nous estimerons les copules à l'aide d'une méthode semi-paramétrique. En effet, faute d'avoir pu les ajuster correctement, nous partons du principe que nous ne connaissons pas les lois des marginales.

7.3.3 Calibrage et estimation de la copule

L'objectif de la section est d'estimer les paramètres des copules candidates à la modélisation de la dépendance entre les branches de risque étudiées. L'étude sera basée sur l'ensemble des copules suivantes :

$$C_\theta = \{Gaussienne, Student, Clayton, Gumbel, Frank, Joe\}$$

Cette étude est donc basée sur deux grandes familles de copules :

- La famille des copules elliptiques avec la copule Gaussienne et la copule de Student ;

- La famille des copules archimédiennes avec la copule de Clayton, de Gumbel, de Frank et de Joe.

Puisqu'il n'a pas été possible de conclure sur les marginales lors de la section précédente, nous procéderons par méthode semi-paramétrique.

- **Méthodologie retenue**

Voici la méthode avec laquelle le calibrage et l'estimation de la copule ont été réalisés :

1. Nous choisissons la famille de copule à tester $C \in C_\theta$.
2. Le(s) paramètre(s) θ de la copule sont estimés par méthode *CML (Canonical Maximum Likelihood)* qui est détaillée section 4.9, page 59 [6].
3. Nous déterminons la statistique qui sera ici une distance.
4. La copule retenue sera celle qui minimisera cette distance.

- **Estimation des paramètres des copules testées**

Afin d'obtenir la forme paramétrique des marginales et de diminuer le risque de modélisation, l'estimation est réalisée à l'aide de la méthode *CML*, méthode semi-paramétrique.

Les copules présentent l'avantage d'estimer les paramètres indépendamment des formes paramétriques des marginales. En effet, l'un des intérêts principaux de la théorie des copules est de séparer la structure de dépendance entre les variables, induite par la copule, des lois marginales [7].

Les paramètres obtenus à l'aide de cette méthode sont restitués dans le tableau ci-dessous :

Copule	RC corporelle/RC matérielle	RC corporelle/Dommage	RC matérielle/Dommage
Normale	0,264	0,296	0,389
Student	0,241 et 7 ddl	0,320 et 8 ddl	0,399 et 15 ddl
Clayton	0,264	0,377	0,443
Gumbel	1,199	1,222	1,320
Frank	1,310	2,112	2,664
Joe	1,290	1,256	1,418

TABLE 7.9 – Paramètres estimés à partir de la méthode *CML*

Il s'agit maintenant de réaliser des tests d'ajustement afin de retenir la copule qui modélise le mieux la structure de dépendance induite par les coûts relatifs à nos couples de branches de risque.

6. L'estimation est réalisée à l'aide du package *Copula* [4] et plus précisément de la fonction *fitCopula* du logiciel R.
7. Tous ces éléments sont détaillés dans la partie théorique relative aux copules (chapitre 4, page 41).

7.3.4 Tests d'ajustement

- **Ajustement statistique**

La statistique de Cramér Von Mises :

La statistique de Cramér-Von Mises est définie par :

$$D_n = \sum_{i=1}^{n} \left\{ C_{\theta n}\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) - C_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \right\}^2$$

où C_n est la copule empirique et C_{θ_n} la copule de paramètre θ_n, simulée grâce à la fonction *pcopula*[8] sous R. Les résultats sont résumés dans le tableau suivant :

Copule	RC corporelle/RC matérielle	RC corporelle/Dommage	RC matérielle/Dommage
Normale	0,018	0,669	0,616
Student	0,016	0,406	0,553
Clayton	0,017	1,641	3,776
Gumbel	0,017	1,315	0,890
Frank	0,015	0,180	0,297
Joe	0,016	3,617	2,902

TABLE 7.10 – Valeur de la statistique de Cramér Von Mises

La statistique de Kolmogorov-Smirnov :

La statistique de Kolmogorov-Smirnov mesure :

$$D_n = \max_{1\leq i\leq n} \left| C_{\theta n}\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) - C_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \right|,$$

avec les mêmes notations que précédemment.

Voici les résultats obtenus à l'aide du logiciel R :

Copule	RC corporelle/RC matérielle	RC corporelle/Dommage	RC matérielle/Dommage
Normale	0,036	0,045	0,021
Student	0,032	0,018	0,020
Clayton	0,034	0,030	0,043
Gumbel	0,030	0,027	0,024
Frank	0,032	0,010	0,012
Joe	0,031	0,043	0,037

TABLE 7.11 – Valeur de la statistique de Kolmogorov-Smirnov

Interprétation des résultats des statistiques :

D_n mesure la distance entre la copule empirique et celle testée et ajustée. Ainsi, plus cette distance est faible, meilleur est l'ajustement et donc la copule à retenir est celle associée à la plus petite distance. Sachant cela, les deux méthodes qui retournent des valeurs cohérentes nous permettent de conclure que :

8. Cette fonction est accessible depuis le lien bibliographique suivant [4].

- La copule de Frank (selon Cramér-Von Mises) et celle de Gumbel (selon Kolmogorov-Smirnov) sont les meilleures parmi celles testées pour le couple RC corporelle/RC matérielle ;
- La copule de Frank est la meilleure parmi celles testées pour le couple RC corporelle/-Dommage ;
- La copule de Frank est la meilleure parmi celles testées pour le couple RC matérielle/Dommage.

L'adéquation de ces résultats a été vérifiée avec la fonction *BiCopSelect* du package *copula*[9] sous le logiciel R. Cette fonction permet, pour des rangs donnés, de retourner la meilleure copule au sens de l'*AIC*. Les mêmes résultats pour les couples RC corporelle/-Dommage et RC matérielle/Dommage ont été obtenus grâce à cette fonction intégrée à R. Pour le couple RC corporelle/RC matérielle, la fonction privilégie la copule de Gumbel que nous retiendrons pour la suite.

- **Ajustement graphique**

Voici les dépendogrammes empiriques et théoriques déterminés avec le logiciel R :

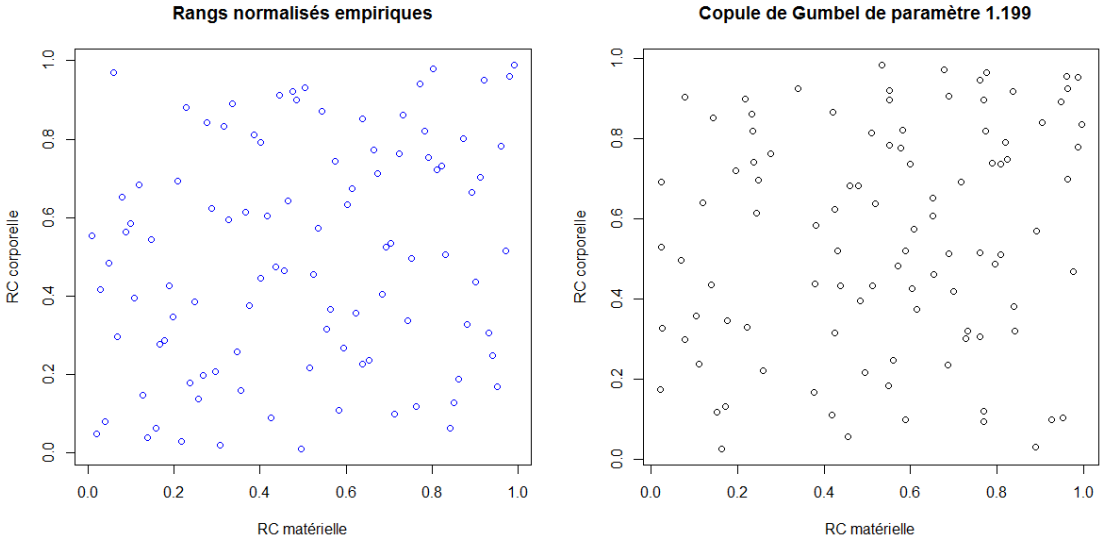


FIGURE 7.28 – Dépendogrammes empiriques et théoriques du couple RC corporelle/RC matérielle

9. Cette fonction est accessible depuis le lien bibliographique suivant [4].

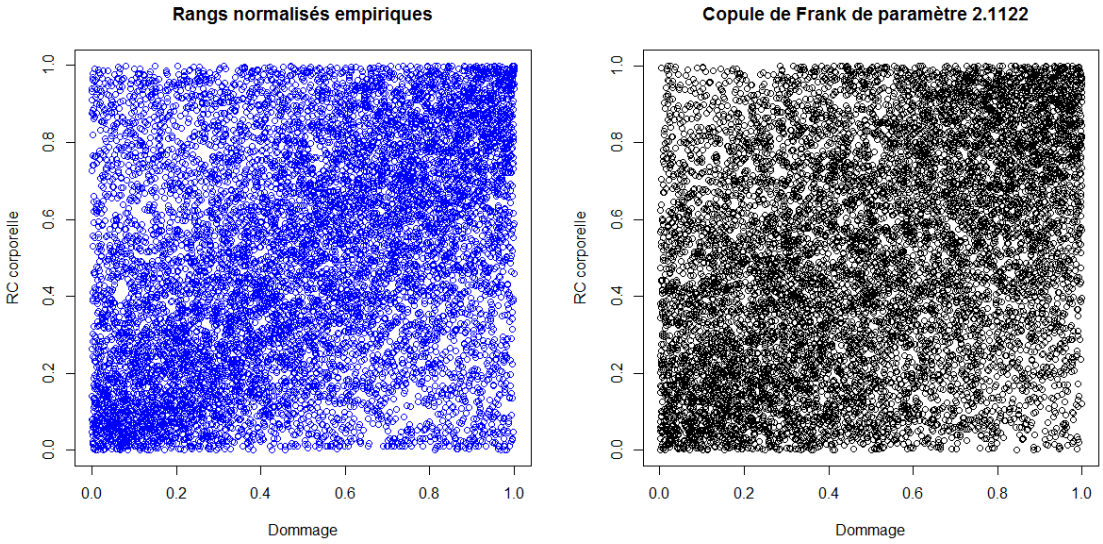


FIGURE 7.29 – Dépendogrammes empiriques et théoriques du couple RC corporelle/Dommage

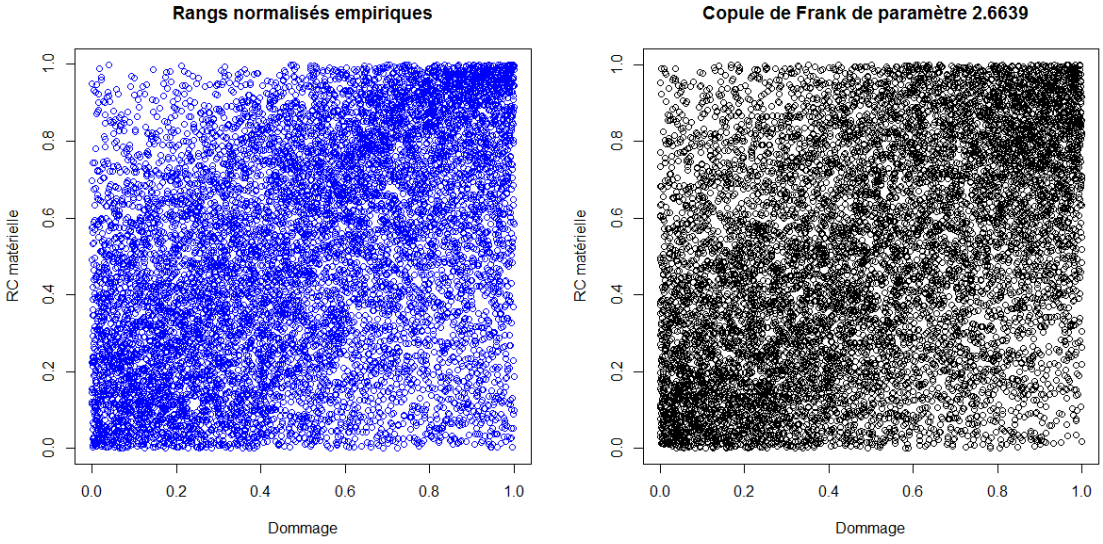


FIGURE 7.30 – Dépendogrammes empiriques et théoriques du couple RC matérielle/Dommage

Globalement, les résultats sont satisfaisants pour les couples RC corporelle/Dommage et RC matérielle/Dommage. La copule de Frank pour ces deux couples reproduit relativement bien la dépendance présente sur la diagonale ascendante, en prenant en compte une concentration légère sur les queues de distribution.

En revanche, il est très difficile de conclure sur la modélisation de la dépendance relative au couple RC corporelle/RC matérielle. Une concentration des points moins marquée sur les bords inférieurs à droite et supérieurs à gauche est respectée. Le dépendogramme des rangs normalisés empiriques ne présente pas à l'origine de dépendance importante. Il est donc difficile d'attendre beaucoup d'un ajustement. Cela est peut-être en partie la conséquence d'un faible nombre d'observations pour ce couple de branches de risque.

Nous pouvons étayer davantage notre analyse avec les graphiques des densités empiriques et des copules ajustées :

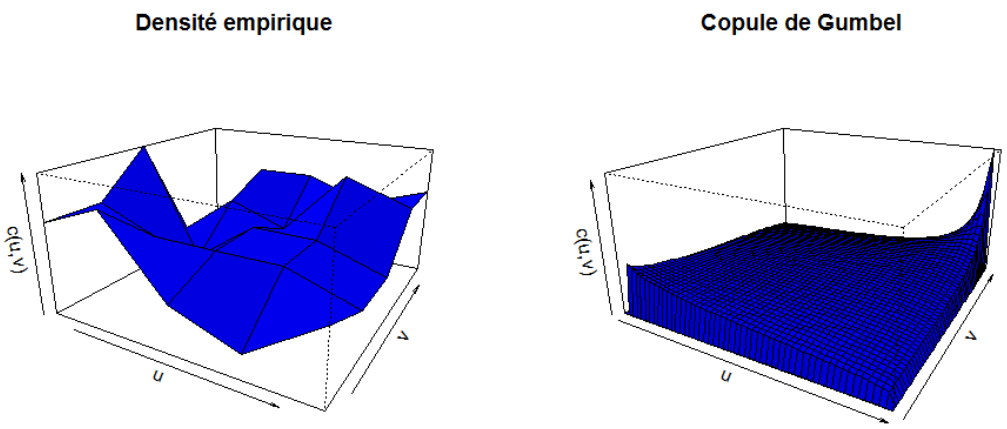


FIGURE 7.31 – Comparaison de la densité empirique et de la copule ajustée pour le couple RC corporelle/RC matérielle

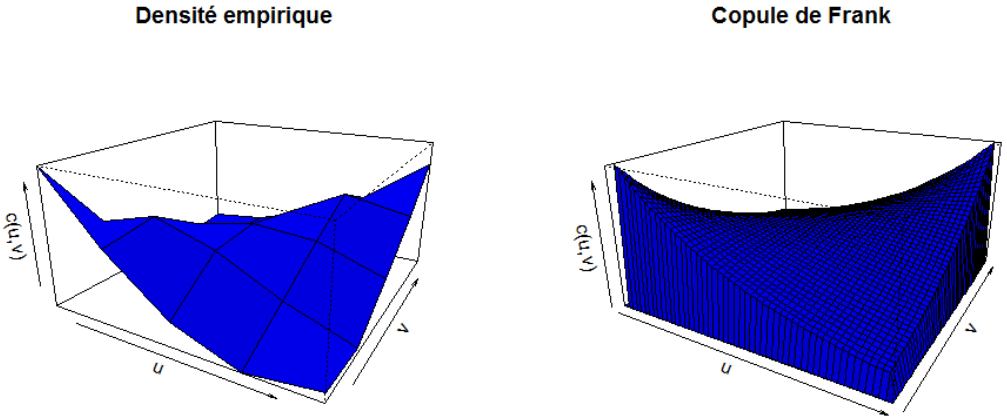


FIGURE 7.32 – Comparaison de la densité empirique et de la copule ajustée pour le couple RC corporelle/Dommage

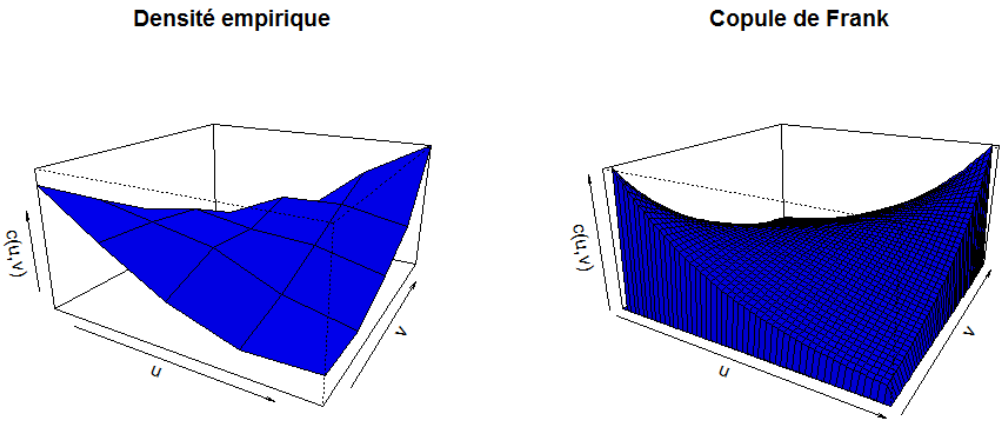


FIGURE 7.33 – Comparaison de la densité empirique et de la copule ajustée pour le couple RC matérielle/Dommage

Comme énoncé précédemment, nous remarquons d'après les graphiques ci-dessus que l'ajustement n'est pas parfait mais relativement correct pour les couples RC corporelle/-Dommage et RC matérielle/Dommage.

Cependant, il est difficile de conclure sur l'ajustement du couple RC corporelle/RC matérielle. En effet, cet ajustement est relativement mauvais mais cela est la conséquence directe d'une densité empirique présentant de fortes irrégularités. Cet inconvénient pourrait être atténué en augmentant la taille de notre échantillon. En effet, en augmentant le volume de données, nous réussirions à épouser davantage la forme de la densité empirique à partir de la copule estimée puisque la densité empirique présenterait certainement moins d'irrégularités.

7.4 Application sur la modélisation du besoin en fonds propres d'un assureur non-vie

Cette application est motivée par l'article de David Cadoux et Jean-Marc Loizeau [2006] [20].

7.4.1 Contexte et hypothèses de modélisation

Il est imposé aux assureurs de justifier d'un niveau minimum de fonds propres pour être en mesure de supporter les mouvements défavorables de résultats non anticipés. De plus, les compagnies d'assurance doivent aujourd'hui gérer leurs fonds propres de façon optimale tout en satisfaisant des intérêts qui divergent. En effet, les autorités de contrôle (l'ACPR en France), les agences de notation (Standard & Poor's par exemple) ainsi que les assurés, souhaitent que la compagnie d'assurance soit la plus solvable possible et qu'elle possède le plus de fonds propres possible. En revanche, les actionnaires veulent un rendement maximal de leur investissement et cela passe par limiter le capital injecté dans la compagnie.

L'objectif de cette application est de quantifier l'impact sur le besoin en fonds propres des dépendances entre branches de risque préalablement modélisées grâce à différentes copules.

L'intérêt est d'estimer l'erreur commise lorsqu'on estime le besoin en fonds propres en supposant l'indépendance entre deux branches de risque issues d'un portefeuille d'assurance non-vie. Différentes mesures de risque vont servir à déterminer le besoin en fonds propres et seront brièvement détaillées dans la section suivante. De plus, le risque considéré est celui relatif aux sinistres bruts de réassurance.

Préalablement, il est nécessaire de détailler le modèle qui prend en compte l'évolution des fonds propres de la société d'assurance considérée. Afin d'obtenir des résultats faciles à interpréter, des hypothèses ont été posées. Ces dernières devraient également permettre l'évaluation simplifiée de la situation financière de la société d'assurance considérée dans cette application.

La charge agrégée des sinistres relatifs à deux branches de risque est en fait la somme des charges des sinistres relatives à ces deux branches.
Il s'agit de déterminer le besoin en fonds propres de deux façons différentes. D'une part en considérant les charges sinistres relatives à deux branches de risque indépendantes et d'autre part, en prenant en compte la structure de dépendance entre deux branches de risque, déterminée à l'aide de la théorie des copules. Dans le premier cas, les charges sont tout simplement agrégées et dans le second cas, les charges sont additionnées en prenant en compte la structure de dépendance.

Ce type de modélisation permet de projeter la situation financière de la société d'assurance sur un horizon de un an. La charge annuelle des sinistres est la seule variable aléatoire du modèle. Elle est brute de réassurance [10].

7.4.2 Mesures de risque

La notion théorique utilisée en pratique pour déterminer le besoin en fonds propres d'une société d'assurance donnée est celle de mesures de risque. Ces dernières permettent de déterminer un niveau de capital pour un portefeuille de branches de risque données.

Il s'agit dans un premier temps de définir ce qu'est une mesure de risque, puis dans un second temps, les propriétés qui nous intéressent relatives à ces mesures de risque. Enfin, les mesures de risque les plus utilisées en pratique pour déterminer un niveau optimal de fonds propres seront détaillées.

Définition

Cette définition s'inspire de celle donnée dans Denuit et Charpentier [2004] [34] et dans Cadoux et Loizeau [2006] [20].

Soit Ω l'ensemble fini des états de nature possibles et X une variable aléatoire réelle qui à un état de la nature ω associe le réel $X(\omega)$. La charge agrégée de sinistres de plusieurs branches de risque (garanties) est ainsi formalisée. Il est possible de caractériser X à l'aide de sa fonction de répartition F_X : $F_X(x) = \mathbb{P}(\omega | X(\omega) \leq x) = \mathbb{P}(X \leq x)$.

On appelle mesure de risque toute application ρ associant un risque X à un réel $\rho(X) \in \mathfrak{R}_+ \cup \{+\infty\}$.
Il est possible d'exprimer le besoin en capital d'une société d'assurance non-vie à partir de cette fonction où X représente le résultat de la société, le besoin en capital étant fonction

10. Ce qui correspond à la non intégration du résultat de réassurance.

de cette variable.

Propriétés

Bien qu'il ne soit pas spécifié quelles propriétés une mesure de risque doit satisfaire, les propriétés énoncées *infra* doivent être satisfaites pour réaliser une estimation du besoin en capital la plus fiable possible [11].

Si on considère deux risques X et Y, alors on peut énoncer les propriétés suivantes :

- **Invariance par translation**

$\rho(X + c) = \rho(X) + c$ pour toute constante c.

- **Sous-additivité**

$\rho(X + Y) \leq \rho(X) + \rho(Y)$.

- **Homogénéité positive**

$\rho(cX) = c\rho(X)$ pour toute constante c positive.

- **Monotonie**

$\mathbb{P}[X < Y] = 1 \Rightarrow \rho(X) \leq \rho(Y)$.

Une mesure de risque vérifiant ces propriétés est considérée comme étant *cohérente* par Artzner et al. [1999]. Contrairement à la Tail-VaR, la VaR n'est pas cohérente puisqu'elle n'est pas sous-additive.

Mesures de risque retenue afin de déterminer le niveau de fonds propres nécessaire

L'objet de ce paragraphe est de décrire les mesures de risque utilisées dans l'application pour estimer le besoin en fonds propres.

- **La *Value at Risk (VaR)***

Cette mesure de risque est équivalente à la notion de Sinistre Maximum Probable (SMP) utilisée en assurance non-vie. La *VaR* est à l'origine très utilisée en finance avant de l'être en assurance.

La *Value at Risk (VaR)* de niveau α [12] associé au risque X est donnée par :

$$VaR_\alpha(X) = \text{Inf}\{x | F_X(x) \geq \alpha\} \text{ [13]}. \tag{7.1}$$

De plus, $VaR_\alpha(X) = F_X^{-1}(\alpha)$ où F_X^{-1} est la fonction quantile de la loi de X. L'avantage de la *VaR* est qu'elle est relativement facile à calculer et à interpréter : $VaR_\alpha(X)$ représente tout simplement la quantité qui va permettre de couvrir le montant de sinistres engendré

11. Il existe d'autres propriétés non présentées dans ce mémoire qui sont spécifiques au calcul du chargement de la cotisation.
12. Exemple : α est égal à $99,5$ % dans la formule standard de Solvabilité II (SII).
13. Cette définition est reprise de Frédéric Planchet et al. [2005] [40].

par le risque X avec une probabilité α.

La notion de *Value at Risk* est directement liée à celle de probabilité de ruine. En effet, si on considère une société d'assurance qui prend en compte un unique risque X et qui a un montant de "ressources disponibles" correspondant à la $VaR_\alpha(X)$, alors la probabilité de ruine associée à cette société est égale à $1-\alpha$.

L'inconvénient de la *Value at Risk* est, comme énoncé *supra*, qu'elle n'est pas sous-additive et qu'elle ne prend pas en compte la sévérité de la ruine. Cependant, ce concept peut avoir du sens lorsqu'on se place dans un objectif de solvabilité. C'est pour cette raison qu'il a été intéressant de la calculer lors de notre application.

- **La *Tail Value at Risk (TVaR)***

Les définitions relatives à ce paragraphe sont issues de Frédéric Planchet et al. [2005] [40].

La *Tail-Value-at-Risk (TVaR)* de niveau α [14] associée au risque X est donnée par :

$$TVaR_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 F_X^{-1}(p)dp. \tag{7.2}$$

La $TVaR$ peut s'exprimer en fonction de la VaR :

$$TVaR_\alpha(X) = VaR_\alpha(X) + \frac{1}{1-\alpha}\mathbb{E}\left[(X - VaR_\alpha(X))^+\right]. \tag{7.3}$$

Le deuxième terme du membre de droite représente la perte moyenne au-delà de la VaR. La $TVaR$ est donc relativement sensible à la forme de la queue de distribution. De ces équations, on a : $TVaR_\alpha(X) < +\infty \Leftrightarrow \mathbb{E}[X] < +\infty$, $\forall \alpha \in]0;1[$.

L'expected shortfall (ES_α) de niveau de probabilité α est la perte moyenne au-delà de la VaR au niveau α et est donnée par :

$$ES_\alpha(X) = \mathbb{E}\left[(X - VaR_\alpha(X))^+\right]. \tag{7.4}$$

Si X représente la charge brute de sinistres, $ES_\alpha(X)$ est le montant de la prime Stop-Loss dont la rétention pour l'assureur est la VaR au niveau α.

La conditionnal Tail Expectation (CTE) de niveau α est le montant de la perte moyenne sachant que cette dernière dépasse la VaR au niveau α et est donnée par :

$$CTE_\alpha(X) = \mathbb{E}\left[X | X > VaR_\alpha(X)\right]. \tag{7.5}$$

Comme énoncé *supra*, la $TVaR$ est une mesure de risque cohérente. Si la fonction de répartition F_X du risque X est continue, alors la $TVaR$ coïncide avec la CTE.

14. Exemple : α est égal à 99 % dans la formule standard du *Swiss Solvency Test (SST)*.

7.4.3 Calcul du besoin en fonds propres en prenant en compte la structure de dépendance entre les branches de risque

Pour chaque couple de garantie, il s'agit de déterminer la charge agrégée de sinistres à l'aide de la copule qui modélise le mieux la structure de dépendance relative à ces couples de garantie. Comme énoncé précédemment, la charge agrégée de sinistres entre deux branches de risque correspond à la somme des charges de sinistres de ces deux branches.

A partir de la copule retenue pour modéliser au mieux la dépendance relative à un couple de branches de risque étudiées, il est possible d'obtenir les distributions agrégées des charges de sinistres associées à ce couple de branches de risque en prenant en compte le type de dépendance. Pour ce faire, trois étapes sont à réaliser :

- Effectuer 10 000 simulations indépendantes d'un vecteur uniforme (u_1, u_2) de loi jointe $C(u_1, u_2)$ à partir de l'algorithme de Marshall-Olkin présenté en annexe .10, page 148, [15],

- Déduire les charges de sinistres relatives à chacune des branches : $X = (X_1, X_2) = \left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$,

- Estimer, pour chacun des 10 000 couples de charges de sinistres simulés $X^{(i)} = (X_1^{(i)}, X_2^{(i)}), i = 1, ..., 10000$, la charge agrégée sur les deux branches de risque : $S_{TOTAL}^{(i)} = X_1^{(i)} + X_2^{(i)}$.

Il est maintenant possible d'obtenir une distribution empirique de la charge de sinistres totale à partir des 10 000 réalisations de S_{TOTAL}.

7.4.4 Calcul du besoin en fonds propres en considérant l'indépendance entre les branches de risque

Pour mémoire, les charges relatives aux branches de risque étudiées suivent des lois Log-Normale.

En considérant l'indépendance entre les charges des branches de risque, la loi de S_{TOTAL} est donnée par le produit de convolution entre les densités de probabilité de chacune des deux branches étudiées et nous avons :

$$\begin{aligned} f_{S_{TOTAL}}(z) &= \int_{-\infty}^{+\infty} f_{S_1}(x) \times f_{S_2}(z-x) dx \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_1 x \sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu_1)^2}{2\sigma_1^2}\right) \times \frac{1}{\sigma_2 (z-x)\sqrt{2\pi}} \exp\left(-\frac{(\ln(z-x) - \mu_2)^2}{2\sigma_2^2}\right) dx \end{aligned} \tag{7.6}$$

La distribution de la charge agrégée pour chaque couple de branches de risque dans le cas indépendant est maintenant déterminée.

La section suivante vise à comparer les distributions obtenues dans le cas où la dépendance est prise en compte et dans le cas indépendant.

15. Ces simulations ont été effectuées à l'aide la fonction *rCopula* [4] sous R.

7.4.5 Comparaison des distributions obtenues en considérant la dépendance et l'indépendance

Afin de comparer les distributions obtenues dans le cas où la dépendance est prise en compte entre les branches de risque et dans le cas où l'indépendance entre les branches est supposée, les quantiles de ces distributions sont calculés pour deux niveaux donnés : 75 % et 99,5 %.

Les copules qui modélisent le mieux la dépendance entre les branches de risque du portefeuille étudié sont archimédiennes. Ainsi, les observations et conclusions relatives à cette application ne concernent que les copules archimédiennes et en particulier la copule de Gumbel et la copule de Frank. Il serait intéressant de réaliser cette application avec des branches de risque dont la dépendance est modélisée par des copules elliptiques.

Les tableaux suivants présentent les résultats du modèle d'évaluation du besoin en fonds propres, calculé à l'aide des mesures de risque retenues et présentées *supra*.

Voici les résultats obtenus à l'aide du logiciel R :

RC corporelle/RC matérielle	$VaR_{75\%}$	$VaR_{99,5\%}$	$TVaR_{75\%}$	$TVaR_{99,5\%}$
Cas dépendant (copule de Gumbel)	13 765	142 199	35 577	220 767
Cas indépendant	13 695	139 599	34 707	217 553
% écart	0,51 %	1,86 %	2,51 %	1,48 %

TABLE 7.12 – Comparaison des valeurs (en euros) de mesures de risque des distributions de la charge agrégée dans le cas dépendant et indépendant pour le couple RC corporelle/RC matérielle

RC corporelle/Dommage	$VaR_{75\%}$	$VaR_{99,5\%}$	$TVaR_{75\%}$	$TVaR_{99,5\%}$
Cas dépendant (copule de Frank)	9 023	102 607	24 369	179 705
Cas indépendant	8 831	100 497	23 702	178 326
% écart	2,17 %	2,10 %	2,81 %	0,77 %

TABLE 7.13 – Comparaison des valeurs (en euros) de mesures de risque des distributions de la charge agrégée dans le cas dépendant et indépendant pour le couple RC corporelle/Dommage

RC matérielle/Dommage	$VaR_{75\%}$	$VaR_{99,5\%}$	$TVaR_{75\%}$	$TVaR_{99,5\%}$
Cas dépendant (copule de Frank)	7 811	86 721	20 814	145 661
Cas indépendant	7 717	85 131	20 055	144 454
% écart	1,22 %	1,87 %	3,78 %	0,84 %

TABLE 7.14 – Comparaison des valeurs (en euros) de mesures de risque des distributions de la charge agrégée dans le cas dépendant et indépendant pour le couple RC matérielle/Dommage

Ces résultats mettent en évidence que la prise en compte de la modélisation des dépendances, en utilisant la théorie des copules, augmente légèrement le besoin en fonds

propres. Cette remarque est valable pour deux mesures de risque (la VaR et la $TVaR$) et pour deux seuils α différents (75 % et 99,5 %).

Dans le cas de la prise en compte de la dépendance entre deux garanties données, le maintien d'un seuil de sécurité équivalent à celui du cas d'indépendance entraîne une hausse de la VaR ou de la $TVaR$ allant de $0,51$ % à $3,78$ %.

7.5 Conclusion et ouverture

Cette étude permet de mettre en évidence que la prise en compte de dépendances positives entre les charges de sinistres relatives à deux branches de risque, modélisées à partir de copules, augmente légèrement le besoin en fonds propres d'une compagnie d'assurance non-vie. Cette légère hausse devrait inciter le lecteur à s'interroger sur l'intérêt de prendre en compte la dépendance entre branches de risque à travers une telle modélisation.

Cependant, il serait intéressant de modéliser, au sein du même portefeuille, deux branches de risque présentant de la dépendance négative, ce qui pourrait modifier les conclusions de cette étude. De plus, seules les copules de Gumbel et de Frank ont été appliquées ici. L'utilisation de copules elliptiques comme la copule de Student permettrait de modéliser des dépendances de queue et ainsi d'approfondir l'analyse.

Aussi, le modèle utilisé dans cette application permet de projeter la situation financière de la société avec un horizon annuel qui est l'année civile. Ce modèle est discret puisqu'il ne modélise pas la situation financière à un instant t quelconque de l'année. Une analyse dynamique de l'évolution des fonds propres sur plusieurs exercices successifs n'est pas réalisée dans ce mémoire mais pourrait faire l'objet d'une seconde étude.

Dans ce sens, le modèle décrit *supra* est une approche simplifiée et ne prétend pas apporter une réponse exhaustive quant à l'étude de la solvabilité d'une société d'assurance non-vie. Pour ce faire, il faudrait être en mesure de quantifier l'impact de la réassurance et d'étudier l'ensemble des risques supportés par la société d'assurance.

Conclusion

Ce mémoire propose d'une part, deux approches de provisionnement ligne à ligne visant à déterminer le montant de Provision pour Sinistres A Payer (PSAP) d'un portefeuille d'assurance automobile. Ces approches passent par l'analyse et la modélisation du coût des sinistres ligne à ligne.
D'autre part, il s'agit de modéliser la dépendance qui existe entre les différentes branches de risque étudiées dans ce portefeuille afin de quantifier l'impact de la prise en compte de cette dépendance sur le besoin en fond propre d'une société d'assurance donnée.

Le choix de l'étude des trois garanties a été motivé par le fait que la branche RC corporelle est une branche caractérisée par des délais de développement relativement longs contrairement aux branches RC matérielle et Dommage. Ces particularités ont été observées au travers de la détermination des fonctions de survie à l'aide de l'estimateur *Kaplan-Meier*.

En ce qui concerne le premier volet de ce mémoire, la notion de données censurées est exploitée afin de réaliser la meilleure estimation possible de PSAP pour chacune des trois garanties étudiées dans ce mémoire. Le recourt à la théorie des modèles de censure a été nécessaire afin de prendre en considération la notion de censure présente dans les données et qui caractérise l'état d'un sinistre (clos ou en cours).

Afin d'effectuer des comparaisons, une méthode déterministe (*Chain ladder*) a été mise en application sur les données agrégées. Puis, le recourt à une méthode stochastique (le modèle de *Mack*) a permis de mesurer l'incertitude liée aux provisions estimées et ainsi de calculer l'erreur de prédiction. La valeur ajoutée de ce modèle est de pouvoir, par proposition d'une hypothèse sur la distribution des provisions, construire un intervalle de confiance pour encadrer les montants de PSAP estimés. Enfin, la connaissance des montants de provisions "dossier" a été très appréciée, permettant l'acquisition d'un référentiel et ainsi de prendre du recul quant aux résultats obtenus par les deux approches ligne à ligne et ceux issus d'une méthode appliquée sur des données agrégées.

Au niveau des résultats obtenus et malgré la particularité des trois garanties étudiées, les analyses des lois marginales des coûts relatifs aux trois garanties convergent. Elles nous amènent à conclure que l'approche 2, tenant compte des sinistres encore en cours à l'aide d'un modèle de censure, est la plus adaptée pour provisionner les types de garanties étudiés. En effet, cette méthode de provisionnement ligne à ligne permet de maximiser la quantité d'informations utilisées par rapport à celle mise à disposition et se caractérise par des résultats relativement proches des montants de provision "dossier" (entre 1,70 % et 3,92 % d'écart par rapport aux montants de provision "dossier").

Comparativement à l'approche 2 (et même aux deux approches de provisionnement ligne à ligne), les résultats associés à *Chain Ladder* sont plus éloignés des provisions "dossier" pour les trois garanties étudiées. Ces résultats conduisent à sous-estimer nettement les montants de PSAP des garanties RC matérielle et Dommage. Cependant, si on regarde de

plus près les intervalles de confiance délivrés par le modèle de *Mack*, on s'aperçoit que ces derniers sont assez "larges", en particulier pour la garantie Dommage. Ceci témoigne d'une grande incertitude quant aux estimations réalisées à partir des triangles de paiements, ainsi que d'une volatilité associée relativement importante.

Pour un sinistre concerné par plusieurs remboursements, nous ne disposions pas des différents montants de règlement ainsi que des dates associées mais seulement du montant des règlements cumulés au 31/12/2009. Pourtant, ces informations auraient pu permettre l'utilisation d'autres modèles [16] et certainement d'améliorer la précision de nos estimations de PSAP. Cependant, le volume de données important relatif au portefeuille étudié nous a permis de réaliser des estimations relativement correctes. Il aurait été intéressant d'utiliser des processus de règlements et des processus d'états en définissant la notion de mesures de provisions cohérente. Seulement, ceci requiert la connaissance des dates des différents règlements, or cette dimension n'est pas présente dans nos données.

Concernant le second volet du mémoire, l'application réalisée permet de mettre en évidence que la prise en compte de dépendances positives entre les charges de sinistres relatives à deux branches de risque, modélisées à partir de copules, augmente légèrement le besoin en fonds propres d'une compagnie d'assurance non-vie pour le portefeuille et les branches de risque considérées [17].

Une dépendance positive a été détectée et modélisée entre les trois couples de branches de risque étudiées. Quant aux lois marginales, la loi Log-Normale pour les charges de sinistres de chacune des trois garanties a été retenue. Ensuite, la simulation, à l'aide des copules retenues pour caractériser le mieux la dépendance entre nos trois couples de garanties (Franck pour les couples RC corporelle/Dommage et RC matérielle/Dommage et Gumbel pour le couple RC corporelle/RC matérielle), nous a permis de déterminer des distributions agrégées de charges de sinistres pour chaque couple de branches de risque étudié.

Les effets de la dépendance entre les coûts des branches de risque sur la solvabilité de l'assureur ont ainsi pu être déterminés. Cette étude montre que considérer l'hypothèse d'indépendance, si elle n'est pas vérifiée, conduit l'assureur à sous-estimer légèrement le besoin en fonds propres lors de l'accumulation des risques lorsque ces derniers sont positivement dépendants.

Cette faible hausse devrait inciter le lecteur à s'interroger sur l'intérêt de prendre en compte la dépendance entre branches de risque pour le portefeuille étudié. C'est pourquoi il serait intéressant de modéliser, au sein du même portefeuille, deux branches de risque présentant de la dépendance négative, ce qui pourrait modifier les conclusions de cette étude. De plus, seules les copules de Gumbel et de Frank ont été appliquées ici. L'utilisation de copules elliptiques, comme la copule de Student, permettrait de modéliser des dépendances de queue et ainsi d'approfondir l'analyse.

Le modèle utilisé dans cette application permet de projeter la situation financière de la société avec un horizon annuel qui est l'année civile. Ce modèle est discret puisqu'il ne modélise pas la situation financière à un instant t quelconque de l'année. Une analyse dynamique de l'évolution des fonds propres sur plusieurs exercices successifs n'est pas

16. Par exemple, les mémoires de Ngoc An Dinh et Gilles Chau [2012] [36] mettent en oeuvre une approche basée sur les dynamiques markoviennes.

17. David Cadoux et Jean-Marc Loizeau [2006] [20] ont abouti à la même conclusion en modélisant la dépendance avec une copule HRT.

réalisée dans ce mémoire mais pourrait faire l'objet d'une seconde étude.

Dans ce sens, la modélisation effectuée est une approche simplifiée et ne prétend en rien apporter une réponse exhaustive quant à l'étude de la solvabilité d'une société d'assurance non-vie. Pour ce faire, il faudrait être en mesure de quantifier l'impact de la réassurance et d'étudier l'ensemble des risques supportés par la société d'assurance. Il serait également intéressant de prendre en compte, dans la modélisation, les différents délais de développement illustrés lors de l'application sur le provisionnement.

Enfin, au regard d'un volume de données insuffisant pour les trois couples étudiés, il n'a pas été possible d'effectuer une analyse de la dépendance en 3 dimensions (d=3). Il serait intéressant de modéliser la dépendance et d'effectuer une analyse du besoin en fonds propres pour le triplet de branches de risque en le considérant conjointement (sous condition de disposer d'un nombre d'observations suffisant).

Bibliographie

[1] Magali Kelle. *Provisionnement pour sinistres à payer : analyses et modélisations sur données détaillées. Bulletin Français d'Actuariat n° 13 / vol. 7 / Janvier 2007 - Juin 2007.* 2007. *2 citations pages 11 et 20*

[2] Site Internet de la FFSA. *http://www.ffsa.fr/sites/jcms/c_51562/fr/.* *2 citations pages 14 et 18*

[3] Site Internet de l'Institut des Actuaires. *http://www.institutdesactuaires.com/gene/main.php?base=081&revue=63&article=537.* *Cité page 15*

[4] Package *copula*. *http://cran.r-project.org/web/packages/copula/index.html*. *5 citations pages 23, 117, 118, 119, et 126*

[5] Site Internet documenté sur le logiciel R. *http://cran.r-project.org/.* *Cité page 23*

[6] Arthur Charpentier. *Computational Actuarial Science with R.* 2015. *Cité page 23*

[7] Site Internet de l'INSEE. *http://www.bdm.insee.fr/.* *Cité page 23*

[8] Ilan Habib et Stéphane Riban. *Quelle méthode de provisionnement pour des engagements non-vie dans Solvabilité 2. Mémoire ENSAE* . 2012. *Cité page 24*

[9] Noémie Rose. *Provisionnement en assurance non vie : utilisation de modèles paramétriques censurés. Mémoire ISUP* . 2009. *2 citations pages 30 et 33*

[10] Sandra Gotlib. *Comparaison des méthodes de provisionnement dans le cadre de la garantie Responsabilité Civile en flotte automobile. Mémoire ISUP* . 2012. *2 citations pages 30 et 33*

[11] Frédéric Planchet et Pierre Thérond. *Modèles de durée : applications actuarielles, Paris : Economica.* 2006. *Cité page 30*

[12] Arthur Charpentier. *http://freakonometrics.hypotheses.org/.* *Cité page 38*

[13] Valérie Lagenebre. *Prise en compte des dépendances entre risques par la théorie des Copules : Impact sur le besoin en fonds propres de l'assureur. Mémoire EURIA* . 2009. *2 citations pages 41 et 47*

[14] Kamal Armel. *Structure de dépendance des générateurs de scénarios économiques, Modélisation et Simulation. Mémoire EURIA* . 2010. *Cité page 41*

[15] Belguise Olivier. *Tempêtes : Etude des dépendances entre les branches Auto et Incendie avec la théorie des copulas. Mémoire ULP* . 2001. *2 citations pages 41 et 42*

[16] Esterina Masiello. *Modeling dependance by copula functions in an insurance context. Mémoire ISFA* . 2010. *Cité page 41*

[17] Gwladys Marylou Noutong. *Les Risk Drivers : modélisation de la dépendance entre branches d'activités d'une compagnie d'assurance Non-Vie. Mémoire ISFA* . 2013. *Cité page 41*

[18] Christian Partrat. *"Corrélation" entre risques d'assurances non-vie : une introduction. Bulletin Français d'Actuariat n° 10 / vol. 6 / pp. 79-91.* 2003. *Cité page 42*

[19] Olivier Junior Karusisi. *Modélisation de la dépendance des défauts par les fonctions Copules. Mémoire EURIA* . 2007. *Cité page 42*

[20] David Cadoux et Jean-Marc Loizeau. *Copules et dépendances : application pratique à la détermination du besoin en fonds propres d'un assureur non-vie. Institut des Actuaires.* 2006. *4 citations pages 46, 122, 123, et 130*

[21] Paul Embrechts avec Alexander McNeil et Daniel Straumann. *Correlation and dependence in risk management : properties and pitfalls.* 2002. *Cité page 46*

[22] Stefano Demarta et Alexander J. McNeil. *The t Copula and Related Copulas.* 2004. *4 citations pages 46, 60, 145, et 148*

[23] Filip Lindskog avec Alexander McNeil et Uwe Schmock. *Kendall's tau for elliptical distributions.* 2003. *2 citations pages 46 et 148*

[24] Jean-Claude Boies et Christian Genest. *Une méthode graphique de détection de la dépendance.* 2003. *2 citations pages 50 et 51*

[25] Roger B. Nelsen. *An Introduction to Copulas.* 1999. *3 citations pages 53, 55, et 58*

[26] Harry Joe. *Multivariate models and dependence concepts.* 1997. *2 citations pages 53 et 60*

[27] Fonction *rank*. *https://stat.ethz.ch/r-manual/r-devel/library/base/html/rank.html* . *Cité page 54*

[28] Sklar A. *Fonction de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris, 8, 229-231.* 1959. *Cité page 55*

[29] Kaitai Fang avec Samuel Kotz et Kai Wang Ng. *Symmetric multivariate and related distributions.* 1990. *Cité page 57*

[30] Christian Genest et R Jock MacKay. *The joy of copulas : Bivariate distributions with uniform marginals.* 1986. *Cité page 57*

[31] Paul Deheuvels. *La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Bulletin de la Classe des Sciences, V. Série, Académie Royale de Belgique.* 1979. *2 citations pages 59 et 110*

[32] Fonction *fitdistr*. *http://stat.ethz.ch/r-manual/r-patched/library/mass/html/fitdistr.html* . *Cité page 74*

[33] Claire Guillaumin. *Détermination d'une méthode de provisionnement pour les créances douteuses. Mémoire Université Paris Dauphine* . 2008. *Cité page 76*

[34] Arthur Charpentier et Michel Denuit. *Mathématiques de l'assurance non-vie, Paris : Economica.* 2004. *2 citations pages 76 et 123*

[35] Nessi J. M. Nisipasu E. Reiz O. Partrat C., Lecoeur E. *Provisionnement technique en assurance non-vie, Perspectives actuarielles modernes, Paris : Economica.* 2007. *Cité page 76*

[36] Ngoc An Dinh et Gilles Chau. *Mesures de provision cohérentes et méthodes ligne à ligne pour des risques non-vie. Mémoire ENSAE* . 2012. *2 citations pages 76 et 130*

[37] Nicoleta Tripa. *Quantification de la variabilité des Provisions pour Sinistres A Payer. Mémoire EURIA* . 2009. *Cité page 77*

[38] Procédure *lifetest*. http://support.sas.com/documentation/. *2 citations pages 78 et 79*

[39] Package *gplots*. http://cran.r-project.org/web/packages/gplots/index.html. *Cité page 110*

[40] Frédéric Planchet avec Pierre Thérond et Julien Jacquemin. *Modèles financiers en assurance*. 2005. *2 citations pages 124 et 125*

[41] Roy Mashal et Assaf Zeevi. *Beyond Correlation : Extreme Co-movements Between Financial Assets*. 2002. *Cité page 145*

[42] Wolfgang Breymann avec Alexandra Dias et Paul Embrechts. *Dependence structures for multivariate high-frequency data in finance*. 2003. *Cité page 145*

[43] Harry Joe et James J. Xu. *The Estimation Method of Inference Functions for Margins for Multivariate Models*. 1996. *Cité page 147*

[44] Marius Hofert. *Sampling Archimedean copulas. Computational Statistics and Data Abalysis 52 (2008) 5163-5174*. 2008. *Cité page 148*

Annexes

.1 Tableaux de la répartition des sinistres par garanties et années de survenance selon le fait qu'ils soient sans suite ou non

	Code de la garantie																								Total	
	A		C		D		E		F		J		K		M		S		T		V		W			
	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%
Indicateur des sinistres sans suite																										
N	4042942	16.88	9227360	38.52	111800	0.47	1482449	6.19	269687	1.13	2075	0.01	6000327	25.05	13932	0.06	2106	0.01	84641	0.35	133088	0.56	204476	0.85	21574883	90.06
O	987016	4.12	798120	3.33	4456	0.02	184632	0.77	288956	1.21	863	0.00	99062	0.41	621	0.00	145	0.00	4383	0.02	7279	0.03	6646	0.03	2382179	9.94
Total	5029958	21.00	10025480	41.85	116256	0.49	1667081	6.96	558643	2.33	2938	0.01	6099389	25.46	14553	0.06	2251	0.01	89024	0.37	140367	0.59	211122	0.88	23957062	100.00

FIGURE 34 – Proportion de sinistres classés sans suite par garanties

	Code de la garantie																								Total	
	A		C		D		E		F		J		K		M		S		T		V		W			
	Indicateur des sinistres sans suite																									
	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Années de survenance																										
1993	1.03	0.23	2.24	0.11	0.02	0.00	0.48	0.05	0.06	0.07	0.00	0.00	1.31	0.01	0.00	0.00	0.00	0.00	0.00	0.00	.	.	0.12	0.00	5.27	0.48
1994	1.04	0.23	2.29	0.11	0.02	0.00	0.49	0.05	0.06	0.09	0.00	0.00	1.47	0.02	0.00	0.00	0.00	0.00	0.00	0.00	.	.	0.08	0.00	5.46	0.51
1995	1.06	0.24	2.34	0.11	0.02	0.00	0.46	0.05	0.06	0.09	0.00	0.00	1.26	0.02	0.00	0.00	0.00	0.00	0.00	0.00	.	.	0.01	0.00	5.21	0.52
1996	1.02	0.26	2.27	0.13	0.02	0.00	0.44	0.05	0.06	0.09	0.00	0.00	1.26	0.02	0.00	0.00	0.00	0.00	0.00	0.00	.	.	0.01	0.00	5.08	0.55
1997	1.05	0.26	2.33	0.13	0.02	0.00	0.42	0.05	0.07	0.09	0.00	0.00	1.36	0.02	0.00	0.00	0.00	.	0.00	0.00	.	.	0.01	0.00	5.27	0.55
1998	1.09	0.27	2.40	0.17	0.02	0.00	0.42	0.05	0.07	0.10	0.00	0.00	1.48	0.02	0.00	0.00	0.00	0.00	0.01	0.00	.	.	0.03	0.00	5.52	0.61
1999	1.12	0.30	2.51	0.24	0.02	0.00	0.41	0.05	0.08	0.10	0.00	0.00	1.47	0.04	0.01	0.00	0.00	0.00	0.25	0.01	.	.	0.06	0.00	5.93	0.76
2000	1.08	0.30	2.48	0.25	0.02	0.00	0.41	0.05	0.07	0.10	0.00	0.00	1.45	0.03	0.01	0.00	0.00	0.00	0.01	0.00	.	.	0.07	0.00	5.60	0.72
2001	1.05	0.30	2.48	0.25	0.03	0.00	0.43	0.05	0.07	0.09	0.00	0.00	1.48	0.02	0.00	0.00	0.00	0.00	0.00	0.00	.	.	0.03	0.00	5.57	0.71
2002	0.98	0.28	2.36	0.24	0.03	0.00	0.40	0.05	0.07	0.07	0.00	0.00	1.54	0.03	0.01	0.00	0.00	.	0.01	0.00	.	.	0.06	0.00	5.46	0.66
2003	0.91	0.25	2.25	0.23	0.03	0.00	0.36	0.05	0.07	0.05	0.00	0.00	1.63	0.03	0.01	0.00	0.00	.	0.01	0.00	.	.	0.06	0.00	5.32	0.62
2004	0.91	0.25	2.20	0.23	0.03	0.00	0.32	0.04	0.07	0.05	0.00	0.00	1.55	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.00	0.03	0.00	5.16	0.61
2005	0.91	0.25	2.12	0.23	0.03	0.00	0.28	0.04	0.07	0.05	0.00	0.00	1.52	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.11	0.01	0.05	0.00	5.11	0.61
2006	0.89	0.23	2.08	0.24	0.03	0.00	0.25	0.04	0.06	0.05	0.00	0.00	1.55	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.11	0.01	0.05	0.00	5.03	0.60
2007	0.90	0.22	2.08	0.24	0.03	0.00	0.22	0.04	0.06	0.05	0.00	0.00	1.53	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.10	0.01	0.04	0.00	4.99	0.58
2008	0.93	0.15	2.02	0.22	0.03	0.00	0.19	0.03	0.06	0.04	0.00	0.00	1.57	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.10	0.01	0.03	0.00	4.95	0.47
2009	0.94	0.10	2.07	0.21	0.03	0.00	0.19	0.03	0.07	0.03	0.00	.	1.62	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.09	0.01	0.10	0.01	5.14	0.39
Total	16.88	4.12	38.52	3.33	0.47	0.02	6.19	0.77	1.13	1.21	0.01	0.00	25.05	0.41	0.06	0.00	0.01	0.00	0.35	0.02	0.56	0.03	0.85	0.03	90.06	9.94

FIGURE 35 – Proportion de sinistres classés sans suite par garanties et par années de survenance

	Code de la garantie																								Total		Pourcentage
	A		C		D		E		F		J		K		M		S		T		V		W				
	Indicateur des sinistres sans suite																										
	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	
	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	Effectif	
Années de survenance																											
1993	246301	56264	537499	26831	5174	235	115842	11906	13225	16375	34	66	314879	2834	22	69	23	1	532	28	.	.	28061	772	1261592	115381	5.75
1994	249118	56235	548778	26593	5199	247	118507	11947	14261	20387	61	92	351524	5379	19	14	29	1	621	20	.	.	19884	634	1308001	121549	5.97
1995	252758	58272	559779	27133	5090	296	110183	11873	14375	22058	105	106	300715	4457	23	37	50	2	835	39	.	.	3448	107	1247361	124380	5.73
1996	243209	61253	543181	29980	5090	304	106310	12176	15032	21348	145	113	301006	5285	3	8	25	1	855	89	.	.	2842	104	1217698	130661	5.63
1997	252438	61981	557733	30751	5144	276	100915	11037	16297	22030	179	90	326433	5062	717	35	17	.	626	37	.	.	2727	100	1263226	131399	5.82
1998	260241	64596	574413	40194	5649	284	101029	11721	17544	23635	203	82	353841	4738	364	10	106	5	1229	52	.	.	8047	230	1322666	145547	6.13
1999	267193	72153	601142	58240	5818	234	98604	12262	18241	24469	208	71	352817	9976	2120	87	55	2	59031	3326	.	.	14464	521	1419693	181341	6.68
2000	257636	71286	595183	59272	5913	162	97612	12125	17759	23732	185	66	347624	6474	1467	43	38	2	2959	76	.	.	15772	352	1342148	173590	6.33
2001	251441	71536	594652	58921	7846	196	102214	11985	15684	20970	193	48	354405	5391	818	50	29	14	909	43	.	.	7148	194	1335339	169348	6.28
2002	234019	66115	565961	56520	7000	189	96248	11503	17512	16140	182	46	368327	7747	2384	66	24	.	1338	48	.	.	13982	282	1306977	158656	6.12
2003	217479	60496	538175	56012	6857	190	85804	10963	16917	12999	130	28	390364	6725	2436	63	25	.	1949	84	.	.	14408	339	1274544	147899	5.94
2004	217550	60104	527733	55748	6760	196	76379	10303	16104	13118	135	22	370881	6215	98	1	17	1	1987	90	10869	473	7442	178	1235955	146449	5.77
2005	217050	58704	509012	55678	8092	380	67836	9773	15747	12450	98	15	363982	6687	843	35	1467	98	661	33	26750	1322	11472	470	1223010	145645	5.71
2006	212220	55892	497603	56848	8096	321	60418	9741	15167	11825	88	8	371148	6808	386	28	104	10	1831	85	25653	1401	12058	431	1204772	143398	5.63
2007	215858	53236	497172	57546	8319	297	53507	9062	15353	11424	68	6	367632	6538	483	21	51	5	1933	90	24360	1379	10275	342	1195011	139946	5.57
2008	222226	35100	483698	52644	7666	329	46051	8197	14688	9829	42	4	376719	5113	1260	41	31	2	1424	63	23263	1307	7808	280	1184876	112909	5.42
2009	226205	23793	495646	49209	8087	320	44990	8058	15781	6167	19	.	388030	3633	489	13	15	1	5921	180	22193	1397	24638	1310	1232014	94081	5.54
Total	4042942	987016	9227360	798120	111800	4456	1482449	184632	269687	288956	2075	863	6000327	99062	13932	621	2106	145	84641	4383	133088	7279	204476	6646	21574883	2382179	100.00

FIGURE 36 – Nombre de sinistres classés sans suite par garanties et par années de survenance

Les intitulés relatifs aux codes des garanties présents dans les tableaux sont les suivants : (A) Responsabilité Civile (matérielle et corporelle) ; (C) Dommage ; (D) Incendie ; (E) Vol ; (F) Recours ; (J) Protection Juridique ; (K) Bris de glace ; (M) Catastrophe Naturelle ; (S) Attentat ; (T) Tempête ; (V) Acte Vandalisme ; (W) Climatique.

.2 L'estimateur de la fonction de survie : une présentation heuristique

Cette annexe est inspirée du cours de modèles de survie de Gilbert Colletaz [2012].

La fonction de survie étant définie comme :

$$S(x) = \mathbb{P}(X > x) = 1 - F(x), x \geq 0$$

Soient deux montants de sinistre x_1 et x_2 telles que $x_2 > x_1$, alors :

$$\mathbb{P}(X > x_2) = \mathbb{P}(X > x_2 \text{ et } X > x_1)$$

étant donné que pour avoir un montant de sinistre supérieur à x_2, il faut que ce montant soit au moins également supérieur à x_1. En utilisant le théorème des probabilités conditionnelles nous obtenons :

$$S(x_2) = \mathbb{P}(X > x_2) = \mathbb{P}(X > x_2 | X > x_1)\mathbb{P}(X > x_1)$$

Ce qui est équivalent d'écrire : $S(x_2) = \mathbb{P}(X > x_2 | X > x_1)S(x_1)$.

De manière générale, en triant les x_i $(i = 1, ..., p)$ dans l'ordre croissant $(x_1 < ... < x_p)$, la fonction de survie peut se définir comme :

$$S(x) = \prod_{i=1}^{p} \underbrace{\mathbb{P}(X > x_i | X > x_{i-1})}_{(a)} \qquad \text{où} \qquad x_0 = 0 < x_1 < ... < x_p = x$$

où (a) peut être estimé par $1 - \frac{d_{x_i}}{n_{x_i}}$ avec d_{x_i} le nombre de sinistres supérieurs à x_i et n_{x_i} le nombre de sinistres compris dans l'intervalle $]x_{i-1}\ ; x_i]$.

Nous avons donc :

$$\hat{S}(x) = \prod_{i|x_i \leq x}^{p} (1 - \hat{q}_i) \quad \text{avec} \quad \hat{q}_i = \frac{d_i}{n_i} \tag{7}$$

Il est possible de montrer que, sous certaines conditions assez faibles, $\hat{S}(t)$, l'estimateur de Kaplan-Meier, a asymptotiquement une distribution normale centrée sur $S(t)$. Une de ces conditions est que la censure soit non informative relativement à l'évènement étudié. Cette condition peut se traduire par le fait que la probabilité de connaître l'évènement étudié à un temps t quelconque ou dans le cas étudié dans ce mémoire, pour un montant x quelconque, est la même pour tous les sinistres censurés et les sinistres non censurés.

.3 Kaplan-Meier comme estimateur du maximum de vraisemblance non paramétrique

Cette annexe est inspirée du cours de modèles de survie de Gilbert Colletaz [2012].

L'intérêt de cette approche est de délivrer une définition plus stricte et plus formelle permettant d'obtenir la dérivation de l'estimateur de Kaplan-Meier. Il s'agit de définir une fonction de vraisemblance sans faire d'hypothèse sur la distribution des temps de survie et d'aboutir à Kaplan-Meier comme étant la solution du problème de maximisation de cette

dernière fonction.

Des conditions de régularité, concernant en particulier la continuité et la différentiabilité de la fonction de survie, seront les seules hypothèses effectuées. L'intérêt de cette démarche sera notamment de pouvoir calculer la variance de l'estimateur de Kaplan-Meier. Ceci sera possible à partir de la matrice d'information de Fisher associée à cette vraisemblance.

Il est supposé que, comme c'est le cas dans nos données, nous nous situons dans un cadre de censure aléatoire à droite. Il convient donc de spécifier et de rappeler certaines hypothèses et notations. Dans le cas présent, les montants cumulés finaux des sinistres ne sont pas toujours observés. Ce que l'on observe est la réalisation de X^* définie par $X_i^* = \min(X_i, C_i)$; si $c_i < x_i$ alors le montant cumulé total du sinistre n'est pas connu et le montant x_i^* pris en compte dans les calculs pour ce sinistre est en fin de compte c_i. Au contraire, si $c_i \geq x_i$ alors $x_i^* = x_i$. On a finalement, $x_i^* = \min(x_i, c_i)$. De plus, on définit une indicatrice indiquant l'absence ($D_i = 1$ si $x_i^* = x_i$) ou la précense de censure ($D_i = 0$) :

$$D_i = \begin{cases} 1 & \text{si } X_i \leq C_i \\ 0 & \text{si } X_i > C_i \end{cases}$$

Il est également supposé que les variables X_i soient indépendantes entre elles. La même hypothèse est posée pour les variables C_i. De plus, le mécanisme de censure est supposé indépendant des coûts unitaires des sinistres : X_i et C_i sont également indépendantes. Les variables aléatoires X_i, pour i allant de 1 à n, sont supposées avoir la même distribution et sont complètement caractérisées par une fonction de densité $f(x)$ et une fonction de survie $S(x)$. De façon similaire, les variables C_i sont distribuées de la même manière, avec comme fonction de densité $k(x)$ et une fonction de survie $K(x)$.

L'objectif est d'estimer les caractéristiques de la distribution des coûts unitaires des sinistres et particulièrement la fonction de survie $S(x)$. Il s'agit donc de caractériser la vraisemblance correspondant à la précédente configuration en écrivant la vraisemblance relative à un sinistre selon qu'il soit censuré ou non :

- Dans le cas d'un sinistre non censuré : la vraisemblance est donnée par la probabilité de survenue de l'évènement (la clôture du sinistre) pour un montant x, soit :

$$\mathbb{P}[X \geq x] - \mathbb{P}[X > x] = S(x^-) - S(x), \text{ avec } S(x^-) = \lim_{dx \to 0^-} S(x + dx). \tag{8}$$

Cette vraisemblance est encore égale à $f(x)$ (plus précisément, à $f(x)dx$ car X est continue).

- Dans le cas d'un sinistre censuré : la vraisemblance correspond tout simplement à $S(x) = \mathbb{P}[X > x]$, la probabilité que le coût unitaire d'un sinistre soit supérieur au montant des règlements cumulés déjà effectués à la d'observation.

Soit d_i, le nombre de sinistres qui connaissent l'évènement étudié (la clôture) pour les montants x_i, pour i allant de 1 à k et m_i le nombre de sinistres censurés sur l'intervalle $[x_i, x_{i+1}[$ à des montants $x_{i,1}, x_{i,2}, ..., x_{i,m_i}$. Cette situation peut être représentée sur l'échelle des montants de la façon suivante :

$$\boxed{0} < \underbrace{x_{0,1} \leq x_{0,2} \leq ... \leq x_{0,m_0}}_{m_0 \text{ sinistres censurés}} < \overbrace{\boxed{x_1}}^{d_1 \text{ non censurés}} \leq \underbrace{x_{1,1} \leq ... \leq x_{1,m_1}}_{m_1 \text{ sinistres censurés}} < \overbrace{\boxed{x_2}}^{d_2 \text{ non censurés}} \leq ... < \overbrace{\boxed{x_k}}^{d_k \text{ non censurés}} \leq \underbrace{x_{k,1} \leq ... \leq x_{k,m_k}}_{m_k \text{ sinistres censurés}} \tag{9}$$

La fonction de vraisemblance s'écrit finalement :

$$\begin{aligned} L &= S(x_{0,1})S(x_{0,2})...S(x_{0,m_0}) \times \prod_{i=1}^{d_1}[S(x_1^-) - S(x_1)] \\ &\times S(x_{1,1})S(x_{1,2})...S(x_{1,m_1}) \times \prod_{i=1}^{d_2}[S(x_2^-) - S(x_2)] \\ &\times ... \\ &\times \prod_{i=1}^{d_k}[S(x_k^-) - S(x_k)] \times S(x_{k,1})S(x_{k,2})...S(x_{k,m_k}) \\ &= \prod_{i=1}^{m_0} S(x_{0,i}) \prod_{i=1}^{k} \left([S(x_i^-) - S(x_i)]^{d_i} \prod_{j=1}^{m_i} S(x_{i,j}) \right) \end{aligned} \tag{10}$$

L'objectif étant de maximiser cette vraisemblance, deux points doivent être soulignés :

- Si on pose $S(x_i^-) = S(x_i)$, alors la vraisemblance s'annule. La fonction de survie doit donc effectuer un saut en $S(x_i)$. De plus, l'objectif étant de maximiser la vraisemblance, $S(x_i^-)$ doit être strictement supérieure à $S(x_i)$. Or, la fonction de survie $S()$ est une fonction monotone non croissante. Nous obtenons donc l'écart maximal en posant $S(x_i^-) = S(x_{i-1})$.

- Les termes du type $S(x_{i,j})$, pour i allant de 0 à k et pour j allant de 1 à m_i sont relatifs aux sinistres pour lesquels le coût unitaire est censuré. Autrement dit, on ne connait pas le coût unitaire final de ces sinistres car leur date de clôture est postérieure à la date d'observation. Afin de maximiser la vraisemblance et puisque la fonction de survie est monotone non croissante, il est nécessaire d'attribuer à ces termes la plus grande valeur possible. Cette dernière valeur correspond à celle prise par la fonction de survie sur le montant de sinistre qui lui est immédiatement inférieur. On a donc $S(x_{i,j}) = S(x_i)$ avec en particulier $S(x_{0,j}) = S(t_0) = S(0) = 1$.

La fonction de survie $S()$ est ainsi une fonction étagée comportant des sauts pour les montants des sinistres non censurés. D'après les points mis en évidence précédemment, la vraisemblance peut encore s'écrire de cette façon :

$$L = \prod_{i=1}^{k}[S(x_{i-1}) - S(x_i)]^{d_i} S(x_i)^{m_i} \tag{11}$$

D'après le théorème des probabilités conditionnelles, on a :

$$\begin{aligned} S(x_i) &= \mathbb{P}[X > x_i] \\ &= \mathbb{P}[X > x_i, X > x_{i-1}] \\ &= \mathbb{P}[X > x_i | X > x_{i-1}] \times \mathbb{P}[X > x_{i-1}] \\ &= \mathbb{P}[X > x_i | X > x_{i-1}] \times S(x_{i-1}) \end{aligned} \tag{12}$$

De plus, on a $S(x_0) = S(0) = 1$, ce qui va permettre de simplifier l'expression ci-dessus. Il en découle que $S(x_1) = \pi_1, S(x_2) = \pi_1\pi_2, S(x_i) = \pi_1\pi_2...\pi_i$, avec π_i la probabilité conditionnelle que X soit supérieure à x_i sachant que X est supérieure à x_{i-1}.

La fonction de vraisemblance peut donc encore s'écrire en fonction des termes π_i :

$$\begin{aligned} L &= \prod_{i=1}^{k} (\pi_1 \pi_2 ... \pi_{i-1})^{d_i} (1-\pi_i)^{d_i} (\pi_1 \pi_2 ... \pi_i)^{m_i} \\ &= \prod_{i=1}^{k} (1-\pi_i)^{d_i} \pi_i^{m_i} (\pi_1 \pi_2 ... \pi_{i-1})^{d_i+m_i} \end{aligned} \tag{13}$$

Enfin, en définissant n_i comme étant le nombre de sinistres non clos à des montants x_i (c'est à dire les sinistres ayant un montant cumulé supérieur à x_i, soit $n_i = \sum_{j \geq i}(d_j + m_j)$), il en découle cette expression pour la fonction de vraisemblance :

$$L = \prod_{i=1}^{k} (1-\pi_i)^{d_i} (\pi_i)^{n_i - d_i} \tag{14}$$

et celle-ci pour la log-vraisemblance :

$$l = \ln L = \sum_{i=1}^{k} d_i \ln(1-\pi_i) + (n_i - d_i) \ln(\pi_i) \tag{15}$$

En déterminant les estimateurs du maximum de vraisemblance $\hat{\pi}_i$, il en découlera ceux de $S(x_i)$ d'après les égalités précédentes ainsi que la propriété d'invariance aux transformations de ces estimateurs. Pour ce faire, les solutions sont obtenues à partir des conditions du premier ordre :

$$\begin{aligned} &\frac{\partial l()}{\partial \pi_i} = 0 \\ \Leftrightarrow &-\frac{d_i}{1-\pi_i} + \frac{n_i - d_i}{\pi_i} = 0 \\ \Leftrightarrow &\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}, \quad i = 1, ..., k \end{aligned} \tag{16}$$

L'estimateur de Kaplan-Meier est finalement donné par l'expression suivante :

$$\hat{S}(x) = \prod_{i | x_i \leq x} \left(1 - \frac{d_i}{n_i}\right) \tag{17}$$

.4 La variance de l'estimateur de Kaplan-Meier

Cette annexe est inspirée du cours de modèles de survie de Gilbert Colletaz [2012].

Comme énoncé dans le corps de ce mémoire, si l'on souhaite apprécier la précision de l'estimation effectuée de la fonction de survie $S(x)$, il est indispensable d'estimer la variance de l'estimateur $\hat{S}(x)$. Pour ce faire, il est possible d'utiliser des résultats dérivés dans le cadre de la théorie de l'estimation par maximum de vraisemblance. En effet, la variance asymptotique de $\hat{\theta}_{MV} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)'$ peut être évaluée par :

$$\hat{V}(\hat{\theta}_{MV}) = -\left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} |_{\theta = \hat{\theta}_{MV}} \right)^{-1}_{i,j=1,...,k} \tag{18}$$

L'idée est ensuite d'obtenir la formule dite de Greenwood couramment utilisée dans la littérature. Pour ce faire, il s'agit de partir de la formule de calcul de l'estimateur Kaplan-Meier précédemment démontrée :

$$\begin{aligned}\hat{S}(x) &= \prod_{i|x_i\leq x}\left(1-\frac{d_i}{n_i}\right) = \prod_{i|x_i\leq x}\hat{\pi}_i \\ \Leftrightarrow \ln\left(\hat{S}(x)\right) &= \sum_{i|x_i\leq x}\ln\left(1-\frac{d_i}{n_i}\right) = \sum_{i|x_i\leq x}\ln\left(\hat{\pi}_i\right)\end{aligned} \tag{19}$$

avec d_i qui vaut 1 lorsque le sinistre est clos et 0 lorsque le sinistre est encore en cours (censure) et n_i le nombre de sinistres non clos à des montants x_i, c'est à dire les sinistres ayant un montant cumulé supérieur à x_i.

D'après la section précédente nous avons :

$$\begin{cases}\dfrac{\partial^2 l()}{\partial \pi_i^2} = -\dfrac{n_i-d_i}{\pi_i^2} - \dfrac{d_i}{(1-\pi_i)^2}, \quad i=1,...,k \\ \dfrac{\partial^2 l()}{\partial \pi_i \partial \pi_j} = 0, \quad i,j=1,...,k \text{ et } i\neq j\end{cases} \tag{20}$$

La matrice de Variance-Covariance est donc une matrice diagonale et son évaluation en $\hat{\pi}_i = 1-d_i/n_i$ aboutit aux résultats suivants :

$$\begin{cases}\mathrm{Var}(\hat{\pi}_i) = \mathrm{Cov}(\hat{\pi}_i,\hat{\pi}_i) = \dfrac{d_i(n_i-d_i)}{n_i^3}, \quad i=1,...,k \\ \mathrm{Cov}(\hat{\pi}_i,\hat{\pi}_j) = 0, \quad i,j=1,...,k \text{ et } i\neq j\end{cases} \tag{21}$$

En appliquant la méthode delta :

$$\begin{aligned}\mathrm{Var}[\ln(\hat{\pi}_i)] &= \hat{\pi}_i^{-2}\mathrm{Var}[\hat{\pi}_i] \\ &= \frac{n_i^2}{(n_i-d_i)^2}\times\frac{d_i(n_i-d_i)}{n_i^3} \\ &= \frac{d_i}{(n_i-d_i)n_i}\end{aligned} \tag{22}$$

Ainsi, on obtient :

$$\mathrm{Var}[\ln(\hat{S}(x))] = \hat{\sigma}^2_{lS_t} = \sum_{i|x_i\leq x}\frac{d_i}{(n_i-d_i)n_i} \tag{23}$$

On obtient la variance de l'estimateur de Kaplan-Meier :

$$\begin{aligned}\mathrm{Var}[\hat{S}(x)] &= \mathrm{Var}[\exp\{\ln(\hat{S}(x))\}] \\ &= [\exp\{\ln(\hat{S}(x))\}]^2 \quad \mathrm{Var}[\ln(\hat{S}(x))] \\ &= \hat{S}(x)^2 \times \sum_{i|x_i\leq x}\frac{d_i}{(n_i-d_i)n_i} \\ &= \hat{S}(x)^2 \times \hat{\sigma}^2_{lS_t}\end{aligned} \tag{24}$$

.5 Intervalle de Confiance sur la fonction de survie

Cette annexe est inspirée du cours de modèles de survie de Gilbert Colletaz [2012].

Construction d'un Intervalle de Confiance ponctuel :

Il s'agit de trouver deux bornes b_{L_x} et b_{U_x} pour avoir : $\mathbb{P}\left[b_{U_x} \geq S(x) \geq b_{L_x}\right] = 1-\alpha$, $\forall x > 0$, où α représente le niveau critique *a priori* fixé et utilisé pour calculer le niveau de confiance.

Tout d'abord, $\sqrt{n}\dfrac{(\hat{S}(x) - S(x))}{S(x)}$ converge vers une martingale gaussienne centrée. Ceci implique que la distribution asymptotique de $\hat{S}(x)$ est gaussienne et centrée sur $S(x)$. Comme démontré *supra*, l'écart-type estimé de la fonction de survie, noté $\hat{\sigma}_{S_x}$, est défini par :

$$\hat{\sigma}_{S_x} = \hat{\sigma}_{lS_x}\hat{S}(x) \tag{25}$$

En notant $z_{\frac{\alpha}{2}}$ le quantile d'ordre $\frac{\alpha}{2}$ de la distribution normale standardisée (loi $\mathcal{N}(0,1)$), l'intervalle de confiance de niveau $100(1-\alpha)\%$ est donné par :

$$IC_{1-\alpha}(S(x)) = \left[\hat{S}(x) - z_{\frac{\alpha}{2}}\hat{\sigma}_{S_x}\ ; \hat{S}(x) + z_{\frac{\alpha}{2}}\hat{\sigma}_{S_x}\right] \tag{26}$$

Un inconvénient de la construction de l'Intervalle de Confiance avec la formule précédente est que les bornes obtenues peuvent être en dehors de l'intervalle $[0,1]$, ce qui serait incompatible avec les valeurs que peut prendre une fonction de survie. C'est pourquoi en pratique, l'intervalle de confiance suivant sera considéré :

$$IC_{1-\alpha}(S(x)) = \left[max(\hat{S}(x) - z_{\frac{\alpha}{2}}\hat{\sigma}_{S_x}; 0)\,; min(\hat{S}(x) + z_{\frac{\alpha}{2}}\hat{\sigma}_{S_x}; 1)\right] \tag{27}$$

Construction de bandes de confiance :

L'idée est à présent de déterminer une région du plan qui englobe la fonction de survie avec une probabilité de $1-\alpha$, ou de manière équivalente, un ensemble de bornes b_{L_x} et b_{U_x} qui encadre $S(x)$ avec une probabilité $1-\alpha$, $\forall x \in [x_L, x_U]$. En pratique, ce sont les bandes de Hall-Wellner et les bandes de Nair (*equal precision bands*) les plus utilisées.

Soit x_k le montant de sinistre maximum observé dans l'échantillon, dans ce cas sont imposées les restrictions suivantes pour les bandes de Nair $0 < x_L < x_U \leq x_k$, puis celles-ci pour les bandes de Hall-Wellner $0 \leq x_L < x_U \leq x_k$. A noter que pour les bandes de Hall-Wellner, la nullité de x_L est permise.

Il n'est pas évident de mettre en avant l'utilité pratique de ces bandes par rapport à l'utilisation plus classique des intervalles de confiance ponctuel, d'autant plus que l'obtention de ces dernières est plus difficile d'un point de vue mathématique[18]. De plus, du fait du caractère joint, pour un montant de sinistre x donné, l'étendue des bandes est plus large que celle de l'intervalle de confiance associé.

- **les bandes de confiance de Hall-Wellner**

Afin d'établir une bande de confiance, Hall-Wellner se place sous l'hypothèse de continuité des fonctions de survie $S(x)$ et $K(x)$ relatives respectivement aux variables X_i et C_i définies

18. Il s'agit d'utiliser le fait que $\sqrt{n}\dfrac{(\hat{S}(x) - S(x))}{S(x)}$ converge vers une martingale gaussienne centrée. Le résultat s'obtient ensuite en passant par une transformation faisant apparaître un pont brownien $\{W^0(y), y \in [0,1]\}$, pondéré par $1/\sqrt{y(1-y)}$ chez Nair, ceci permettant d'obtenir les valeurs critiques adéquates.

précédemment. De cette façon, Hall et Wellner ont démontré que pour tout $x \in [x_L, x_U]$, l'Intervalle de Confiance associé au risque de première espèce α est donné par :

$$\left[\hat{S}(x) - h_\alpha(y_L, y_U)n^{-\frac{1}{2}}[1 + n\hat{\sigma}^2_{lS_x}]\hat{S}(x)\,;\, \hat{S}(x) + h_\alpha(y_L, y_U)n^{-\frac{1}{2}}[1 + n\hat{\sigma}^2_{lS_x}]\hat{S}(x)\right], \tag{28}$$

avec y_L et y_U donnés par $y_i = \frac{n\hat{\sigma}^2_{lS_{x_i}}}{(1+n\hat{\sigma}^2_{lS_{x_i}})}$ pour $i = L, U$ et avec $h_\alpha(y_L, y_U)$ la borne vérifiant $\alpha = \mathbb{P}\left[sup_{y_L \leq y \leq y_U}|W^0(y)| > h_\alpha(y_L, y_U)\right]$.

- **les *equal precision bands* de Nair**

En ce qui concerne la construction des bandes de confiance de Nair, les bornes des Intervalles de Confiance sont modifiées du fait de l'utilisation d'un pont Brownien pondéré (comme détaillé brièvement dans la précédente note de bas de page). En effet, ces dernières sont données pour tout $t \in [x_L, x_U]$ par :

$$\left[\hat{S}(x) - e_\alpha(y_L, y_U)\hat{\sigma}_{S_x}\,;\, \hat{S}(x) + e_\alpha(y_L, y_U)\hat{\sigma}_{S_x}\right], \tag{29}$$

avec $e_\alpha(y_L, y_U)$, la borne qui vérifie cette égalité :

$$\alpha = \mathbb{P}\left[sup_{y_L \leq y \leq y_U}\frac{|W^0(y)|}{\sqrt{y(1-y)}} > e_\alpha(y_L, y_U)\right].$$

En comparant les équations (26) et (29), on remarque très bien que les bornes relatives aux bandes de Nair sont proportionnelles à celles des Intervalles de Confiance ponctuels. De plus, cela correspond simplement à un ajustement du niveau de première espèce associé (seuil de risque utilisé).

.6 Les copules elliptiques

- **La copule Gaussienne**

Tout d'abord, la copule Gaussienne n'est pas adaptée à la modélisation des valeurs extrêmes car elle ne présente pas de dépendance sur les queues de distribution. Le caractère intéressant de cette copule est qu'elle est sous-jacente à la distribution Normale multi-variée.

En effet, le fait de modéliser la structure de dépendance d'un échantillon donné par une copule Gaussienne est cohérent avec la mesure de cette dépendance par le coefficient de corrélation linéaire de Pearson. Néanmoins, dès lors que l'on est en présence d'une dépendance non linéaire ou faisant intervenir des événements extrêmes, il est primordial de privilégier d'autres copules présentées *infra*.

La fonction de distribution de la copule Gaussienne d-dimensionnelle est définie comme suit :

$$C(u_1, ..., u_d) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)), \forall (u_1, ..., u_d) \in [0, 1]^d \tag{30}$$

La fonction de distribution Φ^{-1} est l'inverse généralisée de la distribution Normale centrée réduite uni-variée. La fonction Φ_Σ est la fonction de répartition de la loi Normale centrée réduite et Σ sa matrice de variance-covariance (qui correspond à la matrice de corrélation dans ce cas).

Voici l'expression de la densité de Φ_Σ :

$$\phi_\Sigma(x) = \frac{\exp\left(-\frac{1}{2}x\Sigma^{-1}x'\right)}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}, \forall x = (x_1, ..., x_d) \in \Re^d \tag{31}$$

On peut déduire la densité de la copule Gaussienne d-variée en dérivant la formule qui définit la copule Gaussienne, elle peut s'écrire comme suit :

$$c(u_1, ..., u_d) = \frac{1}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}\beta(\Sigma^{-1} - I_d)\beta'\right) \tag{32}$$

avec I_d la matrice unité de $M_d(\Re)$ et $\beta = \left(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)\right)$.

A noter que la copule Gaussienne ne possède pas de dépendance sur les queues de distribution et donc $\lambda_U = \lambda_L = 0$.

- **La copule de Student**

Cette copule est la copule sous-jacente à une distribution de Student multi-variée. La structure de dépendance relative à la copule de Student est particulièrement utilisée en finance quantitative. En effet, on peut citer de nombreuses références dont Marshal et Zeevi [2002] [41], Breymann et al. [2003] [42] qui montrent que cette copule est plus appropriée que la copule Gaussienne pour mettre en évidence la structure de dépendance d'un ensemble d'indices financiers. La principale cause de cette observation est que la copule de Student est utilisée, en pratique, pour capter les dépendances de queues, situation souvent observée sur les marchés financiers.

La construction de la copule de Student est similaire à celle de la copule Gaussienne. Toutefois, la copule de Student est construite à partir de la distribution de Student centrée réduite. La fonction de densité de la copule de Student d-variée est définie comme suit :

$$c(u_1, ..., u_d) = \frac{f_{\nu,\Sigma}\left(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d)\right)}{\prod_{i=1}^d f_\nu(t_\nu^{-1}(u_i))}, \forall (u_1, ..., u_d) \in [0,1]^d \tag{33}$$

La fonction de distribution t_ν^{-1} est définie comme l'inverse généralisée de la distribution de Student centrée réduite uni-variée à ν degrés de liberté. La fonction $f_{\nu,\Sigma}$ est la densité de probabilité de la loi de Student centrée réduite, f_ν la densité uni-variée de la loi de Student centrée réduite (Σ=1) et Σ la matrice de corrélation.

La densité $f_{\nu,\Sigma}$ est définie comme suit :

$$f_{\nu,\Sigma} = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^d\det(\Sigma)}}\left(1 + \frac{x\Sigma^{-1}x'}{\nu}\right)^{-(\nu+d)/2}, \forall x = (x_1, ..., x_d) \in \Re^d \tag{34}$$

avec Γ la fonction Gamma.

La copule de Student possède une dépendance sur les queues de distribution symétrique. Autrement dit, les dépendances de queue inférieure et supérieure sont égales et on a :

$$\lambda_U = \lambda_L = 2\bar{t}_{\nu+1}\left(\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right) \tag{35}$$

où $\bar{t}_{\nu+1}$ représente la fonction de survie d'une Student à $\nu+1$ degrés de liberté.

En guise de référence, Demarta et McNeil [2004] [22] proposent une méthode de calibrage de la copule de Student. De nombreuses propriétés y sont également reprises.

.7 Les copules archimédiennes

- **La copule de Clayton**

La copule de Clayton permet de modéliser les dépendances extrêmes et permet ainsi de pallier aux insuffisances de la copule Gaussienne.

La copule de Clayton permet de prendre en compte les dépendances positives et a l'avantage de pouvoir représenter des risques dont la structure de dépendance est plus marquée sur la queue inférieure. Pour cette raison, cette copule est utilisée en assurance pour étudier l'impact de la survenance d'événements de faible intensité sur la dépendance entre branches de risque.

Le générateur de cette copule est défini de la façon suivante :

$$\phi(u) = \frac{u^{-\alpha} - 1}{\alpha} \text{ avec } \alpha > 0 \text{ et } u \in]0, 1]. \tag{36}$$

On peut écriture la copule de Clayton en dimension $d = 2$ comme suit :

$$C(u_1, u_2) = \left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-1/\alpha}, \forall\, \alpha \geq -1 \tag{37}$$

On déduit ainsi l'écriture de la copule de Clayton en dimension $d > 2$:

$$C(u_1, ..., u_d) = \left(u_1^{-\alpha} + ... + u_d^{-\alpha} - (d-1)\right)^{-1/\alpha}, \forall\, \alpha \geq 0 \tag{38}$$

- **La copule de Joe**

Le générateur de cette copule est défini de la façon suivante :

$$\phi(u) = -\ln\left(1 - (1-u)^{\alpha}\right) \text{ avec } \alpha \geq 1 \text{ et } u \in [0, 1]. \tag{39}$$

On peut écriture la copule de Joe en dimension $d = 2$ comme suit :

$$C(u_1, u_2) = 1 - \left((1-u_1)^{\alpha} + (1-u_2)^{\alpha} - (1-u_1)^{\alpha}(1-u_2)^{\alpha}\right)^{1/\alpha}, \forall\, \alpha \geq 1 \tag{40}$$

.8 Estimation d'une copule : les approches paramétriques

- **La méthode du Maximum de Vraisemblance (MV)**

Soit $(X_i, Y_i), i = 1, ..., n$, n réalisations indépendantes d'une distribution bivariée. On suppose que la distribution bivariée est caractérisée par deux marginales avec F et G leurs fonctions de répartition respectives, f et g leurs densités respectives et C une copule de densité c.

On suppose également que F, G, et C peuvent être approchées par une loi paramétrique et on note δ, η et α les paramètres des marginales et de la copule respectivement. Par conséquent, le vecteur de paramètres à estimer est $\theta = (\delta, \eta, \alpha)$.

En accord avec le théorème de Sklar, la densité h de la fonction de répartition bidimensionnelle H, avec F_δ et F_η leurs marginales univariées et f_δ et f_η les densités respectives, peut-être représentée comme suit :

$$h(x, y) = c_\alpha\left(F_\delta(x), G_\eta(y)\right) f_\delta(x) g_\eta(y) \tag{41}$$

avec la densité de la copule bidimensionnelle $C(u,v)$ qui est définie par $c(u,v) = \dfrac{\partial C(u,v)}{\partial u \partial v}$.

La fonction de vraisemblance peut alors s'écrire de cette façon :

$$l(\theta) = \sum_{i=1}^{n} log\left\{f_\delta(x_i)\right\} + \sum_{i=1}^{n} log\left\{g_\eta(y_i)\right\} + \sum_{i=1}^{n} log\left[c_\alpha\left\{F_\delta(x_i), G_\eta(y_i)\right\}\right] \tag{42}$$

L'estimateur du Maximum de Vraisemblance (MV) du vecteur de paramètres θ est donné par :

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Theta} l(\theta) \tag{43}$$

avec Θ l'ensemble des valeurs des paramètres possibles.

L'estimateur du maximum de vraisemblance possède de nombreuses propriétés intéressantes comme par exemple la normalité asymptotique. Cependant, la recherche d'une expression analytique de θ se caractérise en général par des calculs assez lourds, ce qui la rend difficile, voire impossible. De plus, la recherche du maximum de vraisemblance à l'aide de procédures numériques peut s'avérer limitée et compliquée car le temps de calcul augmente considérablement dès lors que la taille de l'échantillon ou le nombre de paramètres à estimer augmentent.

Joe et Xu [1996] [43] proposent de calibrer séparément les densités marginales puis de calibrer la copule. C'est l'objet du paragraphe suivant.

- **La méthode *IFM (Inference For Margins)***

Plus la dimension d du modèle multi-varié augmente, plus le nombre de paramètres à estimer par la méthode du maximum de vraisemblance augmente et plus les problèmes de résolution deviennent importants. C'est pourquoi Joe et Xu [1996] [43] proposent une méthode d'estimation en deux étapes, appelée *IFM*.

Dans un premier temps, δ et η, les paramètres relatifs aux marginales sont estimés à part en appliquant la méthode classique du maximum de vraisemblance :

$$\hat{\delta} = \arg\max_{\delta} \sum_{i=1}^{n} log\left\{f_\delta(x_i)\right\} \qquad \hat{\eta} = \arg\max_{\eta} \sum_{i=1}^{n} log\left\{g_\eta(y_i)\right\} \tag{44}$$

Dans un second temps, les paramètres de la copule sont estimés par :

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} log\left\{c_\alpha(F_{\hat{\delta}}(X_i), G_{\hat{\eta}}(Y_i))\right\} \tag{45}$$

Ainsi, la première étape consiste tout simplement à estimer les paramètres des marginales par la méthode du maximum de vraisemblance. Dans un second temps, les estimateurs du maximum de vraisemblance de δ et η sont pris en compte dans la copule. C'est cette vraisemblance totale qui est maximisée selon le paramètre de la copule α.

L'avantage de cette méthode réside dans le fait que chaque maximisation repose sur un nombre de paramètres réduit, ce qui permet de diminuer considérablement les temps de calcul. En guise de référence, cette approche, connue comme la méthode paramétrique du maximum de vraisemblance en deux étapes, a également été appliquée dans un cadre de données censurées par Shih et Louis [1995].

D'autre part, comparer directement la méthode d'estimation par maximum de vraisemblance et la méthode *IFM* est délicat du fait des temps de calcul nécessaires pour obtenir les estimations. Néanmoins, des tests comparatifs sur plusieurs modèles multi-variés sont réalisés dans Xu [1996]. Ils aboutissent à la conclusion que la méthode *IFM* est une méthode efficace et permet d'aboutir à des résultats relativement proches de ceux obtenus avec la méthode classique du maximum de vraisemblance.

.9 Estimation d'une copule : la méthode des moments

L'idée générale est simple : si une mesure d'association k peut s'écrire comme $k = h(\alpha)$ avec une fonction h appropriée, alors $\hat{\alpha} = h^{-1}(\hat{k})$ où $\hat{k}$ est un estimateur non-paramétrique de k basé sur les observations. Deux mesures de corrélation non-paramétrique standards sont, comme énoncé précédemment, le Rho de Spearman et le Tau de Kendall.

C'est une approche basée sur l'inversion de deux mesures de corrélation. Par exemple, une méthode basée sur le Tau de Kendall pour l'estimation des paramètres d'une copule est abordée dans Lindskog [2000], Lindskog et al. [2003] [23] et Cherubini et al. [2004].

De plus, pour certaines copules comme celle de Student, les expressions analytiques liant le Tau de Kendall à leurs paramètres de dépendance sont explicites. L'estimation du Tau de Kendall permet donc, dans certains cas, de calibrer ces copules. Le lecteur intéressé pourra se référer à Demarta et McNeil [2004] [22].

.10 Algorithme de Marshall-Olkin

Marshall et Olkin développent en 1988 un algorithme permettant de simuler des copules archimédienne. Cet algorithme se base sur l'inverse de la transformée de Laplace-Stieltjes du générateur ϕ de la copule. Il existe de nombreuses autres méthodes de simulation des copule archimédiennes (le lecteur intéressé pourra se référer à Marius Hofert [2008] [44]).

Le tableau suivant restitue $LS^{-1}(\phi)$, les inverses des transformées de Laplace-Stieltjes des générateurs des copules de Gumbel et Frank :

Copule	Générateur $\phi(t)$	$LS^{-1}(\phi)$
Gumbel	$(-\ln t)^{\theta}, \theta \geq 1$	$S\left(\frac{1}{\theta}, 1, \left(cos\left(\frac{\pi}{2\theta}\right)\right)^{\theta}, 0; 1\right)$
Frank	$-\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right), \theta \neq 0$	$y_k = \frac{\left(1-e^{-\theta}\right)^k}{k\theta}, k \in \mathbb{N}$

TABLE 15 – Inverses des transformées de Laplace-Stieltjes

A noter que S est la distribution de la loi stable $S(\alpha, \beta, \gamma, \delta; 1)$ où $\alpha \in (0, 2], \beta \in [-1, 1], \gamma \in [0, +\infty)$ et $\delta \in \mathbb{R}$.

L'algorithme de Marshall-Olkin se décompose finalement en trois étapes :

1. Simuler la variable V selon la loi $LS^{-1}(\phi)$
2. Simuler $X_i \approx U[0, 1], i = 1, 2$ avec X_i des variables i.i.d.
3. Retourner les valeurs (U_1, U_2) où $U_i = \phi\left(-\frac{\ln(X_i)}{V}\right), i = 1, 2$