INSTITUT DES ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : ~~Monsieur~~/Madame DERKAOUI Saïda

Titre du mémoire :

Modélisation de la sinistralité grave dans le cadre des revalorisations de primes du partenariat

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

Entreprise :

Nom : Generali

Signature :

Directeur de mémoire en entreprise :

Nom : TOGNI Jérémie

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

Derkaoui Saïda

Membres présents du jury de la filière

Remerciements

Je suis extrêmement reconnaissante envers l'Equité qui m'a permis de réaliser ce mémoire dans les meilleures conditions malgré la situation sanitaire actuelle et de m'avoir prodigué un encadrement et un suivi de qualité.

Je souhaite remercier dans un premier temps mon maître d'alternance, M. Jérémie Togni, *responsable en actuariat des solutions d'assurance automobile*, pour son encadrement, sa patience, ses conseils, sa disponibilité et son investissement. Ses recommandations m'ont permis d'effectuer un travail de qualité.

Je remercie également M. Amin Toussi, *manager du service actuariat*, qui, par sa confiance, m'a permis d'intégrer une équipe dynamique et accueillante. J'ai pu y rencontrer des collaborateurs passionnés qui m'ont guidé et m'ont conseillé. Je tiens à remercier, en particulier, M. Cyril Jamme, Mme Alice Mierzwa, Mme Marie Kathleen Hetelay, Mme Marie Christine Dejean, M. Guillaume Durand et M. Sébastien Lefevre.

Je remercie ma tutrice académique, Mme Maud Thomas, qui fût de très bon conseil. A l'écoute et disponible, ses conseils m'ont permis d'avancer dans mon travail et d'assurer une qualité rédactionnelle dans ce mémoire. De par ses enseignements, j'ai pu développer mes compétences sur le sujet et je l'admire pour ses connaissances qu'elle a pu me partager.

Je remercie également M. Pierre Thérond, pour sa gentillesse, sa disponibilité et ses conseils qui m'ont permis de découvrir plus en détail le domaine de la théorie bayésienne et son application dans divers sujets.

Ma gratitude va envers les personnes qui ont de près comme de loin permis la réalisation de ce mémoire et ont accepté de m'épauler lors de sa rédaction, notamment les personnes citées ici et plus particulièrement M. Antoine Badillet.

Enfin, je remercie mes amis, en particulier M. Kevin Mechouk ainsi que ma famille pour leur aide et leur soutien, sans qui je ne serai pas là aujourd'hui.

Résumé

Dans cette étude, nous nous plaçons dans le cadre des revalorisations de primes faites par l'Equité envers les partenaires d'assurance automobile et nous nous intéresserons, notamment, à la modélisation de la sinistralité grave.

En assurance automobile, la sinistralité grave impacte les résultats des assureurs et est dans la majorité des cas due à la garantie responsabilité civile corporelle. Cette garantie est incluse dans les garanties obligatoires que tout propriétaire d'un véhicule doit souscrire. De plus, depuis la loi Badinter de juillet 1986, les victimes ont droit à une indemnisation complète de leur préjudice sans limitation de montant. Dans le cas de victimes "légères" d'accident de la route, la convention IRCA, Indemnisation et recours corporel automobile, permet de confier l'indemnisation de la victime de la route à sa propre compagnie d'assurance et un recours est effectué en interne par la compagnie adverse. Depuis 2018, le forfait de la convention IRCA reste inchangé, de l'ordre de 1 480 euros, ou 740 euros en cas de partage des responsabilités. Cette convention ne s'applique que pour les dossiers de victimes blessées légèrement. Ainsi, elle ne concerne pas nos dossiers de sinistres graves, qui dépendent des postes de préjudice (esthétique, pretium doloris ...), de l'âge de la victime, du taux AIPP etc...

A L'Equité, un seuil de sinistres graves est appliqué pour l'ensemble des partenaires. Ces dossiers de faible fréquence annuelle représentent une perte importante pour l'assureur mais fait l'objet d'échanges lors des revalorisations tarifaires.

Dans un premier temps, nous envisageons de mettre à l'épreuve le seuil de sinistres graves déjà implémenté dans les modèles de l'équipe et donc, d'appliquer des méthodes de théorie des valeurs extrêmes, notamment les méthodes du Peak Over Threshold, qui permettent d'ajuster nos excès à une loi de Pareto Généralisée. Cette technique va nous permettre de justifier le choix ou non de ce seuil. Deux approches, liées aux données, ont été développées dans ce mémoire : l'une a été retenue pour des raisons opérationnelles mais également de robustesse pour la suite de l'étude.

Dans un deuxième temps, nous cherchons à modéliser la fréquence des sinistres graves, obtenus à partir du seuil choisi, en utilisant différentes approches actuarielles. Nous proposons d'appliquer une prédiction par processus ARIMA sur notre série sur l'ensemble du portefeuille puis d'appliquer la même procédure sur la série chronologique d'un unique partenaire. Les résultats obtenus pour le partenaire n'étant pas assez satisfaisants, nous envisageons d'appliquer des GLM sur le

nombre de sinistres graves par regroupement de partenaire, fait en amont de la modélisation. Afin de limiter le biais dû à ce regroupement, nous appliquerons des méthodes basées sur la théorie bayésienne, à la maille partenaire mais également à la maille partenaire x produits commercialisés.

Nous souhaitons ensuite réaliser une modélisation statistique de la charge grave, via des méthodes de Monte Carlo, qui nous permettra de lier les résultats des deux autres parties. Nous comparerons ces résultats à des résultats obtenus par la théorie bayésienne uniquement appliquée à la charge de sinistralité grave. L'approche retenue sera de différencier les approches selon le type de partenariat.

Mots clés : *Assurance automobile, Responsabilité corporelle, Sinistres graves, Peak over Threshold, Série temporelle, Modélisation linéaire généralisée, Bayésien, Monte Carlo.*

Abstract

In this study, we place ourselves in the context of premium revaluations made by l'Equité towards automobile insurance brokers and we will be interested, in particular, in the modeling of large loss claims.

In automobile insurance, large loss claims have an impact on insurers' results and are in most cases due to public liability coverage. This coverage is included in the mandatory coverages that every vehicle owner must take out. In addition, since the Badinter's law of July 1986, victims are entitled to full compensation for their injury without any limit on the amount. In the case of "light" road accident victims, the IRCA convention, Indemnisation et recours corporel automobile, makes it possible to entrust the compensation of the road accident victim to his own insurance company and recourse is made internally by the opposing company. As of 2018, the IRCA agreement lump sum remains unchanged at around 1,480 euros, or 740 euros in the event of shared responsibility. This convention only applies to cases of slightly injured victims. Thus, it does not concern our files of serious claims, which depend on the items of damage (aesthetic, pretium doloris ...), the age of the victim, the AIPP rate etc...

At L'Equité, a threshold of serious claims is applied for all partners. These low annual frequency cases represent a significant loss for the insurer but are the subject of exchanges during tariff revaluations.

As a first step, we plan to test the severe loss threshold already implemented in the team's models and thus, to apply extreme value theory methods, in particular Peak Over Threshold methods, which allow us to adjust our excesses to a Generalized Pareto Law. This technique will allow us to justify the choice or not of this threshold. Two approaches, linked to the data, have been developed in this thesis : one has been chosen for operational reasons but also for robustness for the rest of the study.

In a second step, we seek to model the frequency of serious claims, obtained from the chosen threshold, using different actuarial approaches. We propose to apply a prediction by ARIMA process on our series on the whole portfolio and then to apply the same procedure on the time series of a single partner. As the results obtained for the partner are not satisfactory enough, we plan to apply GLM on a number of serious claims by grouping partners, done upstream of the modeling. In order to limit the bias due to this grouping, we will apply methods based on Bayesian theory to

the partner grid but also to the partner grid x marketed products.

Then, we want to perform a statistical modeling of the severe load, using Monte Carlo methods, which will allow us to link the results of the other two parts. We will compare these results with results obtained by the Bayesian theory only applied to the severe load. The approach adopted will be to differentiate the approaches according to the type of partnership.

Key words : *Car insurance,Public liability insurance, Large loss claims, Peak over Threshold, Time series, Generalised linear modeling, Bayesian, Monte Carlo*

Table des matières

Table des figures

Introduction

L'Equité, entité filiale de Generali France au sein de laquelle ce mémoire a été réalisé, a connu une forte croissance ces dernières années, en particulier, l'activité d'assurance automobile. Cette entité propose des solutions d'assurance par l'intermédiaire de courtiers et grossistes indépendants.

La tarification en assurance automobile est un des enjeux majeurs dans le rôle des actuaires. Deux méthodes ont toujours été employées : la tarification a priori et la tarification a posteriori. Dans nos relations partenariales, les primes font l'objet de revalorisation à chaque exercice, pour chaque partenaire et pour chaque produit commercialisé. Cette tarification sera la conséquence d'une analyse minutieuse de la sinistralité, d'un point de vue global, pour tout le portefeuille, et à la maille partenaire, mais également la conséquence du niveau de rentabilité annuelle atteint en fonction des objectifs imposés par l'Equité.

Cependant, il convient de distinguer dans nos analyses la sinistralité attritionnelle, correspondant à des événements de fréquence importante mais à moindre coût, de la sinistralité grave, correspondant à des événements de faible fréquence mais aux coûts importants. Au moment des revalorisations, un pourcentage élevé de sinistralité attritionnelle est déjà connu par l'équipe, mais la présence de sinistres graves, à la maille partenaire, reste plus difficile à modéliser et nécessite, pour se faire, de méthodes de statistiques actuarielles.

La prédiction des sinistres graves est une problématique importante et fait l'objet d'échanges entre l'entité et ses partenaires. La première difficulté dans cette modélisation vient d'une instabilité temporelle sur la survenance de ces évènements, nous ne pouvons pas observer le même taux de sinistralité grave chaque année d'exercice pour un partenaire. La deuxième difficulté réside dans le type de modélisation à appliquer et à quelle échelle l'appliquer. Enfin, la dernière difficulté serait l'intégration d'un processus de modélisation simple d'un point de vue opérationnel, avec des résultats robustes et qui prennent en considération un avis d'expert, lié, par exemple, à la connaissance du marché actuel, à l'impact de l'environnement économique sur la sinistralité observée, à la connaissance du partenaire et à la relation partenariale avec l'entité.

La présence de sinistralité grave dans les résultats du partenaire est dans la majorité des cas la conséquence d'un dossier qui fait l'objet d'une garantie en responsabilité civile corporelle. En effet, depuis la loi Badinter de juillet 1986, l'indemnisation des victimes est totale et elle ne fait l'objet d'aucune limitation comme nous le constatons sur d'autres garanties. Cela a pour conséquence,

une charge de sinistralité totalement prise en charge par l'assurance qui peut parfois atteindre le million d'euros, dans les cas les plus graves. De plus, le temps entre la survenance de ces sinistres et la date de consolidation peut prendre plusieurs années, ainsi, le sinistre sera présent dans notre base comme ouvert mais peut ne pas être considéré comme grave.
L'objectif de notre étude est de proposer un modèle de prédiction des sinistres graves à la maille partenaire et de pouvoir répondre à toutes les difficultés que l'on peut observer sur le sujet des sinistres graves. Les données étudiées dans le cadre de ce mémoire sont issues des produits d'assurance "aggravé", " deux roues" et "standard" de différents courtiers associés à la Direction des Partenariats de Generali. L'Équite collabore avec un panel important d'intermédiaire en assurance. Ces intermédiaires peuvent proposer un type de distribution différent, parfois innovant. Le chiffre d'affaire annuel du partenaire caractérise celui-ci, et il peut être distinct d'un partenaire à un autre mais aussi d'un produit à un autre, commercialisé par ce dernier.

Dans un premier temps, nous commencerons, dans ce mémoire, par présenter le contexte de l'étude en nous intéressant au marché de l'assurance automobile en France, à la sinistralité corporelle et sa gestion au sein de l'Équité. Nous donnerons des chiffres clés relatifs à la sinistralité corporelle et nous analyserons la base d'étude utilisée.

Dans un second temps, nous nous consacrerons à la théorie des valeurs extrêmes dont nous rappellerons quelques principes clés. Nous appliquerons la méthode Peak-Over-Threshold et les méthodes de détermination de seuil graphique à notre étude.

Le troisième chapitre sera consacré à la modélisation de la fréquence de nos sinistres graves à partir des éléments obtenus sur le seuil au chapitre deux. Enfin le quatrième chapitre concernera la modélisation du coût moyen de la sinistralité grave où nous appliquerons des méthodes de modélisation Fréquence x Coût que l'on comparera à une méthode de modélisation de la charge par l'approche bayésienne. On appliquera notre étude opérationnellement dans le cadre des revalorisations de primes réalisées chaque année avec le partenaire.

Chapitre 1

Le contexte

1.1 L'Équité, Direction des Partenariats

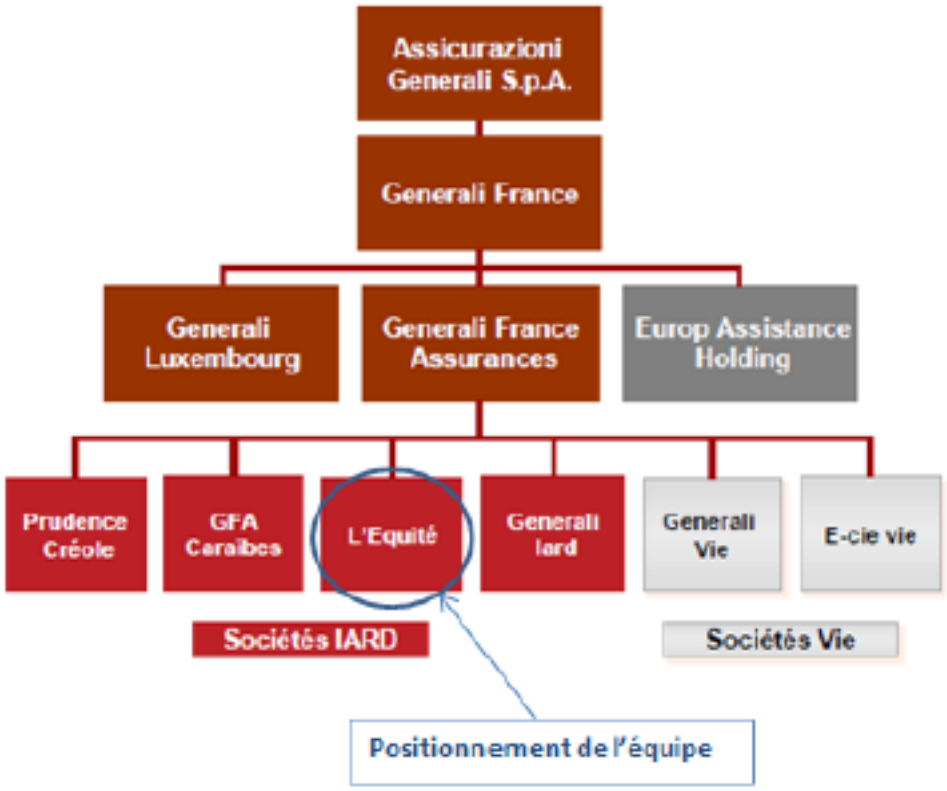


FIGURE 1.1 – Structure de Generali

L'Équité est une entité filiale de Generali France, totalement intégrée à Generali. On la considère comme la Direction des Parapactenariats. Elle s'occupe également de segments de clientèles peu accessibles aux réseaux traditionnels et intervient en complément de ceux-ci.

La Direction des Partenariats regroupe plusieurs pôles dont :

- les solutions d'assurance dommages et des solutions d'assurance de personnes ;
- les opérations d'assurance et de pilotages ;
- le contrôle interne et contrôle des délégations ;
- le service lié au développement.

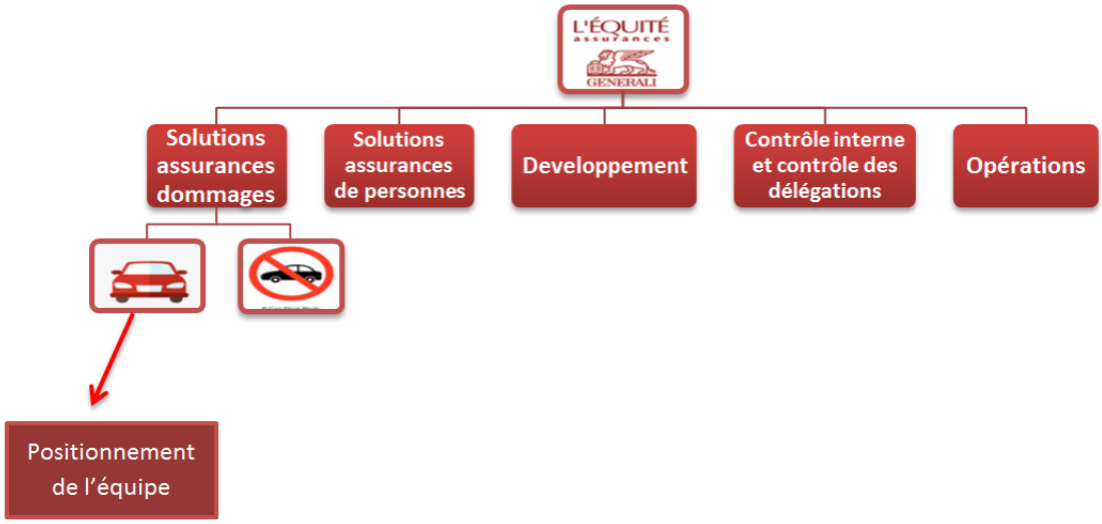


FIGURE 1.2 – Structure de l'Equité

Elle crée et propose en marque blanche, des solutions d'assurance sur mesure qui tiennent compte des spécificités de ses partenaires et intervient sur tout type de segments de marché :

- Assurance dommages de particuliers ;
- Assurance prévoyance accident et santé ;
- Protection juridique.

Le chiffre d'affaire de L'Équité en 2019 avoisine le milliard d'euros, dont environ 390 M d'euros pour l'assurance automobile. Ses partenaires sont :

- des professionnels de la distribution d'assurance ;
- des mutuelles et institutions de prévoyance ;
- des grands comptes et enseignes agissant pour le compte de leurs clients ;
- des enseignes de grande distribution et des commerces organisés ;
- des banques et organismes financiers.

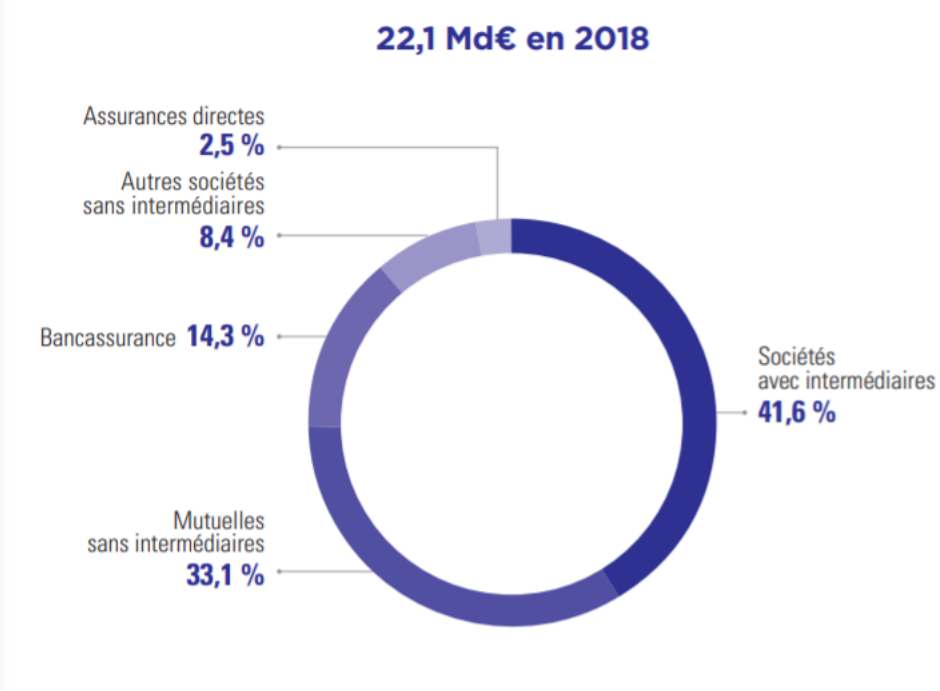
1.2 Le marché de l'assurance automobile

Contexte marché

Le marché de l'assurance en France est composé de :

- Société à intermédiaire, entreprise d'assurance qui utilise principalement des cabinets de courtages et un réseau d'agents généraux pour la distribution de leurs produits d'assurance, détenant 41,6% des parts de marché ;
- Société sans intermédiaire, entreprise d'assurance distribuant leurs produits via un réseau de salariés.

Le chiffre d'affaires de l'assurance automobile, s'établie à 22,1 Md€ en 2018 et représente 39 % de l'ensemble des cotisations des assurances de dommage aux biens et de responsabilité civile.



Chiffres clés

Le marché des véhicules de 1ère et 3ème catégorie

L'automobile des particuliers, avec un chiffre d'affaires de 19,8 Md€ en 2018 se ventile de la façon suivante :

- Les véhicules de 1ère catégorie, par définition les véhicules 4 roues à moteur dont le poids total autorisé en charge (P.T.A.C) est inférieur ou égale à 3,5 tonnes, représentent 17,6 Md€ de cotisations.

— Les véhicules de 3ème catégorie, par définition les véhicules 2 roues, tricycles et quadricycles à moteur, voiturettes et voiturettes électriques, voitures sans permis, représentent 1 Md€ de cotisations.

Ci-dessous, les chiffres de la FFA sur le marché de l'assurance automobile des particuliers en 2018 :

	2018	Variation 2018/2017	Variation 2017/2016
Véhicules de première catégorie	42 562	+0.9%	+1.4%
Dont véhicules assurés en métropoles	41 412	+0.9%	+1.4%
Voitures particulières	37 047	+0.9%	+1.4%
Voitures utilitaires	3 513	+0.7%	+1.2%
Véhicules de 3ème catégorie	4 006	+1.4%	+0.3%
Autres véhicules assurés en mono contrats	2 905	+1.9%	+1.8%

TABLE 1.1 – Nombre de véhicules assurés au 31 décembre 2018

Garantie	Prime moyenne 4 roues	Répartition des cotisations	Taux d'inclusion des garanties	Prime moyenne 2 roues	Répartition des cotisations	Taux d'inclusion des garanties
RC	147€	36%	100.0%	86€	36%	100.0%
Tous risque	496€	100.0%	-	347€	100%	-

TABLE 1.2 – Décomposition de la prime moyenne, focus sur la RC

Sinistralité	1ère catégorie				2 roues			
Garantie	Fréq. 2018	var. 2018/2017	coût moyen 2018	var. 2018/2017	Fréq. 2018	var. 2018/2017	coût moyen 2018	var. 2018/2017
RC	35.4%	-4.3%	///	///	10.7%	-6.2%	///	///
Dont RC corporels	3.4%	-5.7%	///	///	2.6%	-6.4%	///	///
Dont RC matériels	32.0%	-4.2%	1 480€	+2.5%	8.1%	-6.2%	1 285€	+5.8%

TABLE 1.3 – Fréquence de sinistralité en fonction de la garantie, focus sur la RC

On observe une amélioration de la fréquence RC corporels et matériels entre 2017 et 2018.

Sinistres corporels

Notre étude portera sur le marché des particuliers, et plus particulièrement sur les solutions d'assurance destinées aux deux segments, véhicules de 1ère et 3ème catégories.

En 2018, les sinistres corporels indemnisés au titre de la responsabilité civile ont enregistré une amélioration de leurs fréquences de -5,7 % par rapport à 2017.

Année	Variation
2014	+2.2%
2015	- 0.6%
2016	+2.7%
2017	- 2.9%
2018	- 5.7%
Niveau 2018	3.4%

–

tab :lab11

De plus, le nombre d'accidents corporels recensés par les pouvoirs publics est en baisse de 4,9%.

Bilan de l'année 2018	Accidents corporels
Année 2018	55 766
Année 2017	58 613
Différence 2018 / 2017	-2 847
Evolution 2018 / 2017	-4.9%

FIGURE 1.3 – Accidents corporels - Bilan 2018 - *ONSIR*

1.3 Présentation de notre étude et de la problématique

Les sinistres corporels

La garantie RC Automobile corporelle

Facultative à ses débuts en Juillet 1930, dans le cadre du code de l'assurance, la garantie RC Automobile devient obligatoire pour faire face à la hausse du nombre de sinistres et de leurs coûts en février 1958. Il est stipulé que :

> *Art. L211-1. "Toute personne physique ou toute personne morale autre que l'Etat, dont la responsabilité civile peut être engagée en raison de dommages subis par des tiers résultant d'atteintes aux personnes ou aux biens dans la réalisation desquels un véhicule*

est impliqué, doit, pour faire circuler celui-ci, être couverte par une assurance garantissant cette responsabilité, dans les conditions fixées par décret en Conseil d'Etat."

Notons qu'à l'heure actuelle, le conducteur n'est pas considéré comme tiers. Pour être indemnisé de son préjudice corporel en cas de sinistre responsable ou partiellement responsable, il doit souscrire à une garantie, facultative, la garantie protection du conducteur. Cependant, les passagers d'un véhicule, ainsi que les élèves d'un établissement d'enseignement de la conduite, sont considérés comme des tiers.

Gestion des sinistres corporels

La gestion d'un sinistre corporel se décompose en plusieurs étapes clés :

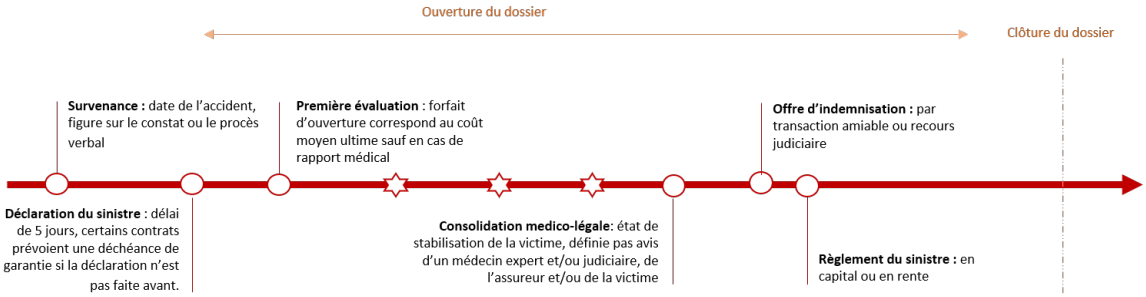


FIGURE 1.4 – Les étapes clés de la gestion d'un sinistre corporel

Les étoiles correspondent aux expertises médicales ou complémentaires. De plus la première évaluation peut être revue à la suite d'une évolution ou de nouvelles informations. Un dossier clôturé peut être réouvert en cas de :

1. changement médical, aggravation ou amélioration de l'état de santé de la victime en lien avec l'accident
2. changement économique (ex : coût horaire de tierce personne)
3. changement situationnel, la victime nécessite une adaptation à une évolution de la situation (ex : décès d'un parent qui l'assiste).

Indemnisation des sinistres corporels

Le taux d'AIPP, Atteinte à l'Intégrité Physique et Psychique, détermine la gravité des séquelles liées à un sinistre corporel. Déterminé par un médecin expert, il mesure la réduction des capacités physiques ou intellectuelles d'une victime d'un accident de la route.

Durée de règlement	Blessés avec AIPP	
	Transaction	Décisions judiciaires
Moins de 6 mois	0.7%	///
6 mois à 1 an	17.9%	0.9%
1 à 2 ans	54.0%	8.7%
2 à 3 ans	16.3%	14.4%
% 3 à 4 ans	4.7%	17.0%
4 à 5 ans	2.4%	15.3%
5 ans et plus	3.9%	43.7%
Ensemble	100.0%	100.0%

La loi Badinter du 5 juillet 1985 élargit le périmètre de la garantie responsabilité civile, et vise à l'amélioration de la situation des victimes d'accidents de la circulation et à l'accélération des procédures d'indemnisations.
La garantie de Generali stipule que :

> *La Compagnie garantit l'Assuré contre les conséquences pécuniaires de la responsabilité civile que celui-ci peut encourir en raison de dommages corporels ou matériels subis par des tiers et dans la réalisation desquels le véhicule assuré est impliqué [...] La Compagnie garantit les frais de défense civile et pénale de l'Assuré dans toute procédure administrative ou judiciaire, pour les intérêts propres de l'Assuré, lorsque la procédure concerne en même temps les intérêts de la Compagnie et ce, pour les risques de responsabilité civile visés au présent article. Cette garantie comprend les frais et honoraires d'enquête, d'instruction, d'expertise, d'avocat ainsi que les frais de procès.*

Les différents produits des solutions automobile

Au sein de l'Equité, il existe différents produits automobiles :

1. le **produit auto standard**, il s'agit d'un produit d'assurance classique, on le retrouve en grande majorité sur le marché automobile avec distribution directe.
2. le **produit deux roues**, il couvre tous les véhicules à moteur deux roues, cyclomoteurs etc...
3. le produit des véhicules haut de gamme et celui des véhicules de collection
4. le **produit auto malussée**, il s'adresse à toutes les personnes ne rentrant pas dans le cadre de souscription des produits d'assurance standard. Il s'agit de produits adaptés aux situations suivantes non-paiement de prime, fausse déclaration, annulation ou suspension de permis etc...

Dans notre étude, seuls les produits surlignés en gras nous intéresseront.

Les différents partenaires

Les contrats d'assurance peuvent être commercialisés par différents intermédiaires d'assurances et, en particulier, nous nous intéresserons aux courtiers. Le statut du courtier est celui d'un

Commerçant inscrit au registre du commerce et des sociétés [qui représente ses] clients, pour le compte desquels [il recherche], auprès des sociétés d'assurances, les garanties adaptées à leurs besoins et [négocie] les conditions de tarif en faisant jouer la concurrence.

Un courtier d'assurance peut commercialiser ses produits directement à ses clients, ou concevoir des produits d'assurances qu'il va ensuite proposer à des réseaux de distributeurs, on parle alors de **courtier grossiste**.

Aujourd'hui, et avec la transformation digitale qui impacte fortement le marché très concurrentiel de l'assurance, nombre de courtiers voient le jour en proposant une stratégie digitale innovante qui révolutionne la traditionnelle distribution en agence (des courtiers totalement digitalisés, proposant des contrats d'assurance en "un clic" ou vendant leurs produits sur des applications smartphones).

Par conséquent, la stratégie d'accompagnement se différencie en fonction du type de partenaire ainsi que de sa position sur le marché. Ces informations sont à prendre en compte dans les revalorisations de primes mais nécessitent un avis d'expert et une connaissance accrue du partenariat que nous n'allons pas développer dans ce mémoire.

Les enjeux du mémoire

La garantie RC corporelle est différente des autres garanties puisqu'elle n'est pas liée aux dommages du véhicule assuré mais aux préjudices causés par ce dernier à des personnes tierces. De plus, il s'agit d'une garantie à déroulement long, certains sinistres ouverts cinq ans plus tôt, ne sont pas encore clos.

Dans le cadre des revalorisations de primes faites tous les ans entre les partenaires et l'Equité, la modélisation des sinistres graves à l'ultime a un impact sur les négociations. Actuellement, la méthode pour déterminer le coût des sinistres graves du portefeuille pour une année d'exercice est d'appliquer environ k points au ratio S/P. On attribue alors $k\%$ de primes acquises dans l'exercice à la charge grave à l'ultime.

Dans notre étude, le coût individuel des sinistres à modéliser dans le portefeuille est net de recours et net de l'impact de la réassurance. L'objectif de la modélisation à l'ultime est de ne pas surestimer le coût des sinistres, qui aurait un impact à la hausse sur les primes renégociées, et de ne pas les sous-estimer au risque de donner un ratio S/P à l'ultime à la baisse pour le partenaire, et donc un niveau de rentabilité élevé.

Ce mémoire a pour ambition de proposer des méthodes actuarielles pour modéliser la survenance des sinistres graves pour chacun des partenaires mais également de déterminer, grâce à cette modélisation, un ratio de sinistralité à l'ultime pour l'exercice en cours. Cette modélisation devra prendre en compte l'hétérogénéité du portefeuille de partenaires.

1.4 Les données, retraitement pour le mémoire

Les bases de données disponibles

Dans cette partie, nous allons créer notre base d'étude en retraitant les bases sinistres, les bases contenant les informations sur les partenaires et les bases des primes. Concernant les traitements de ces bases, nous allons utiliser les logiciels SAS et Excel.

Traitement des bases

La base de sinistre en RC corporelle automobile contient 214 541 sinistres entre 2004 et 2018 en France métropolitaine, en retirant les anomalies dues à une déclaration du sinistre antérieure à sa survenance.

Les sinistres à coûts négatifs ou nuls

Le phénomène des sinistres à coûts nuls et négatifs peut être expliqué par :

- la présence de sinistre en état **sans suite**, dans le cas d'une erreur de déclaration du sinistre ;
- le cas d'un recours en cas de non responsabilité de l'assuré ou d'une responsabilité partielle ;
- des anomalies de règlement de la charge corporelle et matérielle ;
- l'attente des justificatifs de l'assuré ou de la compagnie adverse. L'état du sinistre est considéré en cours.

Nous enlèverons les sinistres dans les cas suivants :

- un règlement corporel négatif, peu importe le montant du coût du sinistre ;
- un montant de sinistre nul ou négatif dont le dossier est classé sans suite ;
- les sinistres annulés (le montant étant nul).

Mauvaise désignation du partenaire

Certains sinistres sont rattachés à aucun partenaire, il a fallu les supprimer de la base, ainsi que les sinistres dits fictifs, c'est-à-dire que le **mouvement de fonds vient d'une saisie manuelle de l'indemnisation**.

Mise en as-if

L'utilisation des données historiques repose sur le fait que le passé est représentatif du futur. La revalorisation des données représente une étape importante avant toute modélisation. Elle consiste à retraiter les charges des sinistres et les primes afin de prendre conscience de l'environnement socio-économique dans lequel on se place. Cette mise en as-if des données nécessite le passage par une revalorisation selon l'indice correspondant à l'inflation en France. En prenant une année de référence N, l'inflation est estimée par un coefficient α. On obtient :

$$\forall n \leq N, C_N^n = C_n^n * (1+\alpha)^{(N-n)}$$

$$\forall n \leq N, P_N^n = P_n^n * (1+\alpha)^{(N-n)}$$

Cependant, d'autres facteurs sont à prendre en compte dans l'indexation des sinistres en RC corporel dont l'augmentation de l'espérance de vie, l'amélioration et l'augmentation des coûts des techniques médicales, l'évolution de la jurisprudence...
Prenons l'exemple de l'indice consommation de soins et de bien médicaux en France :

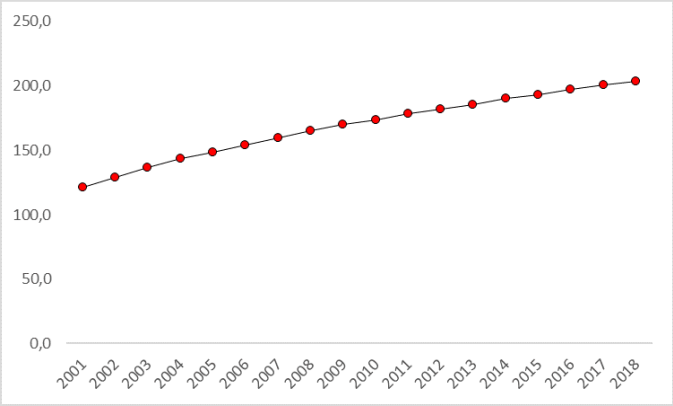


FIGURE 1.5 – Indice à la consommation de soins et de bien médicaux, *source : INSEE*

On observe une tendance à la hausse assez importante et qu'il faudrait prendre en compte dans l'indexation des sinistres.

Visualisation de la sinistralité

Dans cette partie, on s'intéresse à la sinistralité observée de la garantie RC corporelle pour le produit automobile et moto entre 2004 et 2018. Dans la distribution ci-dessous, on représente la distribution du logarithme de coût individuel positif des sinistres en RC corporelle. La médiane semble osciller d'une année à l'autre mais de façon très faible et semble très basse chaque année du fait d'un nombre très important de sinistres attritionnelles. Ces box-plots illustrent des données très asymétriques, la majorité de nos sinistres sont de faibles coûts, de plus, ce comportement reste relativement stable chaque année. Cependant la dispersion de nos données semble instable. Cette instabilité semble due à la présence de nos sinistres extrêmes.

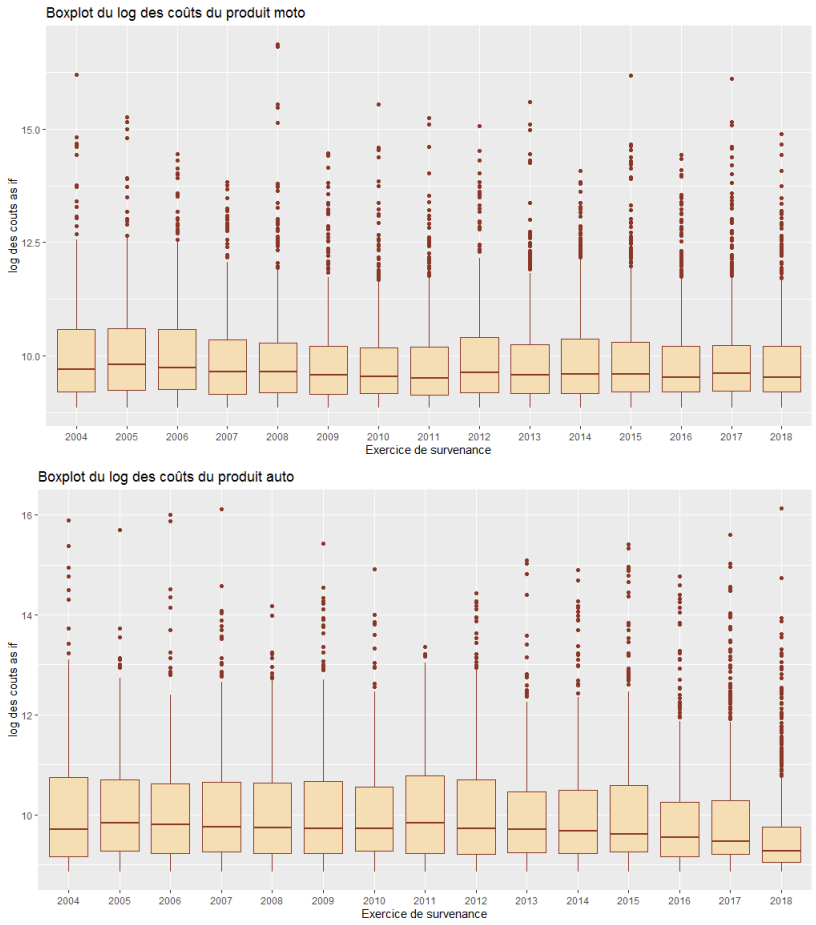


FIGURE 1.6 – Box plot des logarithmes de coûts

Nous remarquons pour le produit auto que la distribution a une tendance baissière à partir de 2015, avec une accentuation de la baisse pour la dernière année. La RC corporelle étant une branche longue, la consolidation des dommages corporels peut prendre des années. Nous avons, par conséquent, peu de vision sur ces sinistres à l'ultime.

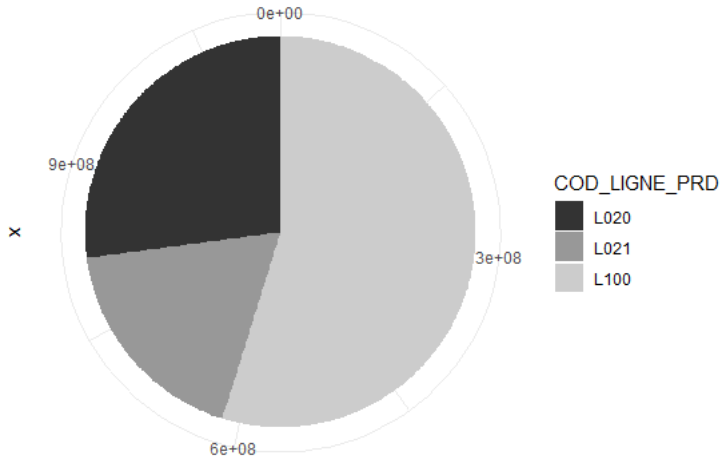


FIGURE 1.7 – Répartition de la sinistralité par produit
Dans la légende, on appelle L020 le produit auto standard, L021 le produit auto malussée et L100 le produit deux roues.

Ici, nous représentons la répartition de notre sinistralité par produit, nous pouvons voir que le poids de la sinistralité des deux roues est très significatif et en lien avec l'activité de l'Equité.
Par la suite, on choisit de représenter géographiquement la sinistralité globale en distinguant la fréquence de sinistre du coût moyen entre 2010 et 2018.

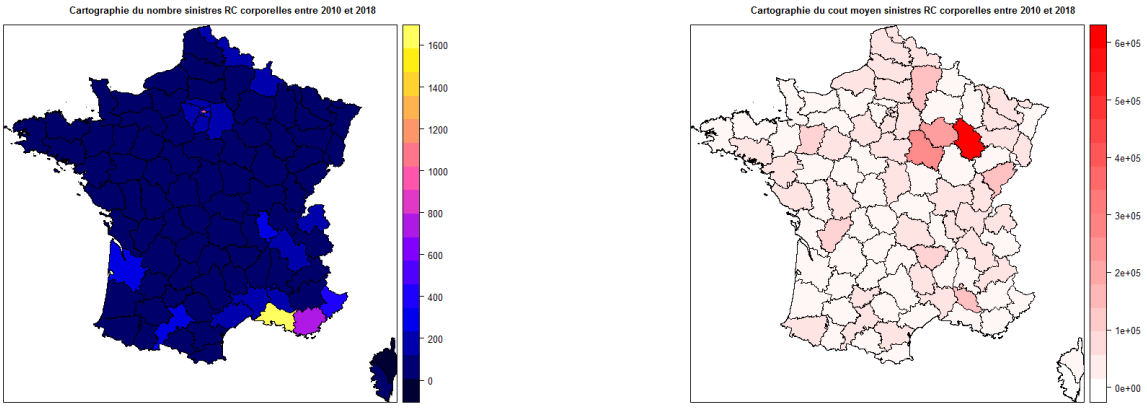


FIGURE 1.8 – Cartographie de la sinistralité

Sur la première carte, on observe une sinistralité plus importante en métropole dont la région parisienne et les départements de la région PACA en lien avec une densité de population plus importante. La cartographie du coût moyen reste plus difficile à analyser par la présence de quelques sinistres graves, on n'observe pas de cluster de sinistres ici.

Enfin, nous représentons ici le nombre de sinistres par an pour chaque partenaire entre 2012 et 2018. Pour des raisons confidentielles, on ne donnera pas les noms des partenaires en question ni le produit commercialisé.

FIGURE 1.9 – Hétérogénéité des partenaires

Nous observons une grande hétérogénéité du nombre de sinistre par an au sein d'un même partenaire mais également entre partenaires (en lien avec leur chiffre d'affaire et le poids du portefeuille).

Classification des partenaires

Parmi les méthodes de statistiques exploratoires multidimensionnelles, dont l'objectif est d'extraire d'une masse de données des "'informations utiles"'. Nous distinguons les méthodes d'analyse factorielle, qui ont pour objectif de visualiser des données et de réduire le nombre de variables, des méthodes de classification automatique. Le but de la classification automatique est de former des groupes d'individus ou de variables afin de structurer un ensemble de données. Nous cherchons à homogénéiser notre portefeuille de partenaires afin d'améliorer la modélisation.
Dans notre étude, nous allons appliquer des techniques de clustering afin d'homogénéiser le plus possible notre portefeuille de partenaire. Nous utiliserons les résultats des partenaires ainsi que leur chiffre d'affaires. Cette classification devra répondre à plusieurs problématiques :

1. Permettre de modéliser un partenaire disposant de peu d'historique à partir des historiques de son groupe ;
2. Regrouper les petits partenaires et avoir un indicateur comparatif de performance du groupe ;
3. Différencier les stratégies de renouvellement et les adapter à la situation du groupe.

Nous allons présenter, en annexe, deux méthodes de classification. Nous les avons appliqués à notre base de données et nous allons présenter les résultats dans cette partie. Cependant les résultats

de classification obtenus ne nous ont pas permis un regroupement optimal des données. Nous choisissons de regrouper les partenaires en fonction de leurs primes acquises à l'exercice, très fortement corrélées à leur sinistralité et aux poids de leur portefeuille sur le portefeuille global.
Nous regrouperons donc les partenaires en trois groupes, en fonction de leurs primes acquises et de leurs sinistralités annuelles.
Si l'on reprend les résultats du K-means, grâce à une analyse en composante principale on peut voir que les groupes peuvent être déterminés en ne prenant en compte que la prime acquise. Les mouvements du portefeuille n'ont pas d'impact sur les résultats de ce dernier sauf dans le cas d'une forte croissante (augmentation de la sinistralité) ou d'une chute importante de la croissance (chute de la sinistralité).
Ce mouvement est très bien représenté par le montant de primes acquises dans l'année. De plus, le dernier cas caractérise souvent un portefeuille en "'Run off"', dont la commercialisation est arrêtée et donc va sortir de notre périmètre de revalorisation de primes.
On va donc procéder à un regroupement en lien avec le mode de déroulement de l'équipe :

1. permettre aux "petits" courtiers de s'insérer sur le marché automobile et de les accompagner dans leur business afin d'atteindre ce niveau de stabilité.
2. Obtenir des "moyens courtiers" qui sont déjà sur le marché un niveau de rentabilité performant, tout en surveillant la volatilité du portefeuille.
3. Surveiller les portefeuilles les plus "gros", qui ont atteint un niveau stable de rentabilité, et permettre un développement sur d'autres produits d'assurance.

Chapitre 2

Modélisation de la sinistralité grave

Sommaire

2.1 Théorie des valeurs extrêmes

La présence de sinistres graves dans notre portefeuille vient perturber l'hypothèse d'homogénéité des risques. Nous allons ici appliquer quelques fondements de la théorie des valeurs extrêmes afin de déterminer un seuil à partir duquel un évènement peut être considéré comme atypique. L'objectif étant d'étudier les évènements rares qui conduisent à des pertes importantes, dans ce chapitre, il va donc nous falloir répondre à plusieurs questions :

1. Comment déterminer le seuil optimal u correspondant aux évènements rares ?
2. Comment estimer les paramètres de la loi de Pareto Généralisée ?

3. Comment vérifier l'adéquation de nos données à la loi retenue ?

La théorie des valeurs extrêmes s'intéresse à la loi du maximum par sa distribution asymptotique grâce au théorème de Fisher Tippett. Cependant, la méthode par estimation des paramètres de la distribution GEV implique une extraction des données par blocs de même taille. Elle peut être réductrice du fait que l'utilisation d'un seul maxima peut conduire à une perte d'information. Ce problème est résolu en considérant tous les *excès*, c'est-à-dire la différence entre les observations au-delà d'un seuil donné et ce seuil en question, cette approche est appelée *méthode de dépassement de seuil* ou *méthode du Peaks Over Threshold*. Nous rappellerons en annexe les principes de la loi du maximum et du théorème de Fischer-Tippett.

La loi des excès

La méthode à dépassement de seuils consiste à utiliser les observations dépassant un certain seuil, que l'on notera u. En considérant toutes les valeurs au-delà d'un seuil donné, on ne risque pas une perte d'informations contrairement à l'approche classique, expliquée en annexe. Cependant, cette méthode nécessite la détermination d'un seuil ni trop faible, pour ne pas prendre un nombre d'excès trop important, ni trop élevé par manque de robustesse dans nos modélisations.
À partir de la distribution F_X, on veut définir une distribution conditionnelle F_u par rapport au seuil u pour les variables aléatoires dépassant ce seuil.

Définition La distribution conditionnelle F_u des excès au-delà du seuil u est définie par :

$$F_u(x) = P(X < x | X > u) = \frac{F(x) - F(u)}{1 - F(u)} \text{ pour } x > u$$

Le théorème de Pickands établit le lien entre le paramètre de la loi du domaine d'attraction maximum et le comportement limite des excès au-delà d'un seuil assez grand. L'indice de queue ξ est identique au paramètre de la loi de Pareto généralisée.
Dans la suite, on notera $Y = X - u$ pour $X \geq u$ et pour n variables aléatoires observées $X_1, ..., X_n$ nous pouvons alors définir $Y_j = X_i - u$ telle que i est l'indice du j^{eme} excès et $j = 1, ..., N_u$.

Théorème de Pickands

Théorème 1. *Une fonction de répartition F appartient au domaine d'attraction maximale G_ξ, si et seulement si, il existe une fonction positive σ telle que*

$$\lim_{u \to x_F} \sup_{0 \leq y \leq x_F - u} |F_u(y) - F^{GPD}_{\xi,\sigma(u)}(y)| = 0$$

où $F^{GPD}_{\xi,\sigma(u)}(y)$ est la fonction de répartition de la loi de Pareto Généralisée (GPD), définie par :

$$F^{GPD}_{\xi,\sigma(u)}(y) = \begin{cases} 1 - (1 - \xi \frac{y}{\sigma(u)})^{-1/\xi} & \textit{si } \xi \neq 0 \\ 1 - \exp(-\frac{y}{\sigma(u)}) & \textit{sinon.} \end{cases}$$

pour

$$\begin{cases} y \in [0, x_F - u] & \textit{si } \xi \geq 0 \\ y \in [0, \min(-\frac{\sigma(u)}{\xi}, x_F - u)] & \textit{si } \xi < 0 \end{cases}$$

Ce théorème est utile lorsque l'on cherche à modéliser les observations qui dépassent un seuil fixé en approchant la loi des excès par une loi de Pareto Généralisée, puisqu'il assure l'existence du paramètre d'échelle $\sigma(u)$ et de queue ξ. La principale difficulté est de choisir ce seuil approprié. L'indice de queue, quant à lui, permet de distinguer les lois à queue épaisse, qui appartiennent au domaine d'attraction de Fréchet, des lois à queue fine ou légère, qui appartiennent au domaine d'attraction de Weibull.

Méthodes de détermination du seuil grave

D'après le théorème de Pickands, pour un seuil u suffisamment élevé, la fonction de répartition $F^{GPD}_{\xi,\sigma(u)}$ peut être approchée par $F_u(y)$ pour de bonnes valeurs de ξ, le paramètre de queue et σ, le paramètre d'échelle. Cela signifie que, nous cherchons à reproduire le comportement d'une Pareto Généralisée.

En pratique, les méthodes graphiques que nous allons présenter permettent de sélectionner plusieurs seuils qui semblent répondre à notre problématique. Il convient alors de les tester et de déterminer le seuil optimal.
Cependant, le choix d'un seuil reste la principale difficulté puisqu'il faut trouver le bon équilibre entre précision et robustesse de notre estimateur. En effet, plus le seuil est élevé, plus la convergence de la loi des excès vers une GPD sera bonne, mais moins la précision le sera car nous diminuons le nombre d'observations.

Le Pareto Quantile Plot et le quantile Plot généralisé

Dans le domaine de Fréchet, les distributions F ont la propriété suivante

$$1 - F(x) = x^{-\frac{1}{\xi}} l_F(x)$$

et

$$Q\left(\frac{1}{x}\right) = U(x) = x^{\xi} l_U(x)$$

Avec l_F et l_U deux fonctions à variations lentes à l'infini et U étant une fonction définie par

$$U(x) = \inf\, y : F(y) \geq 1 - \frac{1}{x}$$

La notion de fonction à variation lente se relie à l'étude d'une fonction au voisinage de certains points. Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction numérique. On dit que :
(a) f est une fonction à variation lente à I'infini, ou une fonction variant lentement à l'infini ssi f est non négative et : $(1)_{\mathrm{a}}$ $f(\mathrm{x} \cdot \mathrm{y})/\mathrm{f}(\mathrm{x}) \to 1$ lorsque $\mathrm{x} \to \infty$, pour tout $\mathrm{y} > 0$.

(b) f est une fonction à variation lente à l'origine, ou une fonction variant lentement à l'origine, ssi f est non négative et : $(1)_{\mathrm{b}} \quad f(\mathrm{x}\cdot \mathrm{y})/\mathrm{f}(\mathrm{x}) \to 1$ lorsque $\mathrm{x} \to 0, \quad$ pour tout $\mathrm{y} > 0$.
Si nous considérons la statistique d'ordre $X_{1,n} \leq ... \leq X_{n,n}$ associée à nos données initiales, le Pareto quantile plot correspond au graphe de $\left(\ln(\frac{n+1}{j}), \ln(X_{n-j+1,n})\right)$. C'est une représentation très utile pour visualiser graphiquement si nos données sont distribuées selon une loi du domaine de Fréchet ou non. Dans ce domaine, le graphe serait approximativement linéaire avec une pente ϵ, pour les petites valeurs de j, c'est-à-dire les points extrêmes.

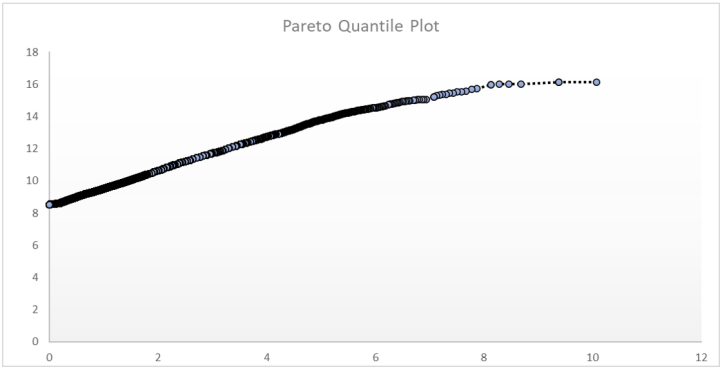


FIGURE 2.1 – Pareto Quantile Plot

Nous pouvons observer, sur le Pareto quantile plot, une pente positive et linéaire. Donc, nos données sont approximativement distribuées selon une loi du domaine de Fréchet.

Une autre approche permettant d'éviter le choix a priori du domaine d'attraction a été proposée par Beirlant *et al.* (1996). Elle consiste à utiliser un quantile plot généralisé défini par le graphe $\left(\ln\left(\frac{n+1}{j+1}\right), \ln\left(UH_{j,n}\right)\right)$ avec $UH_{j,n}$ de la forme

$$UH_{j,n} = X_{n-j,n}\left(\frac{1}{j}\sum_{i=1}^{j}\ln(X_{n-i+1,n}) - \ln(X_{n-j,n})\right)$$

Suivant la pente de ce graphe, nous pouvons en déduire le domaine dans lequel appartiennent les points extrêmes.

1. Dans le domaine de Fréchet, les points extrêmes forment une droite de pente positive.
2. Dans le domaine de Gumbel, la droite est constante
3. Dans le domaine de Weibull, nous constatons une tendance négative.

Cet estimateur $UH_{j,n}$ correspond au produit du k^{eme} plus grand point extrême avec son estimateur de Hill que l'on définira plus tard dans ce mémoire.

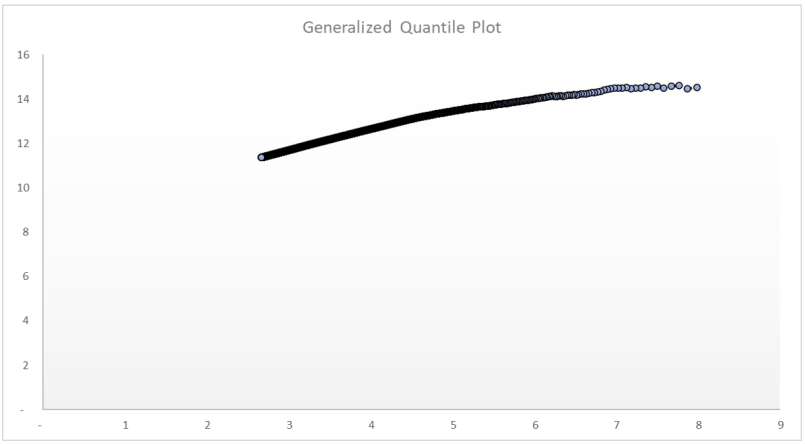


FIGURE 2.2 – Generalized Quantile Plot

D'après le graphique, nous retrouvons le même résultat qu'avec le Pareto quantile Plot.

Threshold Stability Plot

Le diagramme de stabilité des paramètres, aussi appelé Threshold stability plot par Scarrot et MacDonald, examine les estimations du paramètre de forme et d'échelle afin de trouver un seuil approprié. Au-delà d'un certain seuil u auquel le modèle GPD devient valide, les paramètres doivent être stables, c'est-à-dire constants.

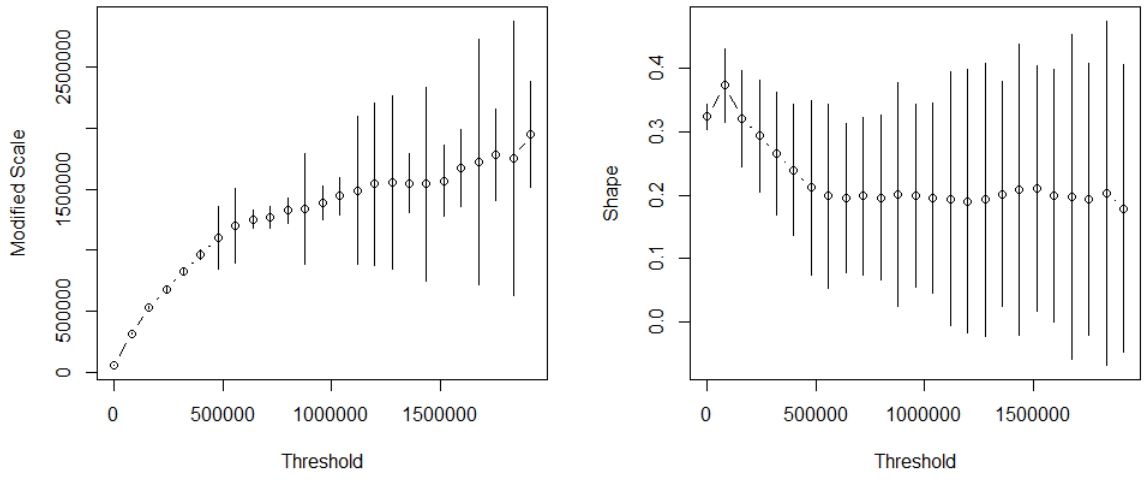


FIGURE 2.3 – Threshold stability plot de nos données

Dans les graphiques ci-dessus, l'axe des abscisses correspond aux seuils u des coûts de sinistres et l'axe des ordonnées correspond respectivement aux paramètres d'échelle et de forme. D'après l'analyse des deux graphes, la stabilisation des paramètres est significative à partir du seuil $5x10^5$, nous obtenons donc les paramètres GPD $(1x10^6, 0.2)$.

Mean Excess Plot

La fonction moyenne des excès permet de décrire la prédiction du dépassement du seuil u lorsqu'un excès se produit et est définie par :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)^+}{\sum_{i=1}^{n} \mathbb{I}_{X_i > u}}$$

où $(X_i - u)^+ = \sup(X_i - u, 0)$
.

Afin de choisir le seuil u, cette approche consiste à tracer l'estimateur empirique de l'espérance résiduelle de X. On choisit u de manière à ce que $\hat{e}_n(u)$ soit approximativement linéaire pour tout $x \geq u$.

Si, à un certain seuil, la fonction moyenne des excès empirique présente une pente positive, alors les données suivent une distribution de Pareto Généralisée avec un paramètre de queue positif.

Lorsque la fonction moyenne des excès empirique présente une pente horizontale, les données suivent une distribution exponentielle.

Enfin, lorsque la fonction moyenne des excès empirique présente une pente négative, les données suivent une distribution à queue légère, c'est-à-dire un paramètre de queue négatif.

Etude sur l'ensemble de la base

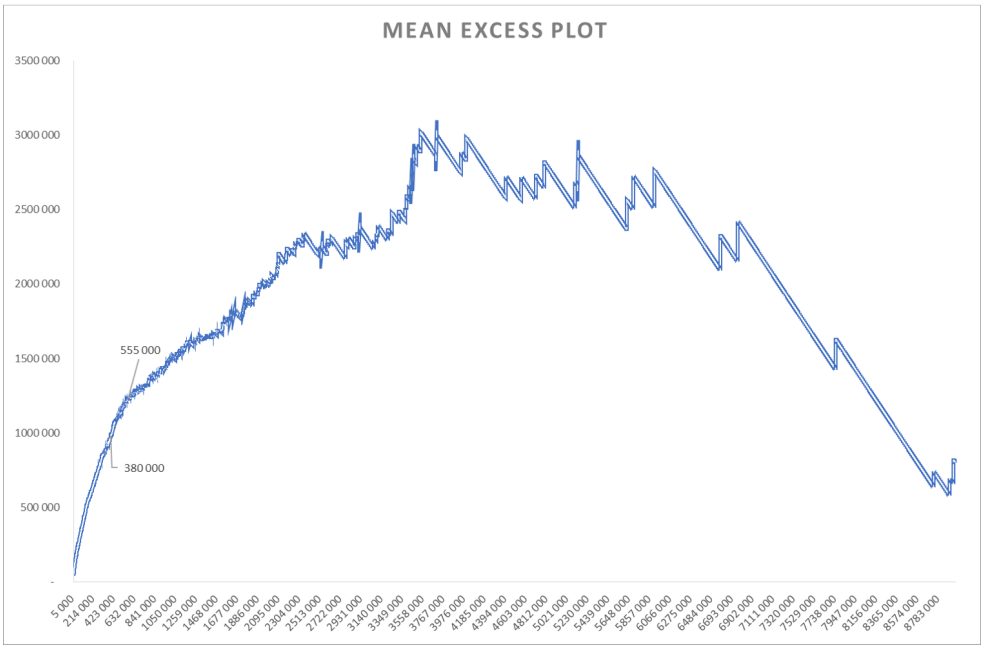


FIGURE 2.4 – Mean Excess plot sur tous les données

L'axe des abscisse représente le seuil u des coûts des sinistres et l'axe des ordonnées correspond à la moyenne des excès définie ci-dessus.

Nous cherchons à déterminer un seuil qui définit parfaitement nos graves. Nous constatons sur ce graphique un changement de pente aux alentours de 3,5 millions. L'instabilité du graphique à partir de ces montants témoigne d'une grande incertitude, en effet, plus nous nous dirigeons vers la droite de l'axe des abscisses, plus nous augmentons l'incertitude liée au manque de données. L'idée ici est donc de déterminer un changement de pente linéaire le plus à gauche de l'axe des abscisses. Notre choix final se porte sur le changement de pente que nous pouvons observer aux alentours de $3x10^5$ et $5x10^5$.

Seuil optimal

Pour un seuil u fixé, si nous considérons un $v \geq u$ alors en notant $Y = [X - u|X > u]$ nous avons l'égalité en loi $[X - v|X > v] = [Y - (v - u)|Y > v - u]$ et comme Y suit une GPD $(\xi, \sigma(u))$:

$$\mathcal{E}\left[X - v|X > v\right] = \frac{\sigma(u) + \xi(v - u)}{1 - \xi}$$

Ainsi, l'espérance des excès est une fonction affine de v, lorsque v est supérieur au seuil de référence u. Cela nous permet de tester si le seuil u est suffisamment élevé.
Nous fixons u à $5x10^5$ et nous cherchons à savoir si ce seuil est optimal graphiquement.

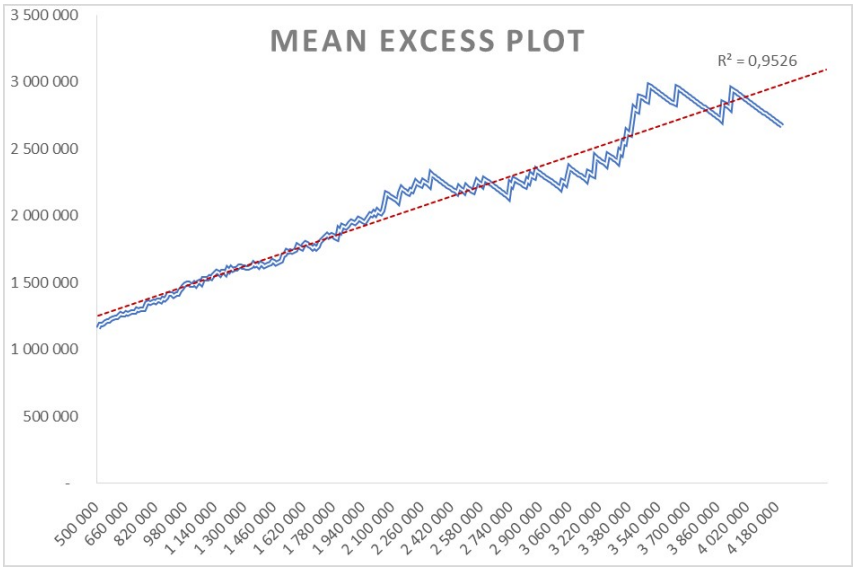


FIGURE 2.5 – Résultat graphique pour la validation du seuil optimal pour le seuil extrême

Nous observons une stabilisation affine de l'espérance des excès au-delà de ce seuil fixé à $5x10^5$. Nous ne regardons pas les valeurs au-delà car nous savons d'après la figure 2.4 que les données à l'extrêmes présentent une pente négative. Cependant, le nombre d'excès est beaucoup trop faible, ce choix ne semble pas pertinent vis-à-vis de notre problématique.

Etude sur les valeurs inférieurs à $5x10^5$ Nous allons observer maintenant le comportement du plot en dessous du seuil précédemment étudié.

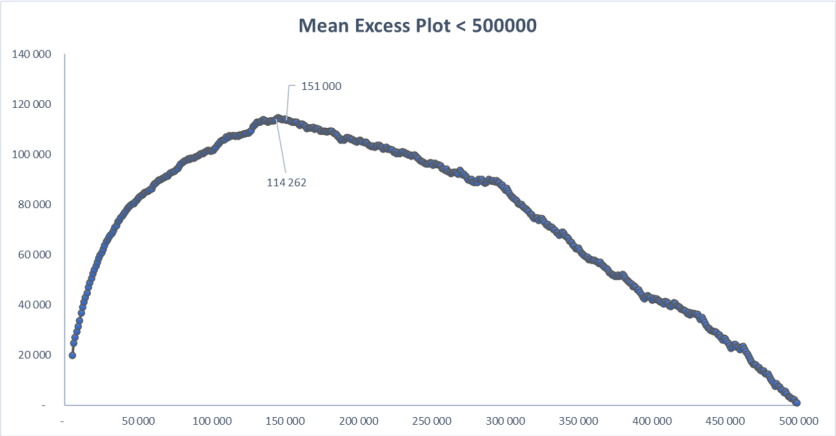


FIGURE 2.6 – Mean Excess plot sur données inférieures à $5x10^5$

L'intérêt de cette approche est qu'elle nous permet de voir le comportement de nos données en tronquant la partie que l'on considère la plus extrême. Les valeurs extrêmes n'influencent plus graphiquement la tendance des valeurs graves.

Nous obtenons une pente décroissante aux alentours de $1.5x10^5$. Ce qui contredit les observations ci-dessus qui considèrent que nos données sont distribuées selon une loi du domaine de Fréchet. En retirant les valeurs les plus extrêmes, nous nous retrouvons dans le cas d'une loi du domaine de Weibull. Nous nous fixons, alors, le seuil à $1.5x10^5$ et nous allons regarder les valeurs supérieures à ce seuil.

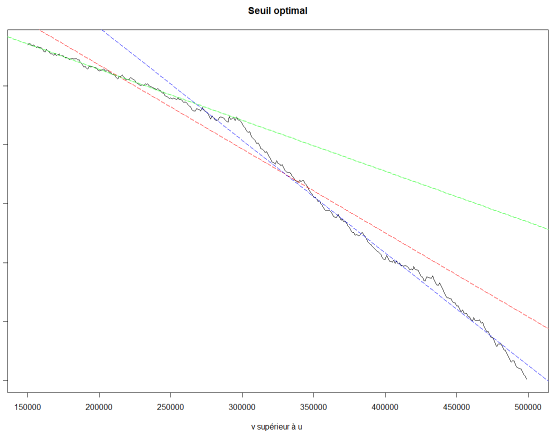


FIGURE 2.7 – Résultat graphique pour la validation du seuil optimal pour le seuil grave

Les résultats sont un peu plus délicats à analyser que dans le cas extrême, nous observons une fonction de moyenne des excès affine en v mais un changement de pente est, également, observé aux alentours de $3x10^5$. Malgré ce changement, nous allons considérer ce seuil comme graphiquement optimal.
Notons u_1 le premier seuil aux alentours de $1.5x10^5$ et u_2 le deuxième seuil aux alentours de $5x10^5$. Par conséquent, l'analyse des graphes du Mean Excess Plot nous amène à penser qu'il pourrait y avoir deux seuils à partir desquels les évènements sont considérés comme graves mais les lois correspondantes appartiennent à deux domaines d'attraction différents :

1. la distribution de $X_i - u_1$ telle que les $X_i \in [u_1, u_2]$ est à queue épaisse : nous pouvons la modéliser par une distribution de Pareto Généralisée avec un paramètre de queue positif
2. la distribution de $X_i - u_2$ telle que les $X_i \geq u_2$ est à queue légère : nous pouvons la modéliser par une distribution de Pareto Généralisée avec un paramètre de queue négatif.

Pour la suite, l'idée sera d'obtenir un estimateur de queue pour trois types de données :

1. les données au-dessus d'un seuil u_1.
2. les données au-dessus d'un seuil u_2
3. les données entre u_1 et u_2

A partir de ces estimateurs obtenus, nous allons déterminer l'approche la plus cohérente pour notre problématique.

Estimation de l'indice de queue des extrêmes

Hill-Plot

Le paramètre u peut être déterminé graphiquement en exploitant la linéarité de la fonction moyenne des excès $e(u)$ pour la GPD. Cependant, plusieurs valeurs de u doivent être testées. En supposant que G appartenait au domaine d'attraction maximale de Fréchet, *Beirlant et al.[1996]* ont suggéré d'opter pour un seuil u qui minimise l'erreur quadratique asymptotique de l'estimateur de Hill de l'indice de queue.
L'estimateur de Hill, basé sur la statistique d'ordre $X_{k,n} \leq ... \leq X_{1,n}$, est défini par :

$$\hat{\xi}_{n,k}^{Hill} = \frac{1}{k-1} \sum_{j=1}^{k-1} ln(X_{j:n}) - ln(X_{k:n})$$

Comme l'estimateur de Hill n'est défini que pour $\xi > 0$, l'idée est de sélectionner u le plus grand possible pour une valeur d'estimation stable. Nous avons sélectionné plusieurs points sur le graphique suivant. Un point va représenter le k^{ieme} extrême qui va définir notre seuil.

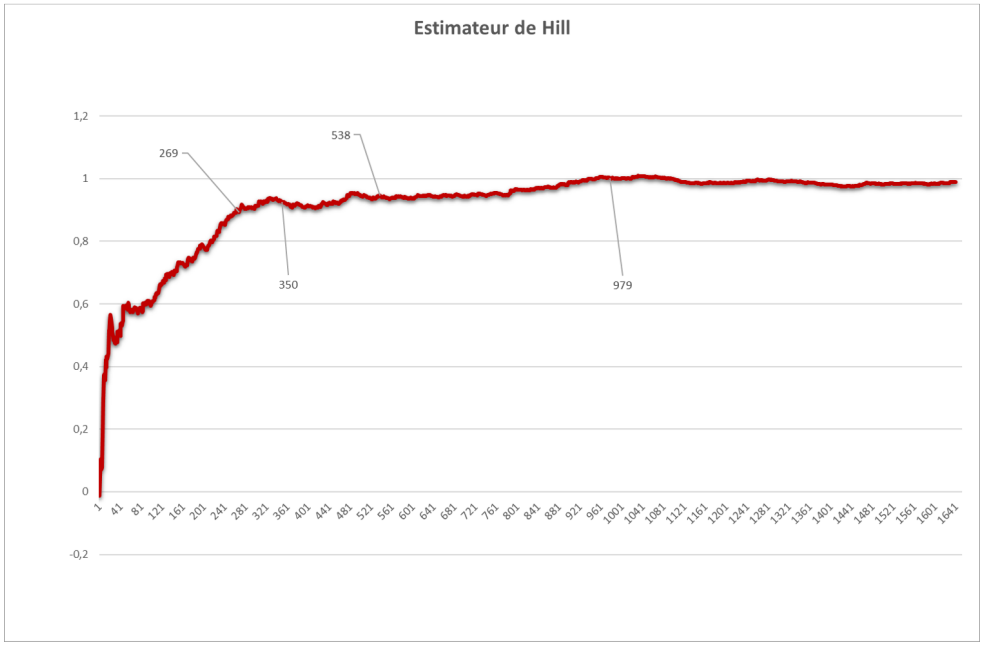


FIGURE 2.8 – Hill Plot

L'axe des abscisses représentent le nombre d'excès k, L'axe des ordonnées représentent l'indice de queue α, c'est-à-dire $\frac{1}{\xi}$
Les points, que nous avons sélectionnés sur ce graphe représentent des étapes de stabilisation de l'estimateur qui semblent intéressantes à analyser. Plus notre nombre d'excès est important, plus notre estimateur est stable mais le choix d'un trop grand nombre d'excès n'est pas intéressant dans notre étude. L'intérêt de cette analyse est de trouver le bon équilibre entre nombre d'excès et

stabilité de l'estimateur. De plus, nous obtenons un grand nombre de seuil à tester et évaluer par la suite.

Pickand's Plot

L'estimateur de Pickands est défini par la statistique :

$$\hat{\xi}_{n,k}^{Pickands} = \frac{1}{\ln(2)} \ln\left(\frac{X_{k:n} - X_{2k:n}}{X_{2k:n} - X_{4k:n}}\right)$$

Contrairement à l'estimateur de Hill, il reste valable quelle que soit la distribution des extrêmes. Cependant, la représentation graphique de cet estimateur en fonction du nombre k d'observations considérées montre un comportement en général très volatile au départ, ce qui nuit à la lisibilité du graphique.

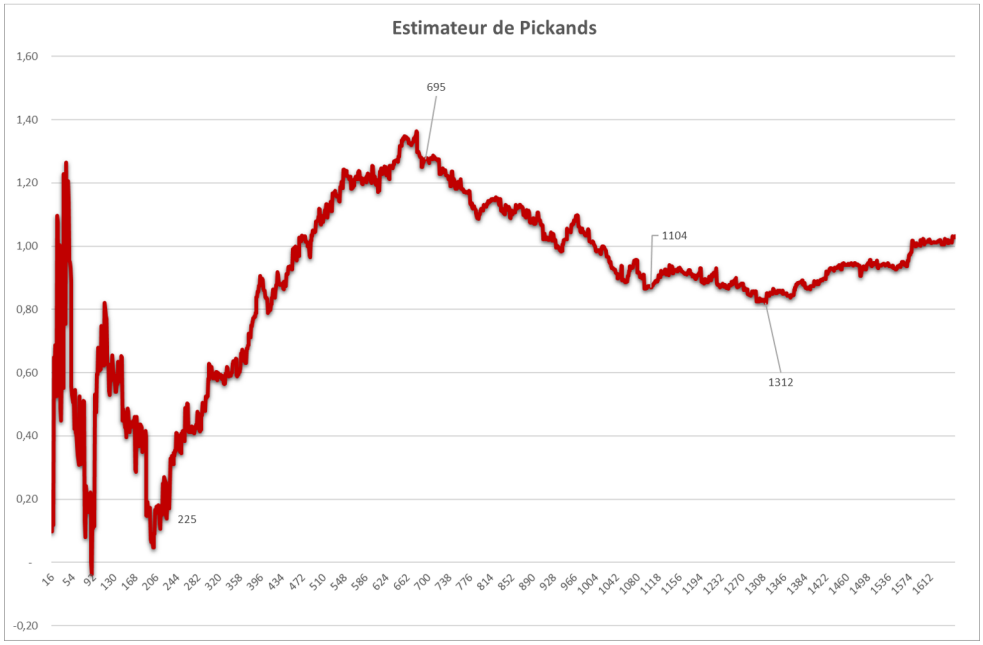


FIGURE 2.9 – Pickand's plot sur tous les données

Comme dit précédemment, l'estimateur de Pickands est très volatile au départ et il est difficile à déterminer une stabilité avec un nombre d'excès moindre. Nous avons, quand même, décidé de nous restreindre à moins de 1500 événements extrêmes, pour choisir nos seuils et leurs estimateurs de queue. Nous pouvons alors observer sur le graphique, les choix de seuil à analyser puis comparer à la suite.

Estimation des paramètres de la loi des excès

Estimateur du maximum de vraisemblance

La fonction de log-vraisemblance pour un échantillon de n_u excès $Y_1, ..., Y_{n_u}$ *i.i.d.* suivant une loi

de Pareto Généralisée est définit par :

$$\mathcal{L}(Y,\xi,\sigma) = -n_u \ln(\sigma) - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{n_u}\ln\left(1+\frac{\xi}{\sigma}y_i\right) \quad \xi \neq 0$$

avec $\left(1+\frac{\xi}{\sigma}y_i\right) > 0$, pour $i = 1, ..., n_u$. Dans le cas où $\xi = 0$, on aura

$$\mathcal{L}(Y,\sigma) = -n_u \ln(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n_u} y_i \quad \xi = 0$$

En utilisant la reparamétrisation $\tau = \xi/\sigma$, on obtient :

$$\frac{1}{\tau} = \frac{1}{n_u}\left(\frac{1}{\hat{\xi}(\tau)}+1\right)\sum_{i=1}^{n_u}\frac{Y_i}{1+\tau Y_i}$$

Cette équation se résout numériquement de manière itérative.
Graphiquement, nous pouvons voir que pour un seuil $1.5x10^5$ et après 60 itérations, nous obtenons nos estimateurs du maximum de vraisemblance.

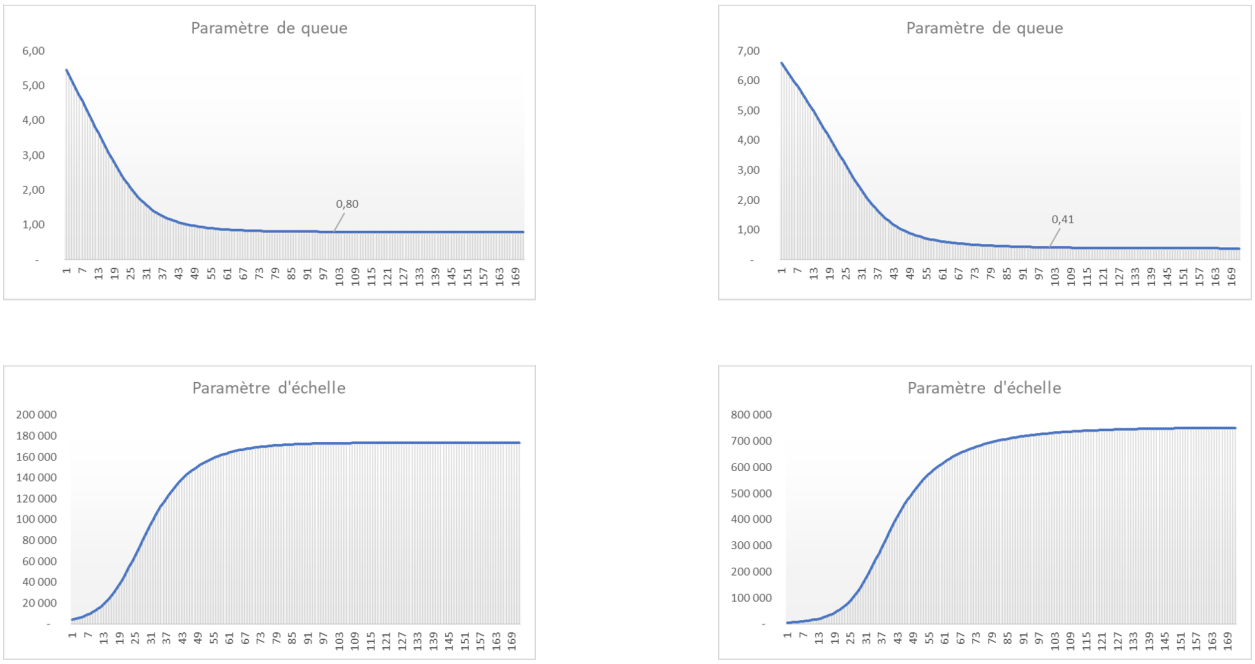


FIGURE 2.10 – Convergence des estimateurs du maximum de vraisemblance par itération avec un seuil à $1.43x10^5$, à gauche, et avec un seuil à $5.37x10^5$, à droite ;

Estimateur des moments et des moments pondérés
Ces estimateurs ont été introduits par Hosking et Wallis. La méthode des moments et la méthode des moments pondérés sont basées sur la comparaison des moments théoriques d'une distribution et leurs versions empiriques. Le moment d'ordre r pour la distribution de Pareto Généralisée existe pour $\xi < 1/r$, ce qui restreint son utilisation.

2.2 Résultats

Dans cette partie, nous allons présenter les résultats de la modélisation des seuils par méthode graphique avec deux approches différentes.

1. Dans le cas où nous ne considèrerons qu'un seuil de graves pour la base de données.
2. Dans le cas où nous ne considèrerons deux seuils distincts, l'un de graves dits ordinaires, l'autre d'extrêmes.

Détermination d'un seuil unique sur l'ensemble de la base

Prenons la première approche, nous représentons les seuils obtenus ainsi que les estimateurs de queues pour chaque méthode présentée plus tôt, et le nombre d'excès associé.

Résultat sur l'ensemble du portefeuille				
u_1	nombre d'excès	ξ^{Hill}	ξ^{MLE}	$\xi^{Pickands}$
143 293	979	1.00	0.80	1.04
251 500	580	0.94	0.74	1.20
408 211	350	0.93	0.50	0.64
537 366	269	0.89	0.38	0.43

Pour chaque cas, nous nous trouvons dans un domaine de Fréchet puisque l'estimateur de queue obtenu, pour chaque méthode, est strictement positif. Pour rappel, cette analyse est en lien avec l'étude par quantile plot et par le mean excess plot. Cependant, nous obtenons souvent des estimateurs très distincts d'une méthode à une autre. Par conséquent, il va donc nous falloir choisir un estimateur et, pour nous aider dans notre décision, nous allons déterminer le niveau d'adéquation de nos données à la loi théorique grâce à des tests statistiques. L'estimateur, qui permettra à la loi de modéliser au mieux nos données, nous donnera à la fois la loi théorique, l'approche à privilégier et le(s) seuil(s) à choisir.

Adéquation aux lois pour le seuil optimal

Dans cette partie, nous allons tester l'ajustement de nos données à la loi Pareto Généralisée pour chaque estimateur par le test de Kolmogorov Smirnov.
Nous rappelons les principes du test :
Soit F_X la fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ et G_n, celle de la loi *à tester*. Le test de **Kolmogorov Smirnov** est défini par la statistique de test :

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_n(x) - G_n(x)|$$

Nous rejetons l'hypothèse H_0 si $D_n > d_{n,\alpha}$ où $\alpha = 0.05$.
Dans le tableau suivant, nous représentons u_1 le seuil choisi, $d_{n,\alpha}$ le seuil de rejet déterminé par la table de Kolmogorov Smirnov, la statistique de test des différents estimateurs. Si cette statistique est plus grande que notre seuil de rejet alors nous rejetons l'hypothèse nulle entre parenthèses.

Test des différents seuils par Kolmogorov Smirnov				
u_1	$d_{n,\alpha}$	Statistique de test - Hill	Statistique de test - MLE	Statistique de test - Pickands
143 293	0.043	0.044 (Rejeté)	0.023 (Accepté)	0.048 (Rejeté)
251 500	0.056	0.048 (Accepté)	0.035 (Accepté)	0.069 (Rejeté)
408 211	0.073	0.078 (Rejeté)	0.051 (Accepté)	0.051 (Accepté)
537 366	0.083	0.106 (Rejeté)	0.021 (Accepté)	0.023 (Accepté)

Le choix semble compliqué puisque les différents seuils peuvent être acceptés en fonction de la méthode d'estimation de queue. Dans tous les cas, l'estimateur du MLE est accepté par le test de Kolmogorov.

Par la suite, nous choisissons de comparer, graphiquement, la fonction de répartition empirique avec les fonctions théoriques obtenues à partir de chaque méthode, pour les deux seuils les plus extrêmes.

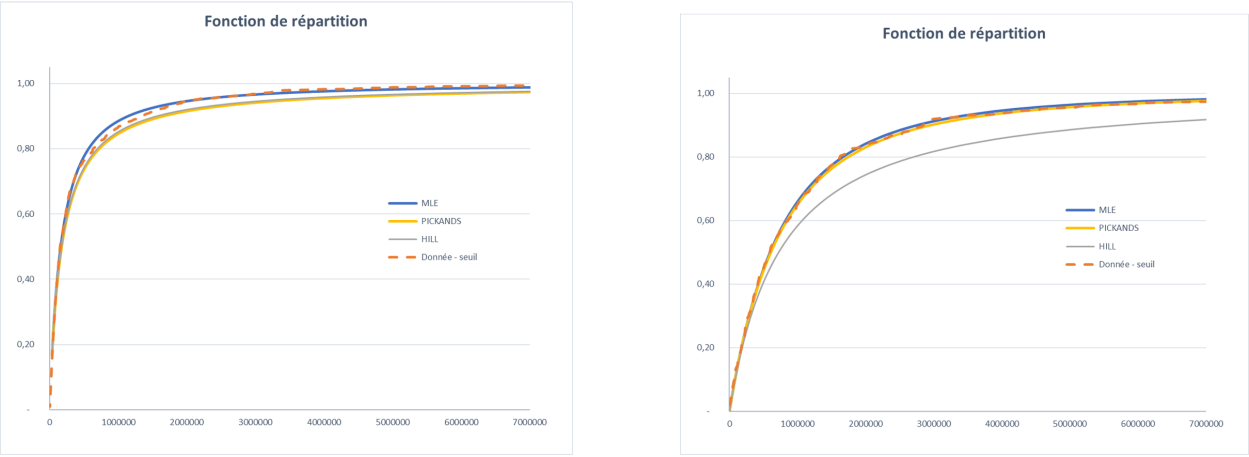
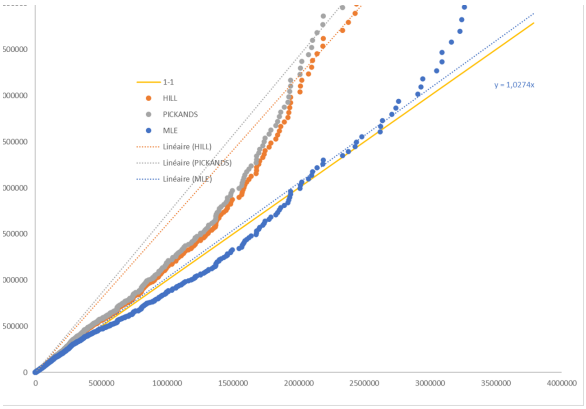


FIGURE 2.11 – Comparaison des estimateurs avec la fonction de répartition pour un seuil à $1.43x10^5$ et un seuil à $5.37x10^5$

Nous faisons la même approche avec le quantile plot, cependant nous excluons l'estimateur de Hill pour le deuxième graphique.



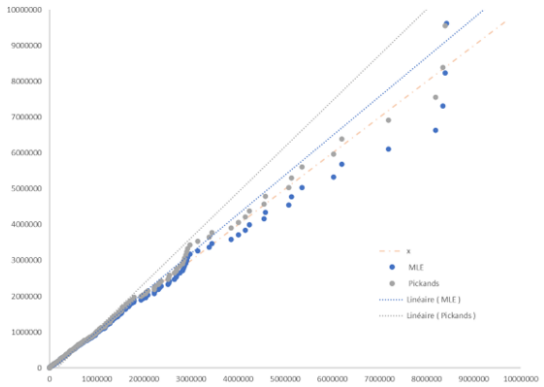


FIGURE 2.12 – Comparaison des estimateurs avec le quantile plot pour un seuil à $1.43x10^5$ et un seuil à $5.37x10^5$

Nous observons une très bonne estimation avec un choix de seuil à $5.37x10^5$. Cependant, nous nous retrouvons dans un cas où il existe peu de données. Il sera difficile par la suite de pallier à ce problème surtout lors de la modélisation de la fréquence des sinistres nous introduirons une grande volatilité. Le deuxième seuil à $1.43x10^5$ est intéressant puisqu'il est très proche du seuil déjà appliqué opérationnellement par Generali. De plus, nous observons un nombre d'excès plus important malgré une estimation de la loi légèrement moins appréciable que dans le premier cas.

Détermination de deux seuils de graves

Nous supposons, ici, que le seuil des extrêmes pour l'ensemble du portefeuille est défini à $5.37x10^5$, en lien avec les résultats graphiques précédents. Grâce au Mean Excess Plot, nous avons pu déterminer une tendance en dessous de ce seuil, avec un changement de pente aux alentours de $1.5x10^5$. L'idée ici va être de supposer qu'il n'existe pas un mais deux seuils de sinistres graves et de :

1. Rechercher ce seuil aux alentours de $1.5x10^5$.
2. Savoir quelles techniques il est possible d'utiliser vis-à-vis du domaine d'attraction.
3. Déterminer si les estimateurs obtenus permettent une adéquation optimale à une loi de Pareto généralisée.

Nous observions dans le Mean Excess plot une tendance décroissante au-dessus de $1.5x10^5$. Si nous choisissons de regarder autour de ce seuil, l'estimateur de Hill ne peut plus être un critère de choix dans notre analyse. Nous allons reprendre le graphe pour generalized quantile plot pour trancher sur le choix de l'estimateur.

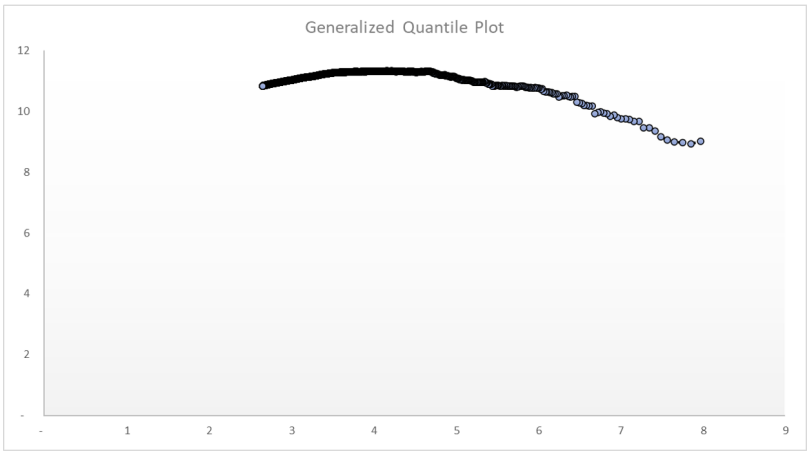


FIGURE 2.13 – Generalized Quantile Plot, données tronquées

En effet la tendance est décroissante, nous allons tracer l'estimateur de Pickands mais cette fois ci pour des valeurs en dessous du premier seuil :

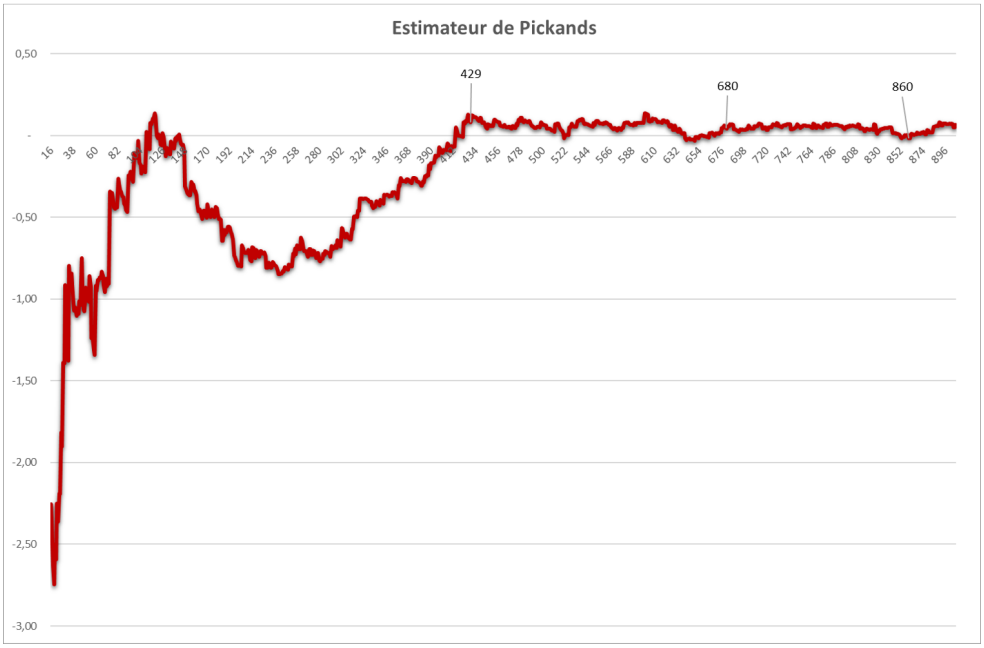


FIGURE 2.14 – Pickands plot, données tronquées

Nous obtenons graphiquement trois seuils potentiels que nous avons décrits dans le tableau ci dessous. Pour chacun de ces seuils, nous avons déterminé l'estimateur du maximum de vraisemblance avec une contrainte :

$$\tau_0 = \xi/\sigma(u) > -\frac{1}{y}$$

avec y la plus grande valeur de notre échantillon, c'est-à-dire $y = X_{n,n} - u$. Si nous prenons une valeur initiale plus grande que 0, nous obtenons un estimateur du MLE très proche de zéro en lien

avec l'estimateur de Pickands. Si nous prenons l'estimateur plus petit que $-\frac{1}{y}$, nous ne pouvons avoir de convergence optimale.

Résultat sur l'ensemble du portefeuille				
Paramètres	u_2	nombre d'excès	ξ^{Pick}	ξ^{MLE}
Ensemble	211 002	429	0.08	-0.49
	148 621	680	0.05	-0.36
	113 668	980	0.18	-0.126

Les résultats obtenus entre les estimateurs de Pickands et ceux du MLE sont assez différents. De plus, ils ne proposent pas de domaine d'attraction identique. L'intérêt de regarder l'adéquation de nos données est de choisir le bon estimateur mais aussi le bon domaine d'attraction. De plus, celui de Pickands ne reflète pas les résultats obtenus avec le MLE et le Generalized quantile plot.

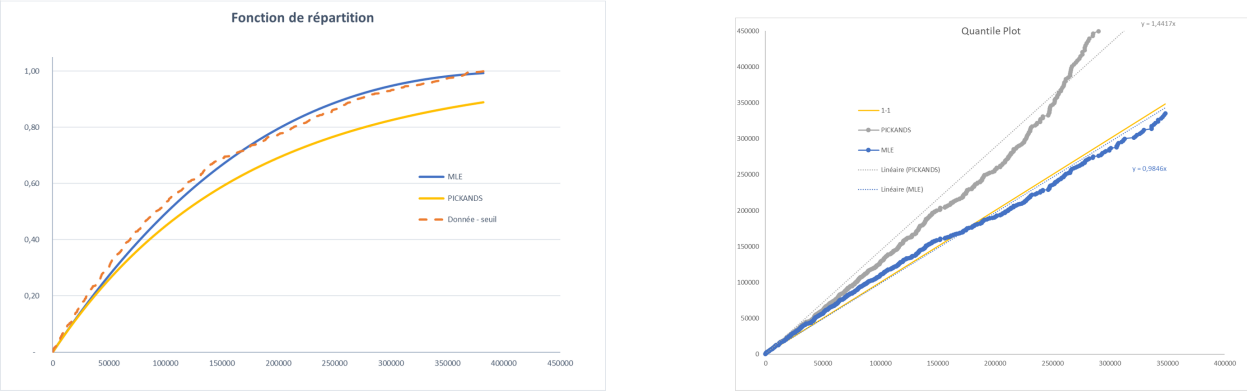


FIGURE 2.15 – Adéquation de loi avec un seuil à 148 621

D'après le test de Kolmogorov Smirnov et avec un seuil à 148 621, l'estimateur du MLE est le plus adéquat à nos données. En effet, la statistique de test calculée est de 0.046 (contre 0.116 avec Pickands) alors que la table de Kolmogorov donne la valeur critique de 0.052. Nous sommes bien dans un domaine d'attraction de Weibull et non de Fréchet comme nous le proposait l'estimateur de Pickands.
Faire le choix de modéliser deux segments au lieu d'un seul revient à soulever le problème de la robustesse de nos modélisations par la suite, notamment sur la fréquence. En effet, nous chercherons à modéliser seulement la survenance de sinistres graves pour l'année 2018. Par la suite, nous ferons donc le choix de modéliser seulement à partir d'un seuil de graves que l'on supposera à $1.44x10^5$.

Adéquation de notre loi sur les années plus récentes

Les résultats obtenus, plus tôt, sont basés sur des données du passé allant jusqu'à 2004. Observons à présent l'adéquation de nos données plus récentes à notre loi modélisée à partir des estimateurs du MLE. Nous représentons ici la fonction de répartition empirique et la fonction d'une Pareto Généralisée modélisée à partir du MLE.

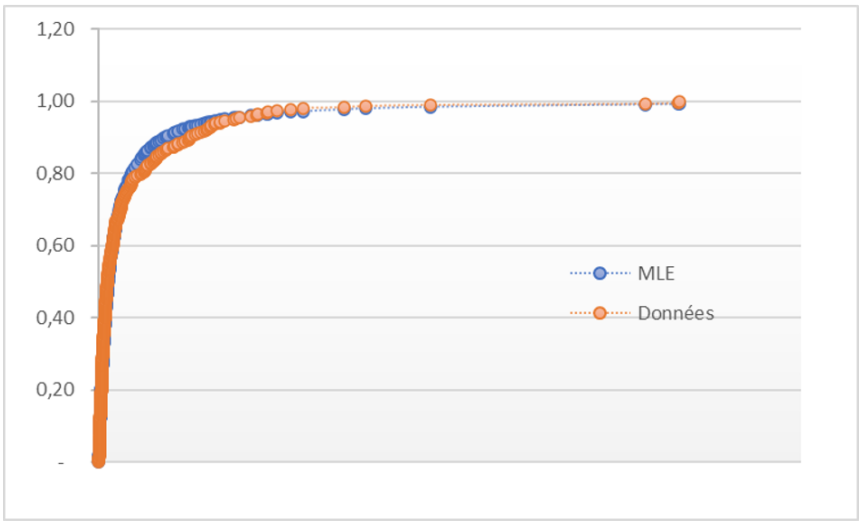


FIGURE 2.16 – Adéquation sur les trois dernières années

En choisissant de comparer notre loi théorique obtenue à partir des estimateurs du MLE à nos données sur les dernières années, nous espérons obtenir une loi qui représente toujours la tendance des prochaines années. En effet, en regardant le graphique ci-dessus, nous observons une très bonne adéquation entre la fonction de répartition théorique et la fonction de répartition empirique. Ces résultats sont rassurants pour la suite. Enfin ceci nous permet d'accentuer le choix de l'approche pour un seuil unique, à $1.44x10^5$ et d'un estimateur par la méthode du maximum de vraisemblance.

Chapitre 3

Modélisation de la fréquence des sinistres graves

Sommaire

L'objectif de ce chapitre est de pouvoir appliquer une méthode d'estimation du nombre de sinistres graves annuellement survenus. Bien que la fréquence soit faible par rapport aux nombres de sinistres attritionnels, leur impact sur le ratio de sinistralité de partenaire est significativement important. Pour rappel, cet effet est particulièrement visible dans le secteur de l'automobile où les corporels peuvent engendrer des dépenses considérables. Dans un premier temps, la fréquence de ces sinistres sera modélisée sur l'ensemble du portefeuille. Nous appliquerons, sur des données chronologiques, des processus ARIMA afin de pouvoir prédire les valeurs futures de notre série. Dans un second temps, nous allons modéliser cette fréquence annuelle sur chaque ensemble de partenaires x produits en appliquant des Modèles Linéaires Généralisés. Les principales raisons de cette approche sont :

— la possibilité de modéliser nos sinistres à partir des informations recueillies sur les partenaires
— la possibilité de détecter une loi autre que Gaussienne pour approcher nos données, notamment la loi de Poisson et Binomiale Négative.

Enfin, nous changeons d'approche de modélisation et nous allons appliquer des principes de la théorie Bayésienne afin d'obtenir une information sur le nombre de sinistres à appliquer sur chaque partenaire que son expérience soit forte soit faible. Contrairement à la modélisation généralisée, nous faisons le choix de ne pas regrouper nos "petits" partenaires en amont de la modélisation.

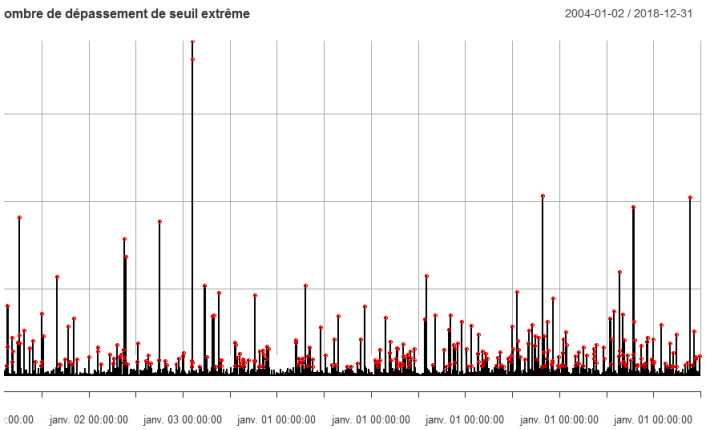


FIGURE 3.1 – série temporelle des sinistres et dépassement de seuil grave

3.1 Projection grâce à des modèles ARIMA - GARCH

L'objectif de cette étude est de pouvoir prédire les valeurs futures de notre variable représentant le nombre de sinistres en fonction des valeurs observées, présentes et passées. Toute variable peut être définie par une fonction qui assigne une probabilité à l'apparition de chacune des valeurs possibles que peut prendre la variable.

Rappel sur les séries temporelles et les processus stochastiques

Définition et structure d'une série temporelle

Un processus stochastique est une famille de variables aléatoires $X_t, t \in \mathbb{N}$. Pour tout $t \in \mathbb{N}$ X_t est une variable aléatoire réelle $X_t : \Omega \longrightarrow \mathbb{R}$. Pour $\omega \in \Omega$, $X(\omega)$ est une réalisation de la variable aléatoire X_t.
Les fonctions $t \longrightarrow X_t(\omega), \omega \in \Omega$ sont les trajectoires du processus.

Nous considérons qu'une série temporelle est un morceau de trajectoire d'un processus stochastique avec n le nombre d'observations, appelé longueur de la série.

Une série temporelle peut être définie à partir de deux types de structures. Elle peut être définie de façon **additive** c'est-à-dire $X_t = a_t + \epsilon_t$ ou de façon multiplicative, c'est-à-dire de la forme $X_t = a_t * \epsilon_t$.
a_t représente la composante déterministe, définie par $a_t = m_t + s_t$ où m_t est la tendance, s_t la saisonnalité, et ϵ_t la composante non déterministe du processus appelée "bruit".

Composante stochastique

Un processus stochastique $(\epsilon_t)_{t\in\mathbb{N}}$ est appelé bruit blanc fort si ses variables ϵ_t sont centrées, indépendantes et identiquement distribuées
Un bruit blanc est dit faible si ses variables ϵ_t sont centrées, décorrélées, et de variances finies constantes $cov(X_s, X_t) = \sigma^2\delta_{s=t}$.

Stationnarité

Pour prédire les valeurs futures d'une série temporelle, il est nécessaire que cette série présente une certaine reproductibilité. Notre série temporelle est un processus fortement stationnaire si pour tout $(t_1, ..., t_n)$ et tout h, la loi de $(X_{t1}, ..., X_{tn})$ est la même que celle de $(X_{t_1+h,...,t_n+h})$, ce qui équivaut à dire que la loi est invariante en temps.
Cette notion de stationnarité forte est très difficile à vérifier en pratique, nous lui préférons la notion de stationnarité faible :
On dit d'une série temporelle qu'elle est un processus faiblement stationnaire si c'est un processus de second ordre, c'est-à-dire admettant un moment d'ordre 2, tel que pour tout t, $E(X_t) = \mu$ et $\forall s, t, h, :$

$$Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h})$$

.
Nous définissons la fonction d'autocorrélation $\rho(h)$ d'un processus X_t tel que

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

où γ représente la fonction d'autocovariance d'un processus et est définie par

$$\gamma(h) = cov(X_t, X_{t+h})$$

La fonction d'autocorrélation empirique $\hat{\rho}$ est définie par

$$\hat{\rho}(h) = \frac{\sum_{k=1}^{n-h} \left(X_{k+h} - \bar{X}\right)\left(X_k - \bar{X}\right)}{\sum_{k=1}^{n-h} \left(X_k - \bar{X}\right)^2}$$

La notion d'autocorrélation partielle

Lorsque nous nous intéressons à caractériser les dépendances d'au moins trois variables aléatoires, il est nécessaire d'introduire la notion de corrélation partielle. Si nous considérons les variables $X_1, ..., X_k$, X_1 peut être corrélée à X_3 car X_1 et X_3 sont toutes deux corrélées à X_2. Nous notons M_h le sous-espace vectoriel engendré par $(X_2, ..., X_h)$ et P_{M_h} la projection sur ce sous-espace. La fonction d'autocorrélation partielle du processus est définie par

$$r(h) = corr\left(X_{h+1} - P_{M_h}(X_{h+1}), X_1 - P_{M_h}(X_1)\right)$$

pour $h \geq 2$. Par convention, nous adoptons $r(1) = corr(X_1, X_2)$.

Transformation de notre série en série stationnaire

Cette première étape est importante dans la prévision d'une série chronologique car elle permet de se ramener à un processus stationnaire. Il existe différents procédés permettant de détecter puis de corriger la tendance et la saisonnalité d'une série temporelle.

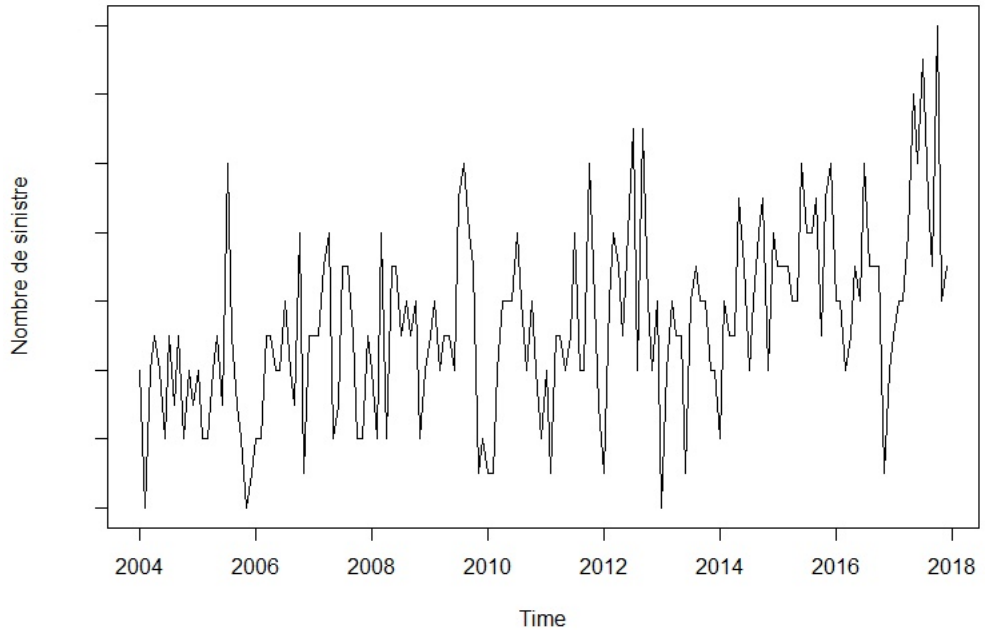


FIGURE 3.2 – Série temporelle du nombre de sinistre

Dans le graphique ci-dessus, nous représentons la série chronologique que nous allons étudier. Cette série correspond au nombre total de sinistres graves, tout portefeuille confondu, entre 2004 et 2018. Le seuil de sinistres graves a été déterminé dans le chapitre précédent et nous avons agrégé la série par mois de survenance.

Estimation paramétrique de la tendance

Après avoir représenté la série, il est souvent possible d'inférer une représentation paramétrique de sa tendance. Nous pouvons procéder par régression linéaire pour estimer cette tendance.
La tendance se calcule grâce à l'estimateur des moindres carrés ordinaires, c'est-à-dire $\hat{m} = (T^tT)^{-1}T^tX$, avec T la variable de temps.

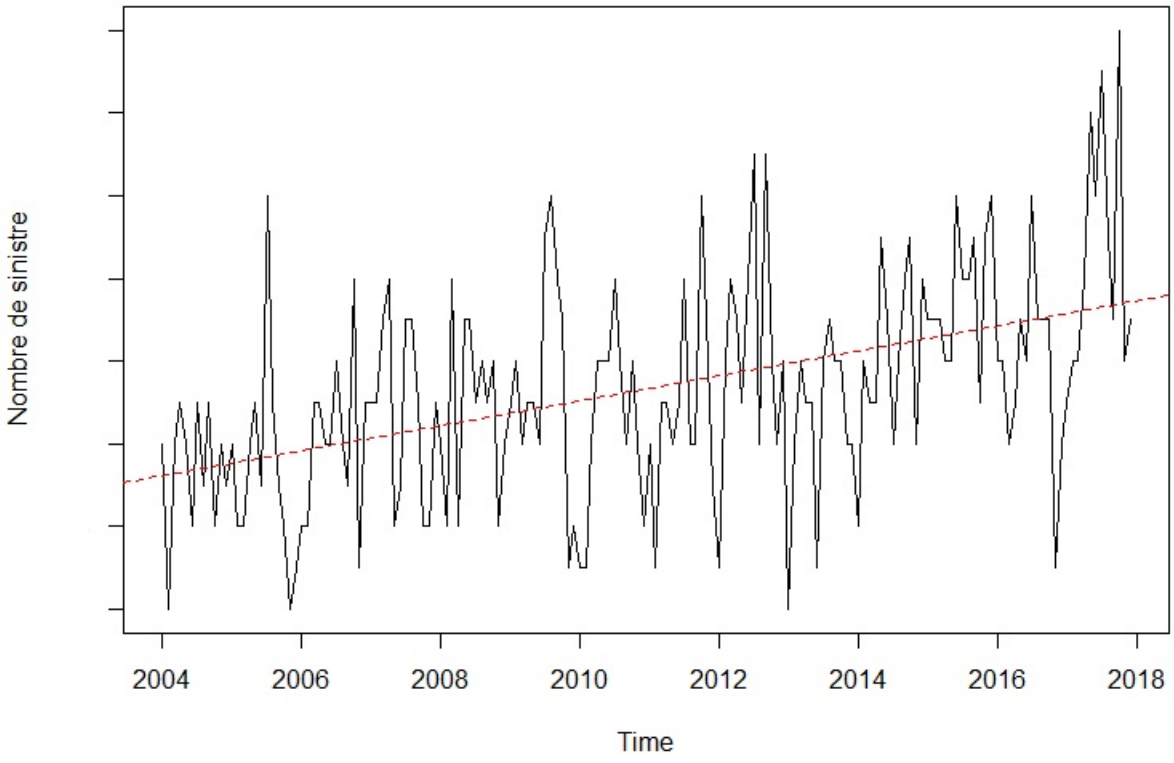


FIGURE 3.3 – Tendance de notre série

La tendance apparaît, dans notre série, de manière évidente. Nous pouvons observer une augmentation de la sinistralité grave en lien avec l'augmentation du chiffre d'affaires de l'organisme d'assurance. Nous avons déterminé une régression linéaire d'ordre 1 à partir des données et estimer le paramètre de pente grâce à l'estimateur des moindres carrés. Nous observons ainsi $\hat{m} = 0.3025$.

Détecter les coefficients saisonniers

Les coefficients saisonniers indiquent la moyenne des variations saisonnières ou leur importance. Après avoir retiré toute tendance de notre série, nous pouvons traiter la composante saisonnière ∇S_t. Comme $\Delta_\tau S_{t+\tau} = \Delta_\tau S_t$ avec τ la période, s'il existe une partie saisonnière à nos données nous pouvons la considérer annuelle, c'est-à-dire $\tau = 12$. Les coefficients de saisonnalité $(\Delta_\tau S_j)_{2\leq t\leq \tau}$ sont faciles à estimer à partir de la moyenne empirique de notre série.

$$\Delta_\tau \hat{S}_j = \frac{\tau}{n} \sum_{1\leq t=k\tau+j\leq n} (\Delta_\tau X_t) \quad 1 \leq j \leq \tau$$

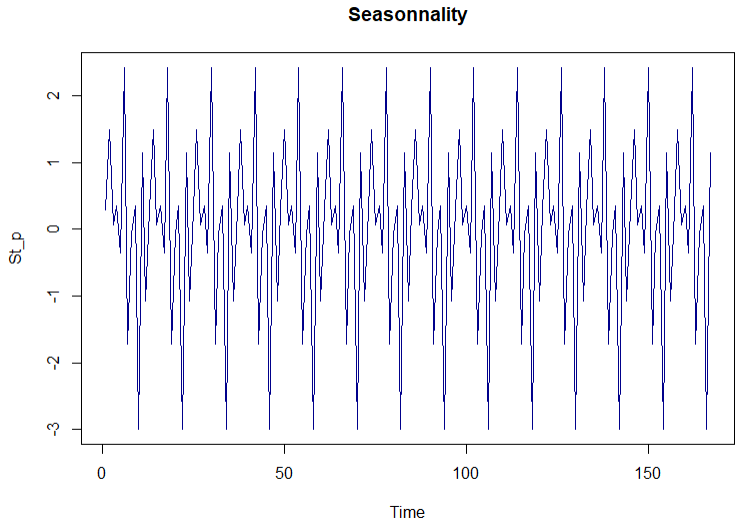


FIGURE 3.4 – Saisonnalité

Différenciation

Notons Δ l'opérateur de différentiation : $\Delta X_t = X_t - X_{t-1}$ et notons l'opérateur de différentiation d'ordre k : $\Delta^k X_t = \Delta(\Delta^{k-1} X_t)$
Supposons que notre processus X_t admette une tendance polynomiale d'ordre 2. Nous obtenons alors un processus ΔX_t d'ordre 1.
Supposons maintenant que ce processus admette une saisonnalité additive de période τ, alors le processus $\Delta_\tau X_t = X_t - X_{t-\tau}$ est un processus désaisonnalisé.

Test de stationnarité KPSS

Afin de déterminer si notre série temporelle est un processus stationnaire, nous choisissons de réalisé le test de Kwiatkowski-Philips-Schmidt-Shin (KPSS). Ce test détermine si notre processus est stationnaire autour d'une tendance moyenne ou linéaire, ou est non stationnaire en raison d'une

racine unitaire. L'hypothèse nulle pour le test est que les données sont stationnaires. L'hypothèse alternative pour le test est que les données ne le sont pas.

Un inconvénient majeur du test KPSS est qu'il présente un taux élevé d'erreurs de type I (il a tendance à rejeter trop souvent l'hypothèse nulle). Une façon de gérer le potentiel d'erreurs de type I élevées consiste à combiner le KPSS avec un test ADF, Augmented Dickey Fuller. Si le résultat des deux tests suggère que la série chronologique est stationnaire, alors c'est probablement le cas.

Nous avons suivi le protocole pour transformer notre série temporelle, appliquer les deux tests sur notre série et déterminer si cette dernière est bien stationnaire. En effet, nous obtenons une série stationnaire que nous représentons ci-dessous.

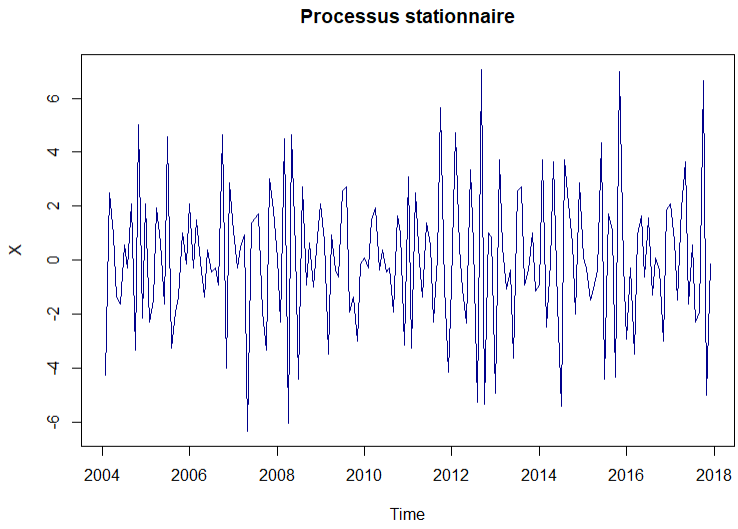


FIGURE 3.5 – Processus stationnaire obtenu après preprocessing

Modèle ARMA

Les processus autorégressifs

Un processus (X_t) est dit autorégressif d'ordre p centré s'il vérifie pour tout $t \geq 0$:

$$X(t) = \epsilon_t + \sum_{j=1}^{p} a_j X_{t-j}$$

avec p entier naturel non nul, $a_p \neq 0$ et ϵ_t qui forme un bruit blanc centré de variance σ^2 et indépendant de $X_{t-1}, X_{t-2}...$, pour tout t. Nous dirons alors que (X_t) est un processus $AR(p)$, c'est-à-dire, X_t "s'explique" par les p observations précédentes.

S'il existe un tel processus stationnaire $(X_t)_{t\geq 0}$ satisfaisant la relation de récurrence de l'équation, définie plus tôt, alors sa fonction d'autocovariance vérifie :

$$pour \;\; h>0, \gamma(h) = \sum_{i=1}^{p} a_i \gamma(h-i)\gamma(0) = \sigma^2 + \sum_{i=1}^{p} a_i \gamma(i)$$

Sa fonction d'autocorrélation vérifie

$$pour \;\; h>1, \rho(h) = \sum_{i=1}^{p} a_i \rho(h-i)$$

et enfin, sa fonction d'autocorrélation partielle vérifie :

$$r(h) = 0 \;\; si \;\; h \geq p+1$$

Les processus en moyenne mobile

Un processus (X_t) en moyenne mobile d'ordre q est un processus qui vérifie pour tout $t \geq 0$:

$$X(t) = \epsilon_t + \sum_{j=1}^{q} b_j \epsilon_{t-j}$$

avec q entier naturel non nul, $b_q \neq 0$ et $(\epsilon_{t)_t}$ qui forme un bruit blanc centré de variance σ^2 et indépendant de $X_{t-1}, X_{t-2}...$, pour tout t.
Nous dirons alors que (X_t) est un processus $MA(q)$.
L'autocovariance d'un processus $MA(q)$ vérifiant l'équation si dessus vérifie, pour $h \geq 0$,

$$\sigma(h) = \begin{cases} \sigma^2 \sum_{k=0}^{q-h} b_k b_{k+h} & si \quad h \leq q, \\ 0 \quad sinon \end{cases}$$

Les processus mixtes ARMA(p,q)

Un processus $(X_t)_{t\in\mathbb{N}}$ autorégressif d'ordre en moyenne mobiles p,q (tels que $p \geq 0$, $q \geq 0$, $p+q \geq 1$) est un processus qui vérifie pour tout $t \geq 0$:

$$X(t) = \sum_{j=0}^{p} a_j X_{t-j} + \sum_{k=1}^{q} b_k \epsilon_{t-k}$$

Nous dirons alors que (X_t) est un processus $ARMA(p,q)$

Tableau des propriétés			
Modèle	$MA(q)$	$AR(p)$	$ARIMA(p,q)$
autocorrélation	$\rho(h) = 0$ si $h > q$	$\rho(h) \to 0$ pour $h \to +\infty$	$\rho(h) \to 0$ pour $h \to +\infty$
autocorrélation partielle	$r(h) \to 0$ si $h \to +\infty$	$r(h) = 0$ si $h > p$	$r(h) \to 0$ si $h \to +\infty$

Nous allons nous servir de ces propriétés pour identifier nos séries temporelles. Nous allons tracer les $\hat{\rho}(h)$ et les $\hat{r}(h)$, c'est-à-dire les autocorrélations empiriques et les autocorrélations partielles empiriques.

Dans le cas du modèle MA(q) :

Nous allons tracer la fonction d'autocorrélation empirique $\hat{\rho}$ pour différentes valeurs de h que nous appelons *lag* dans le graphique. Un $\hat{\rho}(h)$ sous la courbe bleue, représentant le niveau de significativité, n'est pas significatif (au niveau α), et nous supposons alors que la valeur de $\rho(h)$ est nulle. Nous cherchons donc à savoir à partir de quel indice, nous considérons toutes les valeurs de l'autocorrélation nulle.

Nous allons représenter, ci-dessous, la fonction des autocorrélations empiriques :

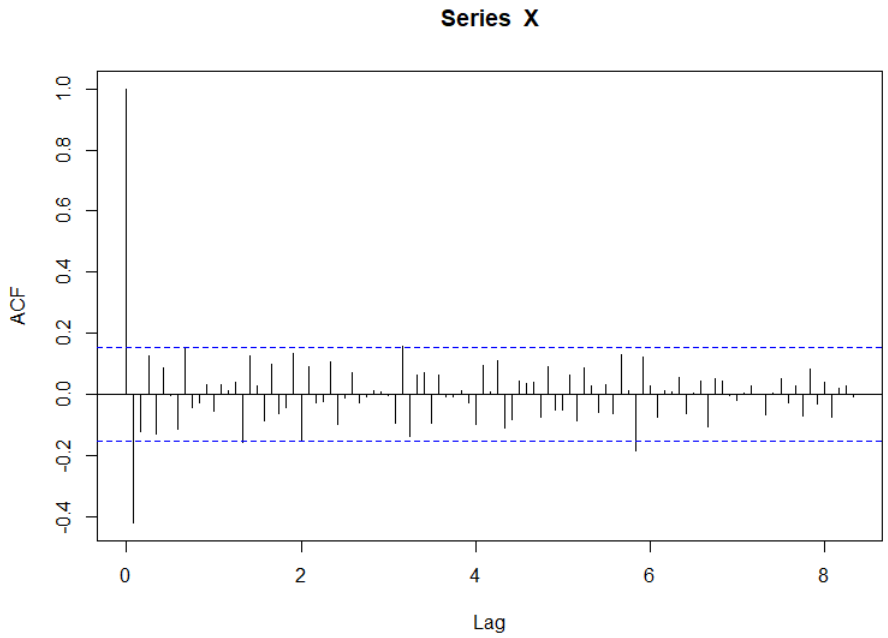


FIGURE 3.6 – autocorrélations empiriques

L'idée de prendre le plus petit h à partir duquel la valeur $\hat{\rho}(h)$ soit nulle n'est pas une chose aisée dans notre cas. Le choix optimal serait de prendre $h = 0$ ainsi nous pouvons dire que nous ne nous trouvons pas dans le cas d'un processus de moyenne mobile.

Dans le cas du modèle AR(p) :

Nous allons faire le même raisonnement avec $\hat{r}(h)$, la fonction d'autocorrélation partielle empirique.

Nous allons représenter, ci-dessous, la fonction des autocorrélations partielles empiriques :

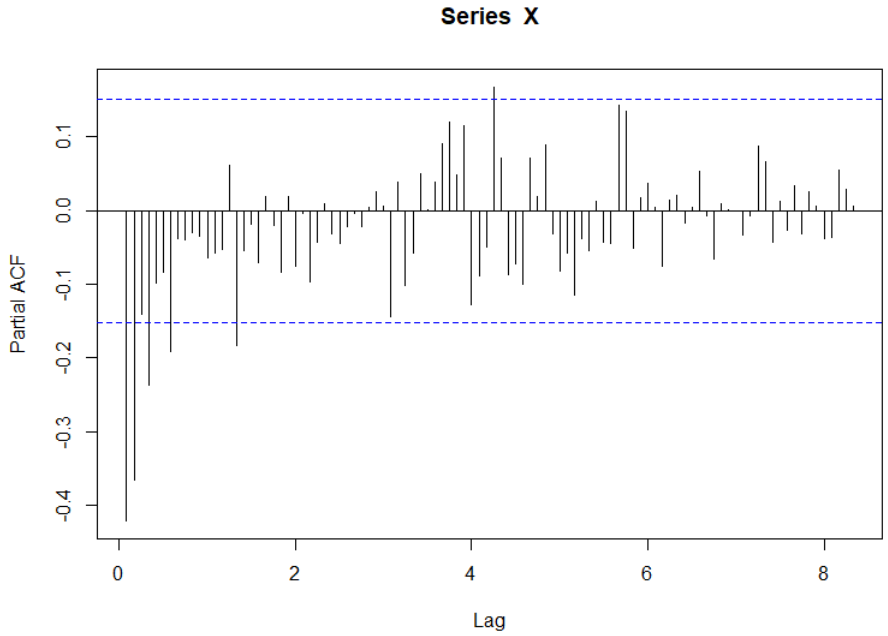


FIGURE 3.7 – autocorrélations partielles empiriques

Nous cherchons donc à savoir à partir de quel indice, nous considérons toutes les valeurs de l'autocorrélation partielle nulle. Nous observons des pics de significativité à h compris entre 1 et 2.
Nous pouvons regarder le BIC ou l'AIC de ces deux modèles et choisir le meilleur, c'est-à-dire celui qui minimise le plus ce critère.

Le critère d'information bayésien est un critère défini par :

$$BIC = -2\ln(\tilde{L}) + k\ln(n)$$

Avec $\tilde{L}$ la fonction de vraisemblance de notre modèle. Ce critère pénalise le nombre de variables présentes dans le modèle.
Pour les deux critères, celui qui représente le meilleur modèle est le modèle $AR(2)$.
L'idée étant de prendre le plus petit h à partir duquel la valeur $\hat{\rho}(h)$ soit nulle
Enfin, dans le cas du modèle ARMA(p,q) :

Afin de trouver les ordres p et q du modèle $ARMA$, une des approches est de considérer un ensemble crédible de modèles ARMA(p,q) et d'utiliser un critère d'information pour sélectionner le meilleur modèle. Ici, nous utilisons le critère du BIC, Bayesian Information Criterion, procédure qui va nous aider à choisir le bon ordre, en regardant les armasubsets (), une fonction dans R, qui nous aide à trouver un certain nombre de sous-ensembles de modèles ARMA.

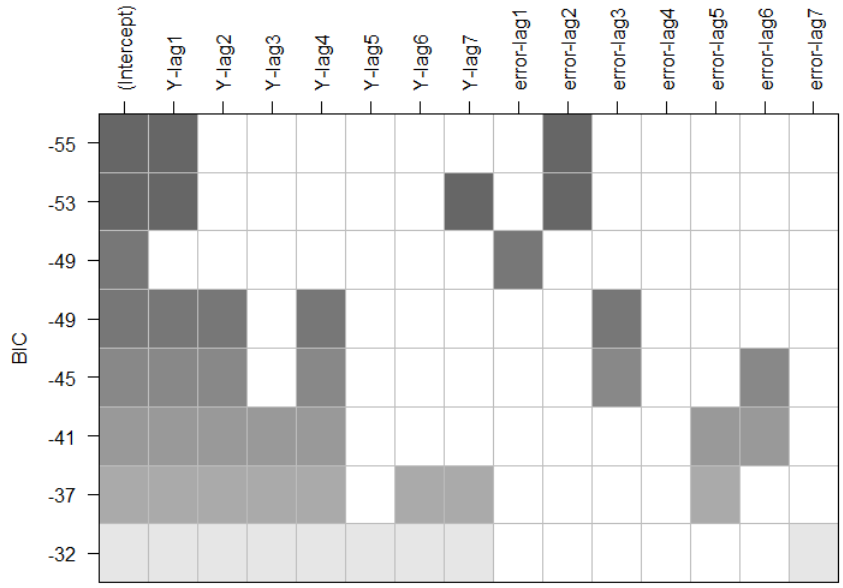


FIGURE 3.8 – Détermination des ordres pour un ARMA avec le critère du BIC

En observant la première ligne, correspondant à une valeur minimale de BIC, nous pouvons voir que le processus qui minimise le plus le critère du BIC est un processus $ARMA(1,2)$. Nous allons donc, pour la suite, utiliser ce modèle pour l'estimation des paramètres et la prédiction.

Estimation des paramètres du modèle ARMA

Dans le cas général, nous pouvons estimer les paramètres de notre modèle par maximum de vraisemblance. En effet, l'hypothèse de la normalité des résidus ϵ_t permet de spécifier une forme fonctionnelle à la vraisemblance du modèle. Si ϵ_t est gaussien, le vecteur des observations $(X_1, ..., X_n)$ est gaussien, lui aussi et, nous pouvons alors calculer sa vraisemblance.

Test sur les innovations

Pour chacun des modèles, si le processus a été correctement estimé, nous pouvons utiliser un test de Ljung-Box afin de tester l'autocorrélation de nos innovations. Si elles obéissent à un bruit blanc, il ne doit donc pas exister d'autocorrélation dans la série. Nous pouvons donc tracer les ACF et PACF et renforcer la vérification avec un test de Ljung-Box.
La statistique Ljung-Box Q permet de déterminer si une série contient des observations, dans le temps, aléatoires et indépendantes. Si ces observations ne sont pas indépendantes, une observation peut être corrélée avec une autre observation, k unités de temps après, établissant ainsi une relation appelée autocorrélation.
Les hypothèses du test sont :

— H_0 $\rho_1 = \rho_2 = \ldots = \rho_r = 0$ pas d'autocorrélation des erreurs d'ordre 1 à r

— H_1 l'un au moins des $\rho_i \neq 0$ il y a autocorrélation des erreurs d'ordre entre 1 et r.

Pour l'effectuer nous récupérons les résidus ϵ_t du modèle de base et nous construisons

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \ldots . + \rho_r \epsilon_{t-r} + X_t$$

Les MCO, moindre carré ordinaire, sur ce modèle donnent des estimations $\hat{\rho}_i$ des ρ_i. Ljung et Box ont montré que sous l'hypothèse H_0 la variable Q' qui vérifie

$$Q' = n(n+2) \sum_{1}^{r} \frac{\widehat{\rho}_i^2}{n-i}$$

suit une loi χ^2 à r degrés de liberté.

Nous allons montrer ici les résultats obtenus pour $ARMA(1,2)$

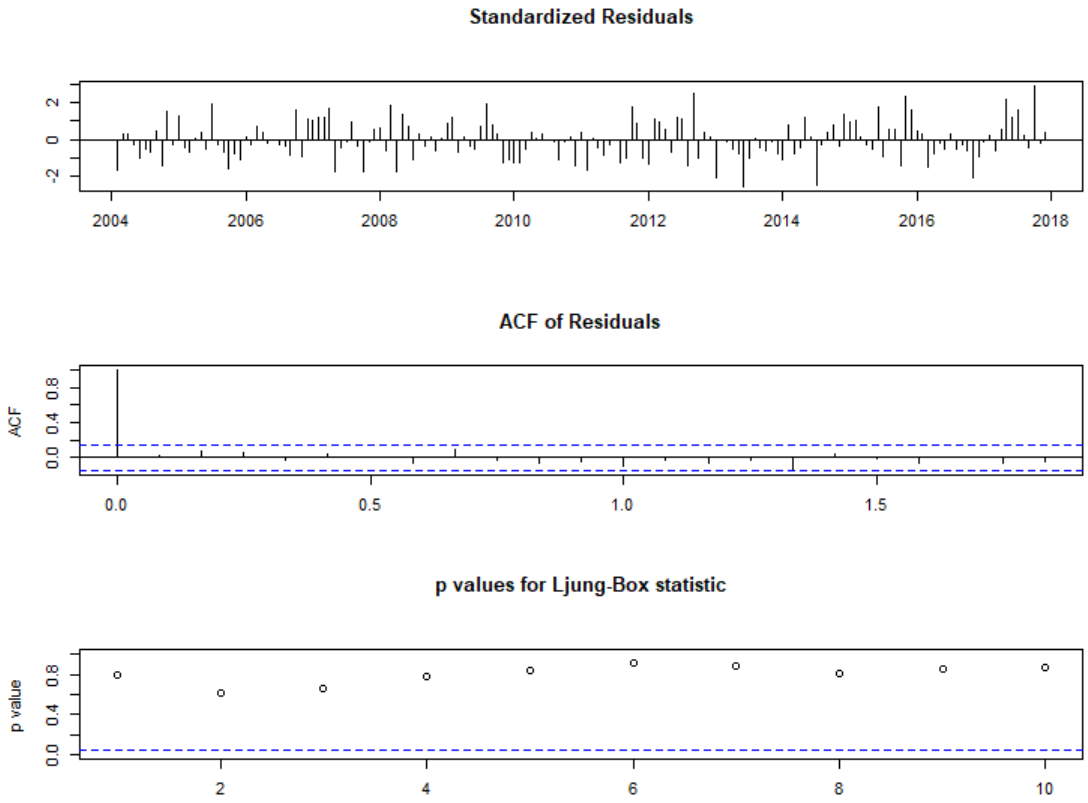


FIGURE 3.9 – Test de Ljung Box sur les ϵ_t

Dans le premier graphe, nous avons tracé les valeurs de nos résidus standardisés et nous pouvons observer que ces résidus sont compris entre $(-2, 2)$ et qu'ils oscillent uniformément le long de l'axe des abscisses, soit la valeur nulle. Le deuxième graphe représente la fonction d'autocorrélation empirique. Nous pouvons voir qu'il ne présente pas de valeur significative. Enfin le troisième

graphique est une représentation de la p-value du test de Ljung Box. Nous acceptons H_0 pour tout ordre r.
De plus, nous allons, pour compléter le diagnostic, observer si nos innovations forment un bruit blanc gaussien. Nous pouvons pour cela utiliser un test d'adéquation comme celui de Shapiro Wilk, nous rappellerons le test de Shapiro Wilk en annexe, et/ou utiliser un Gaussian quantile plot sur ces données.

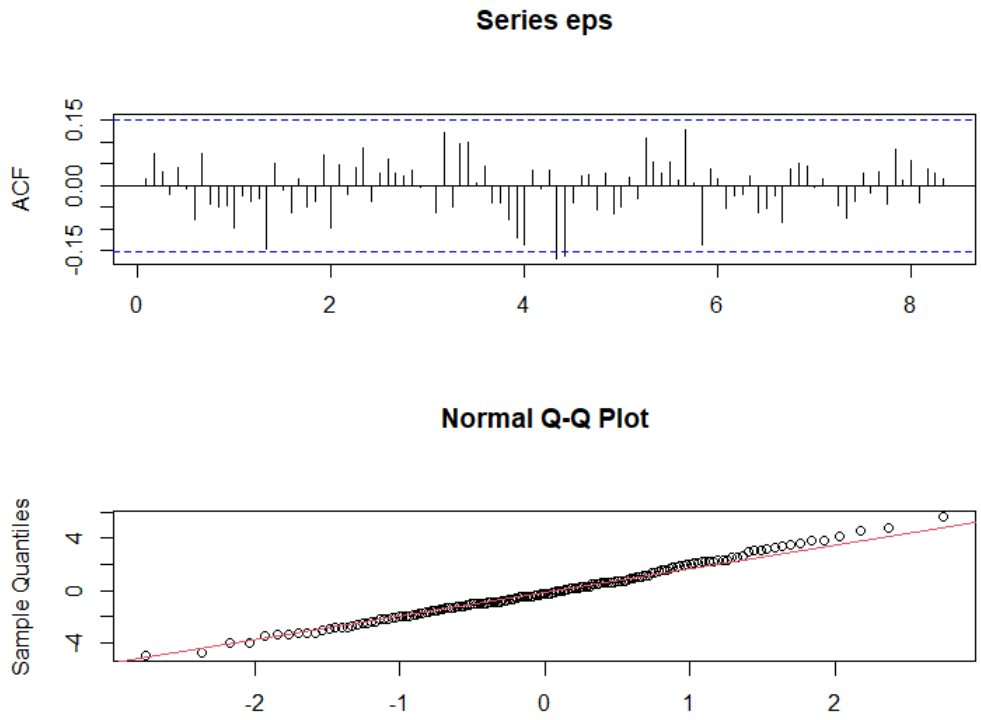


FIGURE 3.10 – Test de normalité sur les ϵ_t

Nous ne remarquons aucune valeur significative dans nos innovations et le normal quantile plot nous apporte une information supplémentaire sur l'adéquation de nos ϵ_t a une loi normale. Enfin le test de Shapiro Wilk nous donne une statistique de test $W = 0.99365$ et une $p - value = 0.6807$. Notre hypothèse de normalité est donc vérifiée.

Prévision

Si le modèle est un processus ARMA, la prédiction pour X_{n+h}, sachant $X_1, ..., X_n$ est

$$\hat{X}_{n,h} = c_1 X_1 + ... + c_n X_n$$

, où les coefficients sont choisis de manière à minimiser l'erreur quadratique

$$E\left[(X_{n+h} - c_1 X_1 - ... - c_n X_n)^2\right]$$

Ce choix $\hat{X}_{n,h}$ entraine l'égalité

$$\hat{X}_{n,h} = E\left(X_{n+h}|X_1, ..., X_n\right)$$

Ainsi l'erreur de prévision à l'horizon 1, $(X_{n+1} - \hat{X}_{n,1})$ est le bruit d'innovation ϵ_1.

Intervalle de confiance

Puisque les ϵ_t sont gaussiens, les $X_t, \hat{X}_{n,h}$ et $\hat{X}_{n,h} - X_{n+h}$ sont aussi gaussiens. Ceci permet de construire facilement des intervalles de confiance.
Supposons $\hat{X}_{n,h} - X_{n+h} \sim \mathcal{N}\left(0, \sigma^2\right)$ Soit $\alpha = 0,01$. Nous cherchons Δ tel que

$$\mathbb{P}\left(X_{n+h} \in \left[\widehat{X}_{n,h} - \Delta; \widehat{X}_{n,h} + \Delta\right]\right) \geq 1 - \alpha$$

(ici, toutes les probabilités sont conditionnelles à $X_1, \ldots, X_n$). Nous calculons

$$\begin{aligned}
&\mathrm{P}\left(X_{n+h} \in \left[\hat{X}_{n,h} - \Delta; \hat{X}_{n,h} + \Delta\right]\right) = \mathbb{P}\left(\left|X_{n+h} - \hat{X}_{n,h}\right| \leq \Delta\right) \\
&= \mathbb{P}\left(\frac{\left|X_{n+h} - \hat{X}_{n,h}\right|}{\sigma} \leq \frac{\Delta}{\sigma}\right) \\
&= 1 - 2\mathbb{P}\left(\frac{X_{n+h} - \hat{X}_{n,h}}{\sigma} \geq \frac{\Delta}{\sigma}\right) \geq 1 - \alpha
\end{aligned}$$

Nous voulons donc Δ tel que :

$$P\left(\frac{X_{n+h} - \hat{X}_{n,h}}{\sigma} \geq \frac{\Delta}{\sigma}\right) \leq \frac{\alpha}{2} = 0,005$$

Puisque $\left(X_{n,h} - \widehat{X}_{n,h}\right)/\sigma$ est de loi $\mathcal{N}(0;1)$,, il suffit de choisir $\Delta/\sigma = 2,58$, à partir d'une table de loi normale, pour que l'inégalité ci-dessus soit vérifiée.

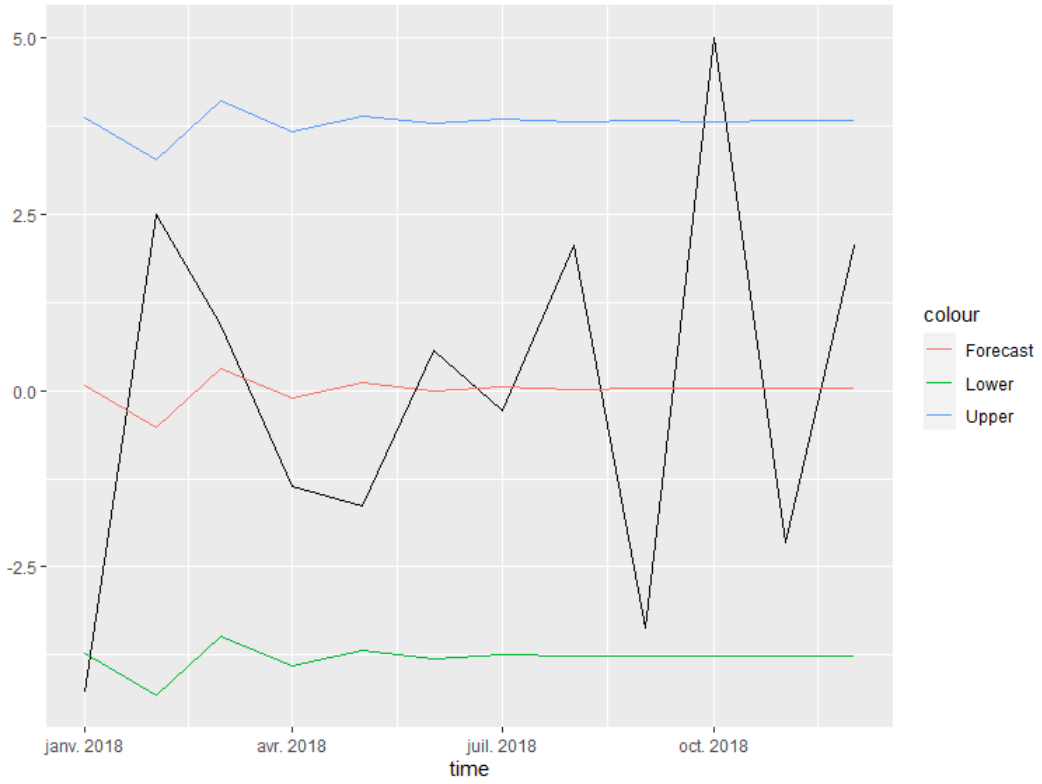


FIGURE 3.11 – Prévision grâce aux modèles ARMA sur la série stationnaire

Nous arrivons à obtenir un ordre d'idées quant à l'intervalle de valeur que peut prendre notre série stationnaire mais l'adéquation reste faible vis-à-vis de nos observations. Par la suite nous allons introduire un modèle qui reprend l'étude sur le processus stationnaire en appliquant une modélisation sur X_t.

Modèles ARIMA et SARIMA

Une fois que nous avons étudié la partie stationnaire de notre processus X_t, on souhaite revenir et pouvoir étudier ce dernier. Ce sont des généralisations des processus ARMA aux cas non stationnaires, avec tendance polynômiale $(ARIMA)$ ou avec une saisonnalité $(SARIMA)$. Ce sont les processus directement utilisés par R.
Notre processus $(X_t)_{t\geq 0}$ est un processus $ARIMA(p,d,q)$ si le processus $Y_t = \Delta_1^d X_t$ est un processus $ARMA(p,q)$

Théorème 2. *soit un processus y admettant une tendance polynomiale d'ordre k :*

$$y_t = \sum_{j=0}^{k} a_j t^j + \varepsilon_t$$

alors le processus Δy admet une tendance polynomiale d'ordre $k-1$

Avec les notations des modèles ARMA, nous pouvons remarquer que $\Delta X_t = (1-L)X_t$ et plus généralement $\Delta^d X_t = (1-L)^d X_t$. Un processus ARIMA est défini ainsi :
Un processus stationnaire X_t admet une représentation ARIMA $(\mathrm{p}, \mathrm{d}, \mathrm{q})$ minimale s'il satisfait

$$\Phi(L)(1-L)^d X_t = \Theta(L)\varepsilon_t, \quad \forall t \in \mathbf{Z}$$

avec les conditions suivantes :

- $\phi_p \neq 0$ et $\theta_q \neq 0$
- Φ et Θ, polynômes de degrés resp. p et q, n'ont pas de racines communes et leurs racines sont de modules > 1
- ε_t est un bruit blanc de variance σ^2

Cependant, notre série temporelle contient une composante saisonnière. En effet notre processus $(X_t)_t \geq 0$ est un processus $SARIMA(p,d,q,T)$ si le processus $Y_t = \Delta_T \circ \Delta_1^d X_t$ est un processus $ARMA(p,q)$
Les processus $SARIMA(p,d,q,T)$ sont donc bien adaptés à notre étude qui présente une saisonnalité de période T et qui ont une tendance polynômiale de degré $d-1$

Résultat pour l'ensemble de portefeuille

Nous allons appliqué ce processus $SARIMA(p,d,q,T)$ sur $\mathbb{R}$ sur l'ensemble de notre portefeuille. Nous arrivons à obtenir une prédiction qui prend en compte la tendance, la composante saisonnière et la composante bruit blanc.

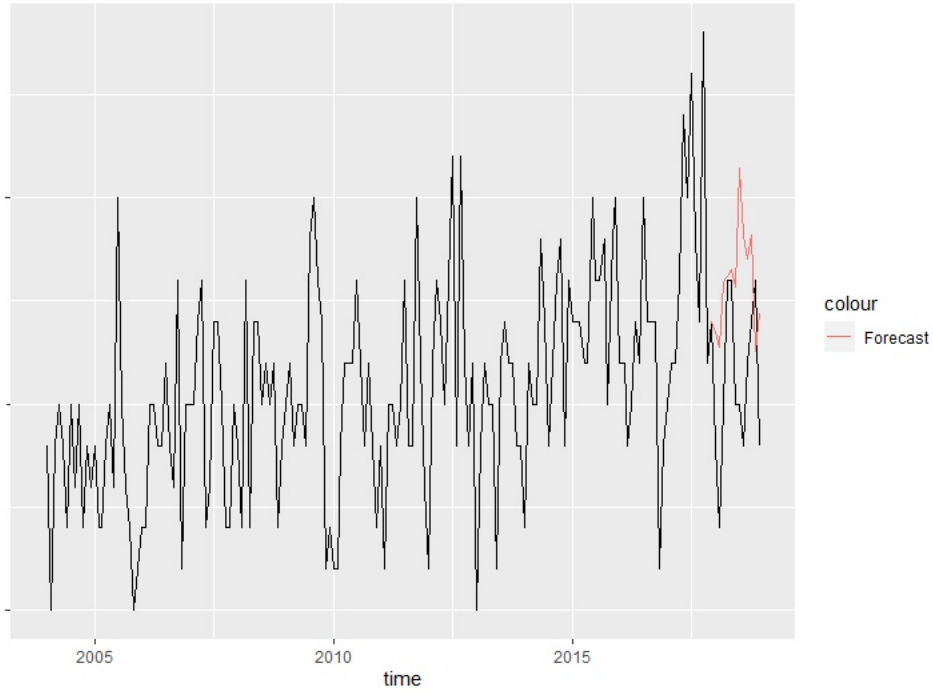


FIGURE 3.12 – Prévision grâce au modèle Sarima sur la série d'étude

D'après la série temporelle, les valeurs observées en 2018 sont faibles vis-à-vis des observations passées. Obtenir une prédiction qui surestime ses valeurs est rassurante. Les estimations obtenues nous paraissent acceptables mais s'appliquent sur l'ensemble des portefeuilles de partenaire. Cela ne nous aidera pas à répondre de manière claire à notre problématique. Nous cherchons donc à appliquer les méthodes que l'on a vues plus tôt sur le portefeuille d'un seul partenaire.

Application de la méthode à un partenaire

L'idée ici va être d'appliquer les mêmes résultats pour nos partenaires. Nous allons présenter les résultats pour un partenaire possédant un portefeuille historique assez important.

En suivant les mêmes étapes que précédemment, le modèle retenu pour la prédiction est un modèle $SARIMA(0, 1, 1, 12)$, bien que les hypothèses sur les innovations ne sont pas validées. En effet, même s'il n'existe pas d'autocorrélation significative dans la série des innovations, l'hypothèse de normalité n'est pas accepté ici. Nous la représentons ci-dessous.

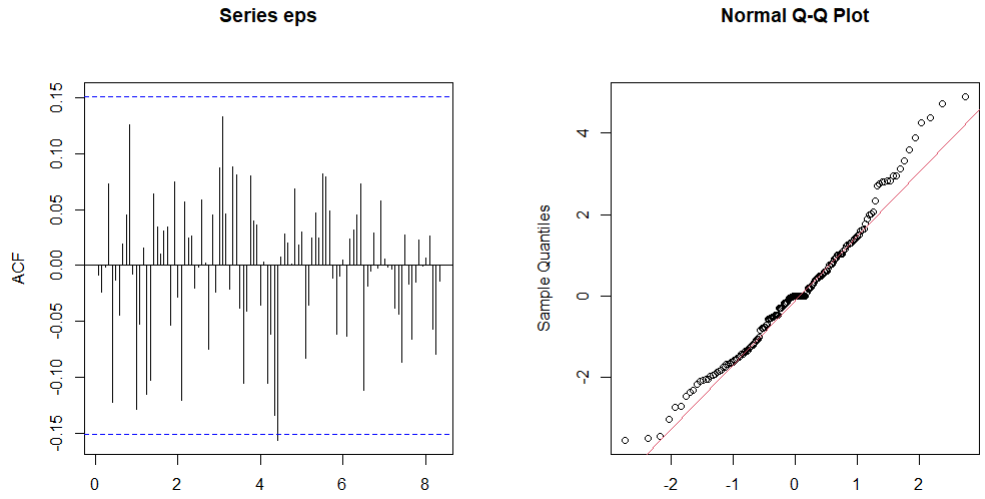


FIGURE 3.13 – Diagnotiques sur les innovations

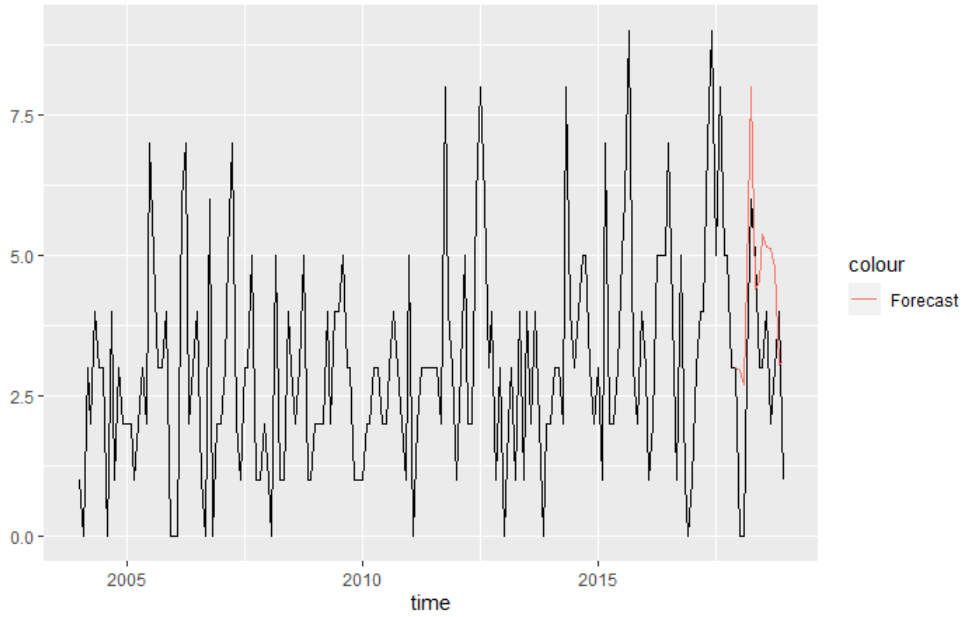


FIGURE 3.14 – Prévision grâce au modèle Sarima sur la série d'un partenaire

On surestime encore nos observations mais les résultats des tests sur les innovations montrent que le modèle s'ajuste mal à nos données surtout, en regardant le quantile plot normal, au niveau de la queue de distribution. L'hypothèse de normalité n'est pas vérifiée.

L'approche par processus ARIMA est une approche intéressante à considérer car elle nous a permis d'obtenir des prévisions robustes sur l'ensemble du portefeuille. Cependant, plus la maille de modélisation est fine plus notre analyse perd en robustesse et ne nous permet pas de répondre à la problématique à la maille des partenaires. En effet, l'omission de certaines informations, comme la dynamique de l'activité propre à chaque partenaire, ne nous permet par d'appliquer cette méthode dans le cas d'un partenaire dont l'activité est instable. Par conséquent, nous nous détournons de l'approche ARIMA pour introduire la notion de modélisation linéaire généralisée qui, d'un point de vue théorique, permettra de répondre à notre problématique.

3.2 Modèles linéaires généralisés

L'objectif des modèles linéaires généralisés est de modéliser la relation mathématique entre une variable réponse Y_i et ses variables explicatives $X_{i,j}$;

Fondements théoriques

Dans cette partie, nous allons rappeler quelques fondements théoriques de la modélisation linéaire généralisée, on rappellera en annexe les principes du modèle linéaire gaussien.

La famille exponentielle

Une variable aléatoire Y a une loi faisant partie de la famille exponentielle si sa densité peut se mettre sous la forme :

$$f(y, \theta, \phi) = \exp \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

Alors l'espérance et la variance d'une telle loi peut se mettre sous la forme :

$$E(Y) = b'(\theta) Var(Y) = a(\phi) b''(\theta)$$

avec θ le paramètre de position et ϕ le paramètre de dispersion. *Nous mettons quelques exemples de loi appartenant à la famille exponentielle en annexe de ce mémoire.*

Fonction de lien

La fonction de lien est une fonction monotone et dérivable qui nous autorise une dépendance non linéaire entre la variable réponse et les variables explicatives. En effet, notons $\mu_i = E(Y_i)$, alors

$$g(\mu_i) = \nu_i \mu_i = g^{-1}(\nu_i) = g^{-1}(x_i^t \beta)$$

Nous cherchons dont à modéliser une transformation de l'espérance de la variable réponse. Il existe plusieurs possibilité de fonction de lien que nous rappelons dans le tableau suivant :

Identité	$g(x) = x$
Log	$g(x) = \ln(x)$
Logit	$g(x) = \ln\left(\frac{x}{1-x}\right)$
Inverse	$g(x) = \frac{1}{x}$

Dans la suite, nous utiliserons la fonction de lien ln pour modéliser le nombre de sinistre.
Nous considérerons le modèle de régression suivant :

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i^t \beta = \eta_i, i \in \{1, \ldots, n\}$$

Avec :

1. n : Nombre d'observations,
2. p : Nombre de variables explicatives,
3. $\mu_i = \mathrm{E}[Y_i]$ où les Y_i sont les variables réponses à expliquer que l'on suppose indépendantes et non identiquement distribuées,
4. $\beta = (\beta_0, \ldots, \beta_p)^t$: Paramètres du modèle à estimer,
5. $(x_{1j}, \ldots, x_{nj})^t, j \in \{1, \ldots, p\}$: j^{me} variable explicative.

Estimation des paramètres par maximum de vraisemblance

L'expression de la vraisemblance s'écrit :

$$L(y_1, \cdot, y_n; \theta, \phi) = \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

Notons $L = L(y_1, \ldots, y_n; \theta_i, \phi)$, nous obtenons l'expression de la log-vraisemblance suivante :

$$\ln(L) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

Afin d'obtenir notre estimateur, il faut maximiser cette expression. Il faut donc la dérivée en fonction des paramètres β_j.

$$\frac{\partial}{\partial\beta_j}\ln(L) = \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)$$

Cas du modèle poissonien

La loi de Poisson est une loi de probabilité discrète. Si Y suit une loi de Poisson, la probabilité qu'il se réalise est :

$$P(Y = k) = e^{-\mu}\frac{\mu^k}{k!}$$

avec $k \in \mathbb{N}$ et $\mu > 0$. L'espérance et la variance d'une loi de Poisson sont identiques : $E(Y) = \mu$ et $V(Y) = \mu$.
Le modèle de Poisson modélise la variable aléatoire Y par une loi de Poisson de paramètre μ telle que $ln(\mu) = x'\beta$, avec x la matrice des variables explicatives.
μ :

$$L = \prod_{i=1}^{n} e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!}$$
$$\ln(L) = \sum_{i=1}^{n} -\mu_i + y_i\ln(\mu_i) - \ln(y_i!)$$
$$\frac{\partial}{\partial\beta_j}\ln(L) = \sum_{i=1}^{n} \frac{\partial}{\partial\mu_i}\{-\mu_i + y_i\ln(\mu_i) - \ln(y_i!)\}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}$$

Nous obtenons :

$$\frac{\partial}{\partial\beta_j}\ln(L) = \sum_{i=1}^{n}(y_i - \mu_i)x_{ij}$$

Ainsi les équations de vraisemblance sont :

$$\sum_{i=1}^{n} (y_i - \mu_i)\, x_{ij} = 0 \forall j = 1, \ldots, p$$

La résolution de ces équations requiert une méthode itérative telle que la méthode de Newton-Raphson que nous développerons en annexe.
Nous utilisons une régression de Poisson sur un modèle lorsque celui-ci possède une espérance égale à sa variance et la forme de la distribution ne dépend que du paramètre μ.
Remarque : plusieurs problèmes peuvent survenir lorsqu'on utilise cette méthode :

- une déviance beaucoup plus grande que le degré de liberté de la loi du χ^2, ce qui montre un problème d'ajustement ;
- une surdispersion (une variance des données supérieures à la variance théorique) des données pouvant entraîner :
  - de fausses estimations de l'erreur standard, de la statistique du χ^2 et de la p-value ;
  - une sous-estimation de l'erreur standard et une surestimation de la statistique du χ^2 ;
  - des paramètres non biaisés mais qui seront déclarés significatifs trop souvent.

La prise en compte de la dispersion, modèle binomial négatif

La modélisation d'une variable de comptage est souvent réalisée par le biais d'une loi de Poisson. Cependant cette loi est construite sur une hypothèse forte et difficile à vérifier en pratique qui est l'équidispersion des données. En effet, on retrouve souvent la présence de sur-dispersion. Ainsi, il existe plusieurs modèles alternatifs permettant de prendre en compte ce phénomène de sur-dispersion. Dans cette étude, nous avons fait le choix du modèle binomial négatif :
La loi binomial négative est une loi discrète qui permet de calculer la probabilité d'avoir n succès parmi k expériences identiques et indépendantes. Si X suit une loi binomiale négative, la probabilité que X se réalise est :

$$P(X = k) = \binom{n + k - 1}{k} p^n (1 - p)^k$$

avec $n, k \in \mathbb{N}$ et $p \in [0, 1]$. L'espérance et la variance d'une loi binomiale négative valent : $E(X) = \frac{n}{p}$ et $V(X) = \frac{n(1-p)}{p^2}$
Le modèle Binomiale négative généralise le modèle de Poisson. Il permet d'éviter une surdispersion des données.
Nous supposons que $Y|\Theta$ suit une loi $\mathcal{P}(\lambda\Theta)$, où Θ suit une loi Gamma de paramètre identique α, nous obtenons la loi binomial négative.

Validation de notre modèle

Test sur la statistique de Pearson

Pour tester la qualité de l'ajustement, on peut utiliser le test de Pearson. Définissons la statistique de Pearson de manière générale :

$$\chi^2_P = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{\mathbb{V}(\hat{Y}_i)}$$

Chaque terme de la somme est un écart au carré entre la valeur observée et la valeur prédite divisée par la variance de y_i. Le χ^2 de Pearson normalisé vaut $\frac{\chi^2}{\phi}$ Cette statistique est distribuée approximativement selon une loi du χ^2 à n-p degrés de liberté.

Test sur les résidus de Pearson

L'approche la plus simple pour obtenir des résidus est de calculer la différence entre les valeurs observées et les valeurs prédites et de diviser cet écart par l'écart type des valeurs observées. Les résidus de Pearson suivent approximativement une loi normale centrée réduite. Nous considérons qu'un individu qui a un résidu $|p_i| \geq 2$ doit nécessiter une attention particulière.

Test de la déviance et de la déviance résiduelle

Après avoir estimé tous les paramètres, il faut vérifier que le modèle est bien ajusté aux données. Pour cela, nous pouvons comparer le modèle estimer avec un modèle plus général, appelé modèle saturé, qui a autant de paramètres que de variables explicatives, la même distribution et la même fonction lien que le modèle estimé. Le but est de comparer la vraisemblance de ces deux modèles. Posons L la vraisemblance du modèle saturé et l celle du modèle estimé.

Pour cela nous allons utiliser :

— la déviance du modèle :

$$\Delta = 2\ln(\lambda) = 2(L - l)$$

$$\Delta = 2\sum_{i=1}^{n} \frac{Y_i(\Theta_i - \theta_i) + a(\Theta_i) - a(\theta_i)}{\phi}$$

— les déviances résiduelles :

$$\delta_i = sign(y_i - \mu_i)\sqrt{\delta_i^2}$$

où

$$\delta_i^2 = 2\frac{Y_i(\Theta_i - \theta_i) + a(\Theta_i) - a(\theta_i)}{\phi}$$

Si toutes les hypothèses du modèle sont vérifiées, Δ converge en loi vers une loi du χ^2 à n-p degrés de liberté, où p est le nombre de paramètres à estimer.
Concernant les résidus de déviance, les observations telles que $\delta_i \geq 2$ peuvent indiquer un défaut d'ajustement.

Les critères AIC et BIC

Comme pour les modèles sur série temporelle, les critères AIC et BIC permettent de comparer les modèles entre eux. Ici nous utiliserons le critère de AIC qui est défni par la formule suivante :

$$AIC = -2ln(L) + 2k$$

où L est la vraisemblance maximisée et k le nombre de paramètres.

Bootstrap

Une autre méthode de validation de modèle est le Bootstrap. Elle est utilisée pour estimer et valider la distribution du modèle lorsque que nous ne connaissons pas la distribution théorique ou quand la taille de l'échantillon est trop petite pour une étude statistique simple.
Il existe deux manières d'utiliser le bootstrap avec des modèles statistiques :

1. Ajustez le modèle de nombreuses fois en sélectionnant les observations à inclure au hasard avec remplacement, de sorte que certains points de données soient exclus et d'autres apparaissent plus d'une fois dans un ajustement de modèle particulier.
2. Ajustez le modèle une fois et calculez les résidus et les valeurs ajustées, puis mélangez les résidus plusieurs fois et ajoutez-les aux valeurs ajustées dans différentes permutations, en adaptant le modèle aux nombreux ensembles de données différents. Dans les deux cas, vous obtiendrez une distribution des valeurs de paramètres pour le modèle à partir de laquelle vous pourrez dériver des intervalles de confiance.

Nous cherchons à faire une prédiction sur chaque point de notre modèle. Et donc à créer un intervalle de confiance pour le prédicteur, qui va dépendre de l'erreur d'estimation des paramètres.

Application

Base d'étude

Pour modéliser nos sinistres avec un GLM, nous allons utiliser les résultats du premier chapitre afin de distinguer le nombre de sinistres graves aux sinistres attritionnels.
Nous divisons notre base en deux bases distinctes :

1. une base d'apprentissage, correspondant aux années d'exercice 2004 à 2016 ;
2. une base de test, correspondant à l'année d'exercice 2018

Pour notre base apprentissage, nous utilisons comme variables explicatives les informations annuelles connues du partenaire (primes acquises, nombre d'assuré aux portefeuilles, taux de résiliations, taux d'affaires nouvelles et le type de produit commercialisé) ainsi que le nombre de sinistres attritionnels observés et connues à date. Dans l'outil Excel produit, il est important de pouvoir choisir et modifier le nombre de variables explicatives pour le modèle, ainsi on laisse la liberté à l'utilisateur de modifier à sa guise la base de données.

Modélisation

Dans un premier temps, nous observons l'adéquation du nombre de sinistres à une loi de Poisson et une loi binomiale négative :

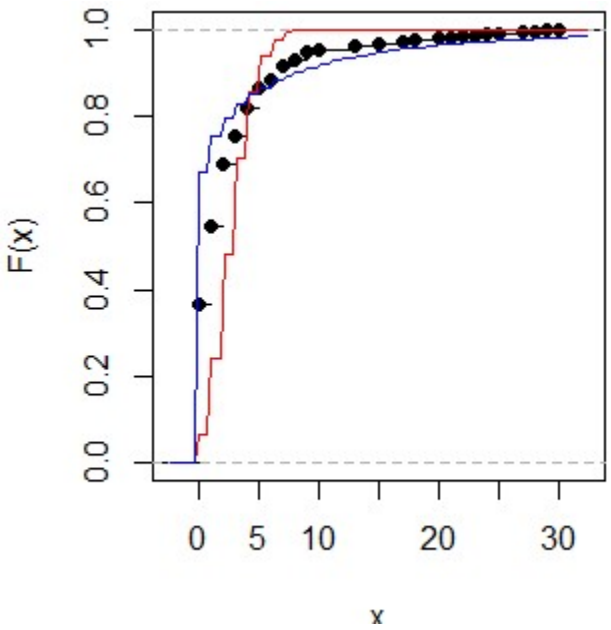


FIGURE 3.15 – Adéquation de nos donnée avec *en rouge* une loi de poisson et en *en bleu* une loi binomiale négative

Nous pouvons observer ici qu'en corrigeant la dispersion, nous obtenons une meilleure approximation de nos données par la loi binomiale négative. Pour apporter un argument autre que graphique, nous testons l'adéquation de nos données à une loi de Poisson et binomiale négative grâce à un test du χ^2. Ce test sera détaillé en annexe.

Test du χ^2 pour les sinistres graves			
Distribution	Statistique du χ^2	degrés de liberté	p-value
Poisson	138 178	29	0
Binomiale négative	39.9	28	0.066

TABLE 3.1 – Résultat du test d'adéquation du khi deux pour les sinistres graves

Au vu des résultats, il semble que la loi binomiale négative semble mieux s'adapter à nos données que la loi de Poisson. Cependant, nous continuerons, par la suite à tester et comparer les deux modèles.

Sortie du modèle poisson

Deviance Residuals :

Min	$1Q$	$Median$	$3Q$	Max
-117.089	-16.001	-2.440	7.438	120.817

Null deviance : 720916 on 216 degrees of freedom
Residual deviance : 167993 on 212 degrees of freedom
AIC :647 138
Number of Fisher scoring iterations : 4

Notre *Null Deviance* est très grande, cela signifie que le modèle Null, modèle qui est caractérisé par aucun effet facteur, n'explique pas assez bien les données. De même avec la *Residual deviance*, le modèle proposé explique très mal et n'est pas adapté à nos données.

Sortie du modèle binomial négatif

Deviance Residuals :

Min	$1Q$	$Median$	$3Q$	Max
-118.4	-15.9	-5	7.297	85.483

Null deviance : 711274 on 216 degrees of freedom
Residual deviance : 133816 on 212 degrees of freedom
AIC : 633585
Number of Fisher scoring iterations : 1

Après avoir modéliser nos sinistres, nous comparons les résultats obtenus en fonction du nombre de sinistres graves de la base test.

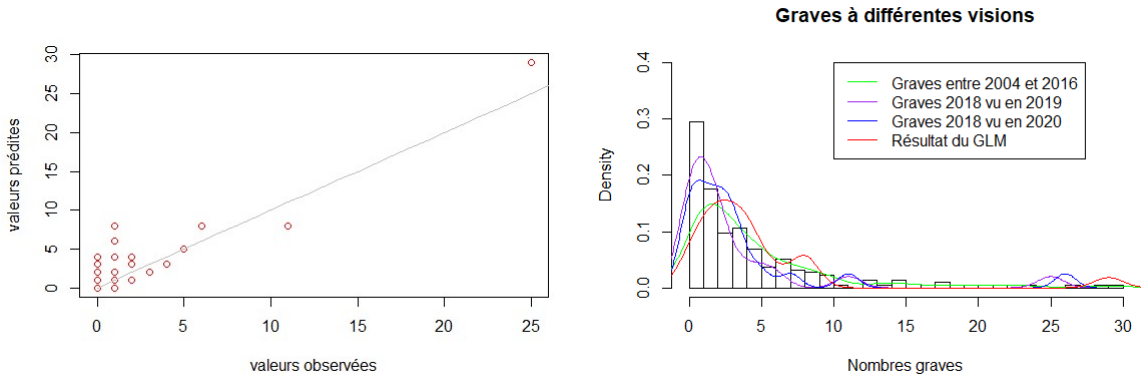


FIGURE 3.16 – Résultat des sinistres graves, modèle poisson

Dans le premier graphique, nous représentons les résultats de notre modèle, estimés à partir des variables explicatives de la base test, en fonction des valeurs que l'on observe à fin 2018. Nous traçons également une courbe $1-1$ afin de déterminer la qualité d'ajustement de ces résultats graphiquement. Dans le deuxième graphique, nous observons la distribution du nombre de sinistres observés toutes années confondues et nous obtenons les densités suivantes :

1. la distribution, en vert, du nombre de sinistres graves entre 2004 et 2016.

2. la distribution, en violet, des graves de l'exercice 2018 vues en 2019.
3. la distribution, en bleu, des graves de l'exercice 2018 vues en 2019.
4. les résultats du GLM.

Nous pouvons voir ici que le modèle surestime un peu le nombre de sinistres pour des partenaires avec peu de sinistralité. En regardant les différentes visions, nous pouvons voir que la distribution change entre 2019 et 2020 ce qui est due au type de garantie.
Nous réalisons la même approche que pour le modèle binomial négatif.

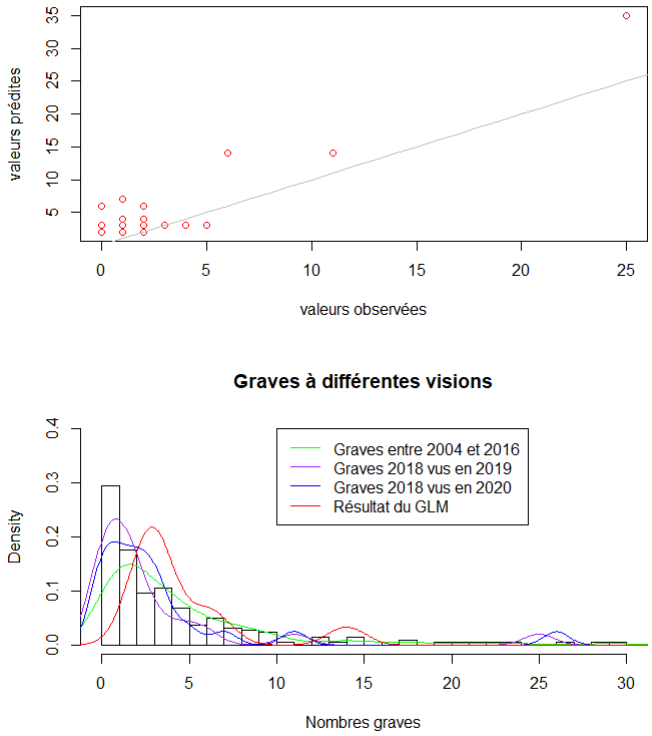


FIGURE 3.17 – Résultat des sinistres graves, modèle binomial négatif

La sinistralité semble être surestimée, de manière plus importante que pour le modèle poissonien et en regardant la distribution du nombre de sinistres graves, le modèle a tendance à surestimer nos nombres de manière plus générale.
D'un point de vue opérationnel, le modèle binomial négatif se veut plus prudent que le modèle de poisson, voir trop prudent et peut sur évaluer la sinistralité attendue et réellement observée au moment des revalorisations.
En regardant le critère du BIC pour ces deux modèles, le modèle retenu devrait être celui qui minimise le BIC.

Modèle	BIC
Binomiale négative	633 609
Poisson	647 155

Ici, nous constatons que le modèle choisi par le critère de l'AIC est le modèle binomial négatif. C'est-à-dire le modèle le plus prudent.

Enfin, nous traçons un graphique des résidus de déviance jackknife par rapport aux valeurs ajustées et un graphique QQ normal des résidus de déviance normalisés.

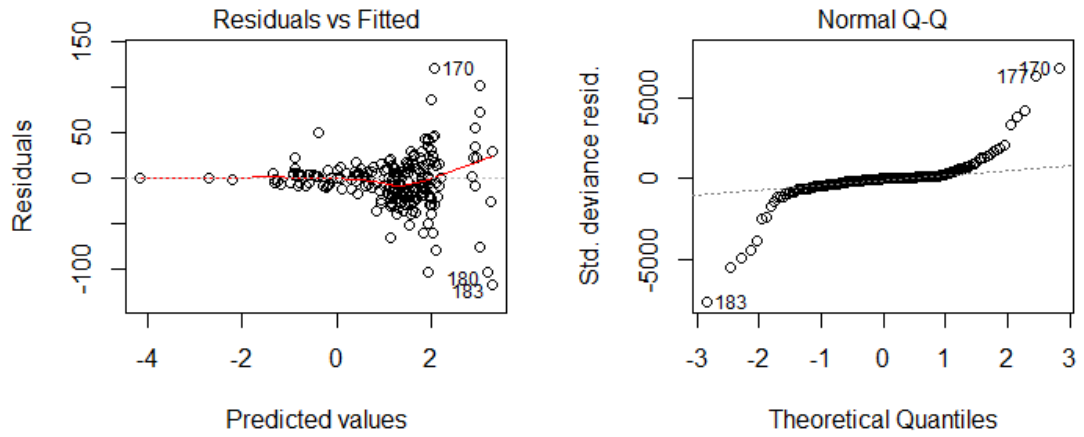


FIGURE 3.18 – Analyse du modèle poisson sur les sinistres graves

Nous observons, dans le premier graphe, que plus nous allons vers la droite plus la variance augmente, ce qui traduit l'existence potentielle d'hétéroscédasticité dans nos résidus. De plus, le modèle fait référence à trois valeurs aberrantes. Enfin, en regardant le graphique à droite, il n'y a pas une bonne adéquation de nos résidus de déviance à une loi normale.

Nous allons faire la même étude sur notre modèle binomial négatif :

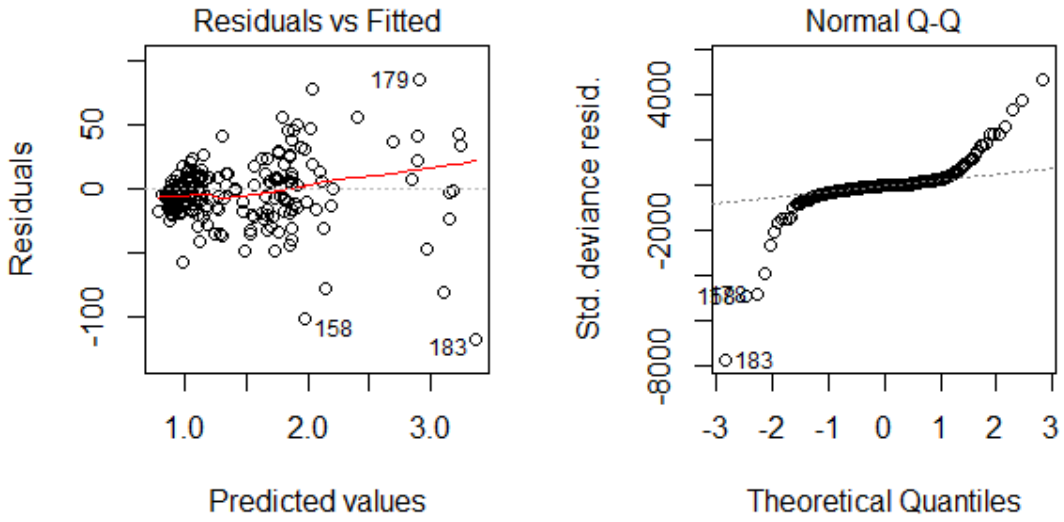


FIGURE 3.19 – Analyse du modèle binomial négatif sur les sinistres graves

Nous pouvons voir, dans le premier graphe, que plus on va vers la droite plus la variance augmente, ce qui traduit l'existence potentielle d'hétéroscédasticité dans nos résidus. De plus, le modèle fait référence à trois valeurs aberrantes, la valeur 183 est aberrante dans les deux modèles de régression. Enfin, en regardant le graphique à droite, il n'y a pas une bonne adéquation de nos résidus de déviance à une loi normale.
Enfin, nous avons appliqué une analyse par bootstrap afin de voir la qualité d'ajustement de notre modèle et d'obtenir un intervalle de confiance pour notre prédicteur.

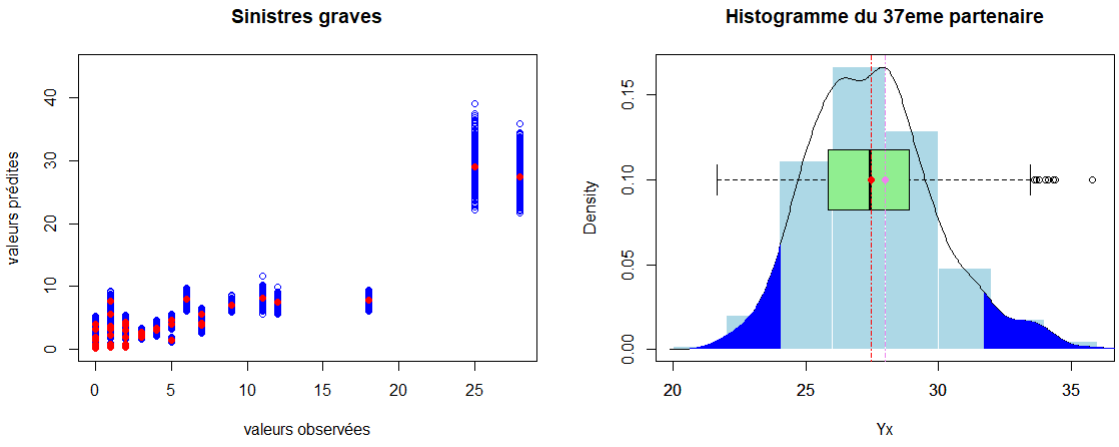


FIGURE 3.20 – Résultats du bootrap appliqué pour les graves au GLM poisson

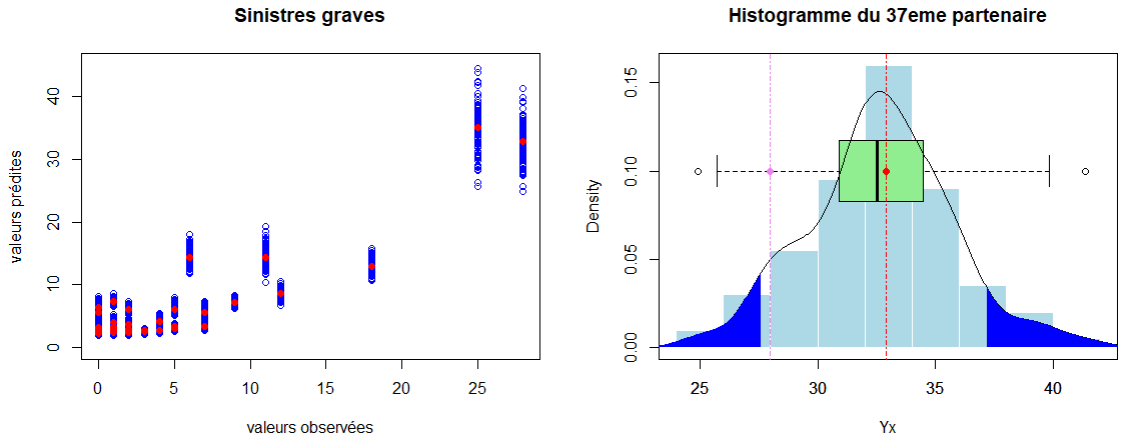


FIGURE 3.21 – Résultats du bootrap appliqué pour les graves au GLM binomiale négatif

Moins il y a de données similaires dans notre base d'apprentissage, plus l'intervalle de confiance est grand. Cependant, nous avons examiné la distribution des prédictions pour un partenaire pour lequel une sinistralité annuelle importante : Nous pouvons voir que l'intervalle de confiance reste faible, de l'ordre de +/- 5 sinistres graves en plus, compte tenu du nombre de sinistres qu'il peut avoir chaque année. Dans les deux modèles la valeur observée *(en rose)* est inclue dans l'intervalle de confiance, avec une prédiction plus précise pour le modèle poisson.
L'approche par GLM nous a permis d'approcher notre variable à une maille assez fine, en regroupant les plus petits partenaires en amont de la modélisation, on arrive à un résultat robuste. Cependant, les hypothèses sur la déviance et les résidus de notre modèle n'ont pas été vérifiées et incorporent un biais dans la modélisation. Par conséquent, nous allons considérer une approche différente de celle présentée et tourner notre réflexion vers la théorie bayésienne.

3.3 Théorie bayésienne

La tarification automobile est établie sur la loi des grands nombres. En mutualisant les risques, les actuaires peuvent estimer la charge de sinistralité attendue. Il n'est pas possible d'utiliser cette notion sur une seule observation, mais nous pouvons lui affecter une charge collective. Nous pouvons prendre le cas d'un assuré sans sinistre, la loi des grands nombres lui affectera une prime nulle.
Whitney (1918) est le premier à proposer l'approche de précision de par

> *la nécessité, par équité envers le risque individuel, de pondérer l'expérience du groupe d'une part et l'expérience individuelle d'autre part.*

Ainsi l'assureur peut réclamer la prime de crédibilité

$$p_j = (1-\alpha)p_{groupe} + \alpha\bar{p}_j$$

où α est le facteur de crédibilité et $\bar{p}_j$ la charge annuelle moyenne de sinistre de l'assuré j.

La théorie de Mowbray, la crédibilité complète

Avant Whitney, Arthur H. Mowbray, en 1914, définit une prime pure "'fiable"' comme

> *Une prime pour laquelle la probabilité est forte qu'elle ne diffère pas de la vraie prime par plus d'une limite arbitraire*

Ce qui signifie que :

$$Pr\left[(1-k)E[S] \leq S \leq (1+k)E[S]\right] \geq p$$

- avec k petit (généralement 5%)
- p est près de 1, habituellement, 0.90 0.95 ou 0.99
- S représente l'expérience du contrat.

Dans ce cas de figure, un contrat d'assurance est considéré crédible si son expérience est stable. Il y a cependant peu d'applications légitimes de la crédibilité complète.

La théorie de Whitney, la crédibilité partielle

Whitney, en 1918, propose de pondérer l'expérience individuelle et la prime collective par un facteur de crédibilité z en une prime de la forme

$$\pi = zS + (1-z)m$$

Il apparaît alors la nécessité de pouvoir tenir compte en partie de l'expérience individuelle d'un contrat se trouvant sous le seuil de crédibilité.
La formule pour le facteur de crédibilité proposé par Whitney est :

$$z = \frac{n}{n+K}$$

où K est une constante déterminée au jugement pour une meilleure stabilité du modèle afin d'éviter les fluctuations d'une année à l'autre. Cependant, le facteur de crédibilité n'est basé que sur un facteur taille n, ce qui rend la tarification peu précise et peu fiable. Pour ces raisons, la théorie de Whitney fut abandonnée pour la crédibilité bayésienne.

Approche bayésienne

Cas continu

Considérons le partenaire j auquel correspond la charge annuelle grave, $Y_{j,1}, ..., Y_{j,T}$ et le paramètre de risque Θ_j. Ce paramètre de risque n'est pas observable, il est supposé constant dans le temps. On suppose :

- la loi marginale de Θ_j admet une densité $\mu(\theta_j)$
- les variables $Y_{j,1}, ..., Y_{j,T}$ sont, conditionnellement à Θ_j, indépendantes entre elles et distribuées identiquement avec pour densité $f(y_{jt}|\theta_j)$

On obtient ainsi que les variables $Y_{j,1}, ..., Y_{j,T}, \Theta_j$ admettent pour densité :

$$\mu(\theta_j) \prod_{t=1}^{T} f(y_{j,t}|\theta_j)$$

L'utilisation de la règle de Bayes nous permet, à partir de l'information disponible en date T, d'enrichir notre connaissance de Θ_j et de $Y_{j,T+1}$
La densité prédictive pour le risque futur $Y_{j,T+1}$ est donnée par :

$$f(y_{j,T+1}|\sigma\{Y_{j,1}, ..., Y_{j,T}\}) = \frac{\int \mu(\theta) \prod_{t=1}^{T+1} f(y_{j,t}|\theta) d\theta}{\int \mu(\theta) \prod_{t=1}^{T} f(y_{j,t}|\theta) d\theta}$$

L'estimateur de Bayes pour la période $[T, T+1]$ est donné par :

$$E[Y_{j,T+1}|Y_{j,1} = y_{j,1}, ..., Y_{j,T} = y_{j,T}] = \int y_{j,T+1} f(y_{j,T+1}|\sigma\{Y_{j,1}, ..., Y_{j,T}\})$$

Cet estimateur est optimal au sens des moindres carrés.

Cas discret

Dans le cas discret, on considère le partenaire j auquel correspond le nombre annuel de sinistres graves, $Y_{j,1}, ..., Y_{j,T}$ et le paramètre de risque Φ_j. Si la variable aléatoire Φ_j prend que des valeurs discrètes, la distribution a priori est exprimée par $Pr[\Phi_j = \phi_j]$.
L'estimateur bayésien minimisant l'erreur quadratique moyenne est défini par :

$$E[Y_{j,T+1}|Y_{j,1} = k_{j,1}, ..., Y_{j,T} = k_{j,T}] = \sum k_{j,T+1} Pr[k_{j,T+1}|\sigma\{Y_{j,1}, ..., Y_{j,T}\}]$$

Modèle de Bühlmann, crédibilité de précision

Bühlmann restreint l'approximation de cet estimateur aux fonctions linéaires des observations, c'est-à-dire de la forme :

$$c_0 + \sum_{t=1}^{n} c_t Y_{it}$$

Chaque partenaire est caractérisé par :
- un niveau de risque θ_i réalisation d'une variable aléatoire Θ_i
- des observations $(Y_{i,1}, ...Y_{i,n}) \equiv Y_i$

Les hypothèses du modèle de Bühlmann sont les suivantes :

1. Les partenaires (Θ_i, Y_i), $i = 1, ..., I$ sont indépendants, les variables aléatoires $\Theta_i, ...\Theta_I$ sont identiquement distribuées et les variables Y_{it} ont une variance finie.
2. Les variables aléatoires Y_{it} sont telles que

$$E[Y_{it}|\Theta_i] = \mu(\Theta_i) \;\; i = 1, ...I$$

$$Cov(Y_{it}, Y_{ik}|\Theta_i) = \delta_{tk}\sigma^2(\Theta_i), \;\; t, k = 1, ..., n$$

La première hypothèse définit l'indépendance inter partenaires. La deuxième représente l'homogénéité temporelle et "indépendance" intra partenaires :

— $\mu(\Theta_i)$ constante dans le temps
— observations conditionnellement non corrélées

Théorème 3. *Pour un portefeuille tel qu'illustré et sous les deux hypothèses introduites précédemment, la meilleure approximation linéaire non homogène $\mu(\Theta_i)$ est*

$$\pi^B_{i,n+1} = \alpha \bar{Y}_i + (1-\alpha) E[Y_{j,n}]$$

Le facteur de crédibilité est défini par $\alpha = \frac{n}{n+K}$, $K = \frac{E[\sigma^2]}{Cov(\Theta_{i,n+1}, Y_{i,t})} = \frac{s^2}{a}$.

Estimation des paramètres

Nous devons estimer les paramètres de structure du portefeuille :

1. $m = E[\mu(\theta)]$, moyenne du portefeuille ;
2. $s^2 = E[\sigma^2(\Theta)]$, variabilité moyenne du portefeuille, homogénéité temporelle ;
3. $a = Var[\mu(\Theta)]$, variance entre les moyennes des partenaires, homogénéité du portefeuille.

Nous obtenons alors notre estimateur empirique en remplaçant chaque paramètre inconnu par son estimateur :

$$\hat{\pi}^B_{i,n+1} = \hat{\alpha} \bar{Y}_i + (1-\hat{\alpha}) \hat{m} \hat{z} = \frac{n}{n + \hat{s}^2/\hat{a}}$$

où

$$\hat{m} = \frac{1}{In} \sum_{i=1}^{I} \sum_{t=1}^{n} Y_{it}$$

$$\hat{s}^2 = \frac{1}{I(n-1)} \sum_{i=1}^{I} \sum_{t=1}^{n} (Y_{it} - \bar{Y}_i)^2$$

$$\hat{a} = \frac{1}{I-1} \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 - \frac{1}{n} \hat{s}^2$$

Modèle de Bühlmann-Straub

Le modèle de Bühlmann-Straub est une généralisation du modèle de Bühlmann intégrant des pondérations sur les observations.
Dans sa forme la plus générale, on associe un poids ω_{it} à chaque donnée notée S_it et définie par :

$$S_{it} = \frac{Y_{it}}{\omega_t}$$

Ainsi, nous nous attendons à ce que l'expérience d'un partenaire possédant un plus gros portefeuille d'assuré soit plus stable dans le temps qu'un partenaire possédant un plus petit portefeuille. Pour que le modèle reflète cette idée, les hypothèses sont le suivantes :

1. Les partenaires (Θ_i, S_i), $i = 1, ..., I$ sont indépendants, les variables aléatoires $\Theta_i, ...\Theta_I$ sont identiquement distribuées et les variables S_{it} ont une variance finie.
2. Les variables aléatoires S_{it} sont telles que

$$E[S_{it}|\Theta_i] = \mu(\Theta_i) \;\; i = 1, ...I$$

$$Cov(S_{it}, S_{ik}|\Theta_i) = \delta_{tk}\frac{\sigma^2(\Theta_i)}{\omega_{it}}, \;\; t, k = 1, ..., n$$

.

Nous obtenons

$$Var[S_{it}|\Theta_i] = \frac{\sigma^2(\Theta_i)}{\omega_{it}}$$

Tout comme le modèle de Bühlmann, pour un portefeuille tel qu'illustré et sous les deux hypothèses introduites précédemment, la meilleure approximation linéaire non homogène de $S_{i,n+1}$ est

$$\pi^{BS}_{i,n+1} = \alpha_i \bar{S}_{iw} + (1 - \alpha_i)E[S_{i,n}]$$

où

$$\alpha_i = \frac{\omega_{i\Sigma}}{\omega_{i\Sigma} + K} = \frac{\sum_{t=1}^{n} \omega_{it}}{\sum_{t=1}^{n} \omega_{it} + K}, \;\; K = \frac{s^2}{a}$$

De plus, on a :

$$s^2 = E[Var[S_{it}|\Theta_i]]\omega_{it}$$

$$Cov(S_{it}, S_{i\omega}) = a + \frac{s^2}{w_{i\Sigma}}$$

Estimation des paramètres

Comme pour le modèle de Bühlmann, nous devons estimer les paramètres de structure du portefeuille, soit m,s^2,et a.
Pour m, l'estimateur intuitif de m est

$$S_{\omega\omega} = \sum_{i=1}^{I} \frac{\omega_{i\Sigma}}{\omega_{\Sigma\Sigma}} S_{i\omega} = \sum_{i=1}^{I}\sum_{t=1}^{n} \frac{\omega_{it}}{\omega_{\Sigma\Sigma}} S_{it}$$

où

$$\omega_{\Sigma\Sigma} = \sum_{i=1}^{I}\sum_{t=1}^{n} \omega_{it}$$

Pour s^2, nous obtenons l'estimateur

$$\hat{s}^2 = \frac{1}{I(n-1)} \sum_{i=1}^{I}\sum_{t=1}^{n} \omega_{it}(S_{it} - \bar{S}_{i\omega})^2$$

$$\hat{a} = \frac{\omega_{\Sigma\Sigma}}{\omega_{\Sigma\Sigma}^2 - \sum_{i=1}^{I} \omega_{i\Sigma}^2} \sum_{i=1}^{I} \omega_{i\Sigma}(S_{i\omega} - \bar{S}_{\omega\omega})^2 - (I-1)\hat{s}^2$$

Application à nos données

Dans un premier temps, nous allons déterminer la sinistralité future de chaque partenaire en prenant en compte sa sinistralité historique. Nous segmenterons chaque partenaire par produit d'assurance ce qui nous permettra de détecter les performances du partenaire x produit. Le modèle que nous utiliserons dans cette étude sera basé sur le modèle de Bühlmann-Straub en prenant en compte la prime acquise par le partenaire. Contrairement à l'approche par modélisation généralisée, nous choisissons de modéliser le nombre de sinistres à la maille la plus fine, c'est-à-dire par partenaire, ce qui signifie qu'il n'y a pas de regroupement en amont de la modélisation.

Modèle de Bühlmann-Straub

Dans un premier temps nous allons calculer le ratio S/P de chaque partenaire et estimer notre modèle à partir de ce dernier. Dans un deuxième temps on fait une prédiction à partir des paramètres des modèles pour chaque partenaire. Enfin, nous multiplions cette prédiction par la prime acquise annuelle du partenaire en 2018.
Nous montrons ici les résultats de notre sortie de modèle :
Structure Parameters Estimators
collective Estimator : 4.5011

Partenaires	*Moy.ind.*	*Poids*	*Fact.Cred.*	*Estim.*	*Obs.*
1	12.46	535290277	0.996647	12.43	6
7	5.12	29710859	0.942852	5.09	4
13	2.90	80462093	0.978109	2.94	6
14	0	0	0.000000	4.50	0
15	0	19030	0.010457	4.45	0
19	2	8249694	0.820823	2.45	0

Pour chaque partenaire, on obtient la moyenne individuelle, le poids, le facteur de crédibilité et le nombre de sinistres crédibilisé. Lorsque l'on n'a pas d'information sur le partenaire (expérience nulle) le modèle va automatiquement choisir le facteur de crédibilité de groupe. Le modèle différencie bien un acteur avec de l'expérience dont la somme du nombre de sinistres grave est nulle d'un acteur sans expérience, respectivement les partenaires 15 et 14.

Comme pour les modèles linéaires généralisés, on va tracer la prédiction du modèle en fonction des valeurs observées en 2018.

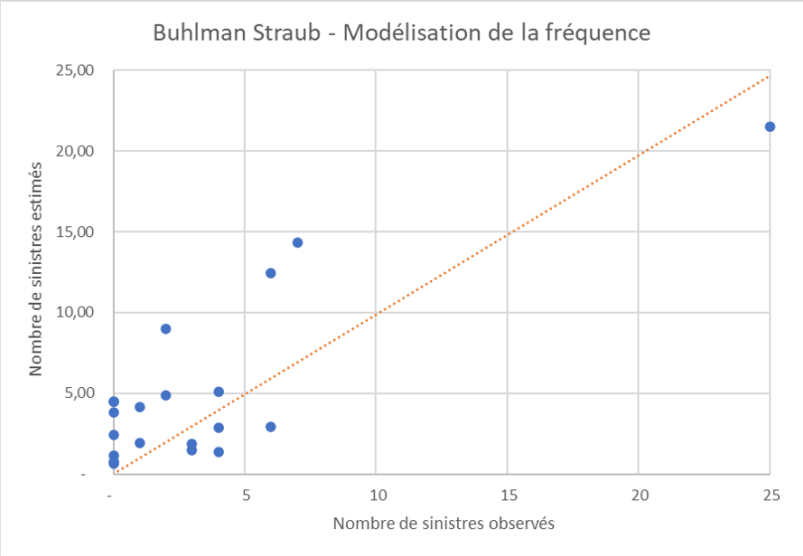


FIGURE 3.22 – Résultat de l'approche bayésienne, poids utilisés : la prime acquise annuelle

Nous pouvons voir une concentration autour de la droite 1-1. Cependant, certaines valeurs vont être beaucoup trop surestimées. Nous pouvons observer que pour certaines valeurs, nous estimons 4 fois plus de sinistres graves ce qui pourraient peser lourd sur les résultats annuels du partenaire

Approche par partenaire x produit, modèle de Jewell

Les modèles hiérarchiques introduits par Jewell prennent en considération la subdivision du portefeuille en catégorie et en sous catégories. L'idée serait de prendre en compte le produit auquel appartiennent les données exploitées.

Sur une période de N années, nous observons un portefeuille de p partenaires. Ce portefeuille est divisé en catégorie C_i, représentant le type de produit. Nous nous intéressons à la sinistralité de chaque partenaire k composant le portefeuille. Ainsi, on note $X_{i,k,t}$ l'observation de sinistralité d'un partenaire k auquel on associe un poids $W_{i,k,t}$ correspondant à la prime acquise du partenaire l'année t.

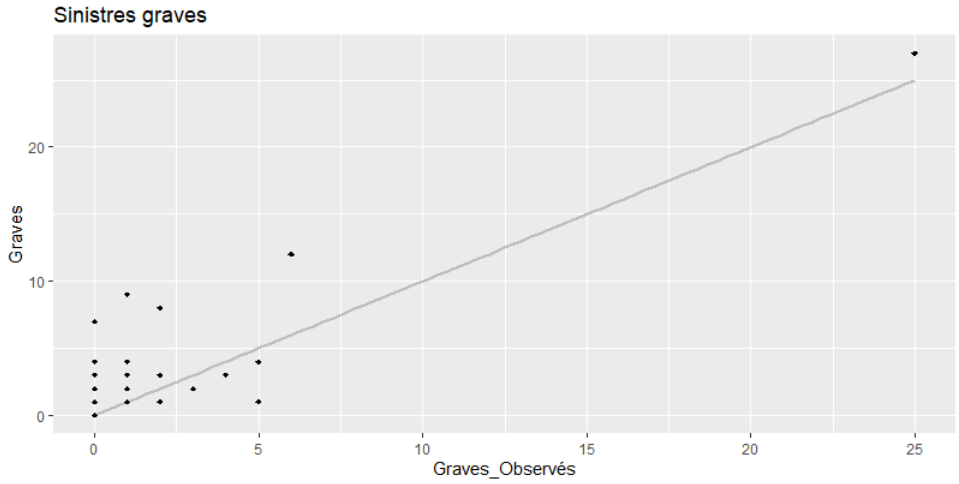


FIGURE 3.23 – Résultat de l'approche bayésienne avec segmentation partenaire x produit

D'après les résultats de prédiction, le produit est une variable importante à prendre en compte dans notre modélisation. En effet, d'un point de vue statistique, le modèle va prédire de manière plus précise et, d'un point de vue opérationnel, l'analyse et les négociations se feront par type de produits commercialisés par le partenaire.

Nous appliquons le *modèle à un niveau* au portefeuille afin de déterminer les facteurs de crédibilité relatifs aux types de produits commercialisés par le partenaire.

Modèle pour les graves	
Produit	facteur de crédibilité
Auto standard	0.424
Auto malussée	0.356
2 Roues	0.338

Les résultats de crédibilité du modèle sur le produit montrent qu'il y a plus de partenaire "stable" positionné sur le marché de l'auto standard et moins de partenaire "stable" sur le marché du deux roues.

Chapitre 4

Modélisation de la charge grave

Sommaire

4.1 Modèle Fréquence x Cout

En assurance auto, et de façon générale en assurance non vie, l'approche Fréquence x Coût moyen est largement utilisée. Les modèles les plus utilisés sont les modèles linéaires généralisés et les modèles de Machine Learning, où des variables explicatives sont requises.

Les survenances des sinistres étant supposées aléatoires, d'autres outils de modélisation basés sur les approches probabilistes peuvent être utilisés sans variables explicatives.

Comme vue précédemment, la théorie de la crédibilité est aussi une méthode largement utilisée en tarification actuarielle qui permet de tarifer un contrat au sein d'un portefeuille en se basant sur l'historique des sinistres du contrat mais aussi sur l'historique de groupe. La prime associée est alors composée d'une part relative à la sinistralité individuelle, et d'une autre à la sinistralité globale du portefeuille.

Dans notre étude, nous allons suivre cette approche Fréquence x Coût moyen en nous concentrant sur les différents modèles proposés ci-dessus.

La fréquence est le nombre de sinistres divisés par l'exposition pour un groupe de police d'assurance commercialisé par le partenaire, dans notre étude, l'exposition est la même pour toute. Pour modéliser le nombre de sinistres, on retrouve le plus souvent les lois de Poisson et Binomiale négative. Nous avons vu, dans le chapitre précédent, plusieurs méthodes de modélisation de la fréquence de nos sinistres graves.

Le coût moyen de sinistres représente les indemnités moyennes versées (ou qui reste à verser) à une tierce personne. Ce coût moyen est très volatile du fait de la présence des sinistres graves.

Ainsi, **la charge annuelle de sinistralité** pour un partenaire est obtenue en multipliant la fréquence par le coût moyen.
Soit C_i^T la charge sinistre du partenaire i durant l'exercice T, soit N la variable aléatoire à valeurs entières représentant le nombre de sinistres affectant le partenaire durant l'exercice T, on définit la suite $X^T = (X_k^T)_{k\geq 0}$, variables aléatoires réelles représentant le coût individuel d'un sinistre, avec $X_{0,i}^T = 0$.

On a alors,

$$C_i^T = \sum_{k=0}^{N} X_{k,i}^T$$

Remarques : Le modèle est facilement applicable sous réserve de deux hypothèses fortes :

- Indépendance entre la fréquence et le coût des sinistres ;
- Indépendance et stationnarité des montants de sinistres.

La charge sinistre

Pour notre étude, nous allons décomposer la charge sinistre pour nos trois intervalles de coût de sinistres, attritionnels $A_{k,i}^T$, graves $G_{k,i}^T$.
On a alors :

$$C_i^T = \sum_{k=0}^{N_A} A_{k,i}^T + \sum_{k=0}^{N_G} G_{k,i}^T$$

avec $A^T_{0,i} = G^T_{0,i} = 0$, N_A et N_G les variables aléatoires représentant respectivement les sinistres attritionnels et graves.

Sachant qu'on se trouve dans une étude de modélisation de la charge grave pendant les périodes de revalorisation. On va considérer qu'environ 80% de l'ultime des sinistres attritionnels est connu à date, puisque la date de revalorisation diffère entre chaque partenaire.

Visualisation du coût moyen historique

Ici, on représente mensuellement le coût moyen des sinistres graves tout portefeuille confondu.

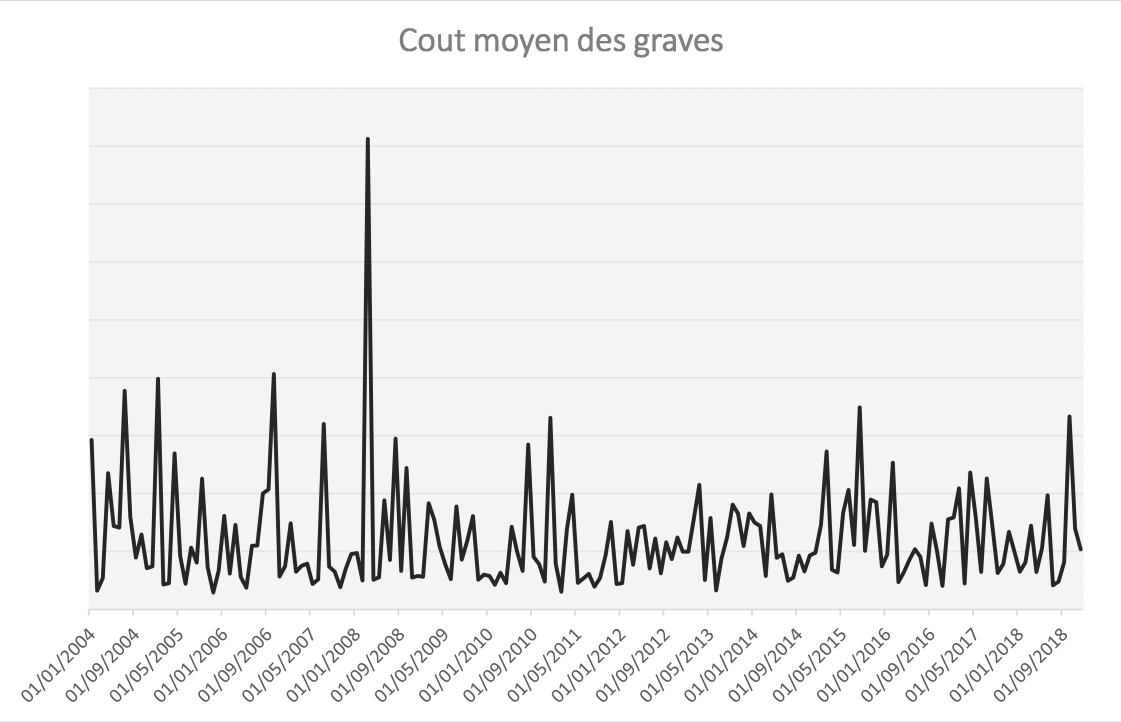


FIGURE 4.1 – Coût moyen grave, visualisation mensuelle

L'idée ici est d'obtenir une estimation des graves sur 2018, cependant il sera difficile de capter un sinistre atypique comme on peut l'observer pour 2008, le pic étant très éloigné de la tendance générale.

Méthode de Monte-Carlo

La méthode de Monte-Carlo consiste à simuler un grand nombre de scénarios à partir des lois de fréquences et de coût individuel des sinistres. Elle se décompose selon les étapes suivantes :

1. on tire, dans la loi du nombre de sinistres, un nombre de sinistres N ;

2. puis, N fois de suite, on tire, dans la loi du montant des sinistres, un montant de sinistres S ;

3. pour obtenir une réalisation de la loi de la charge de sinistres, il suffit de faire la somme des N nombres S tirés.

La simulation d'un scénario permet de générer les variables aléatoires de la charge individuelle de sinistralité grave.

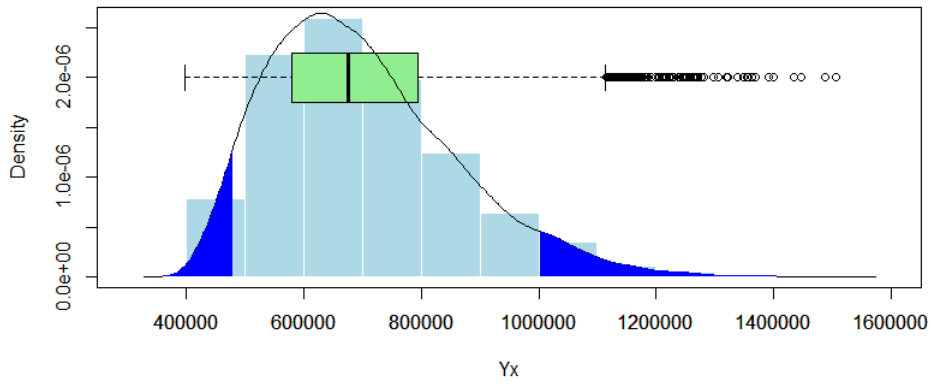


FIGURE 4.2 – Illustration de la méthode Monte Carlo, à l'aide d'un histogramme

Dans l'exemple ci-dessus, nous générons la charge de sinistre grave annuelle d'un partenaire à partir de la distribution de Pareto Généralisée. Ainsi, nous allons pouvoir obtenir des intervalles de confiance et les quantiles de cette distribution. L'idée était de ne pas forcément se positionner sur la moyenne des charges mais plutôt de surévaluer ou sousévaluer le risque en fonction de la connaissance par un expert de l'activité du partenaire.

Application

Nous allons regarder de plus près les résultats de la méthode de modélisation suivante :

1. Une approche par modélisation linéaire généralisée pour la fréquence.

2. Une approche par Monte Carlo pour les coûts graves.

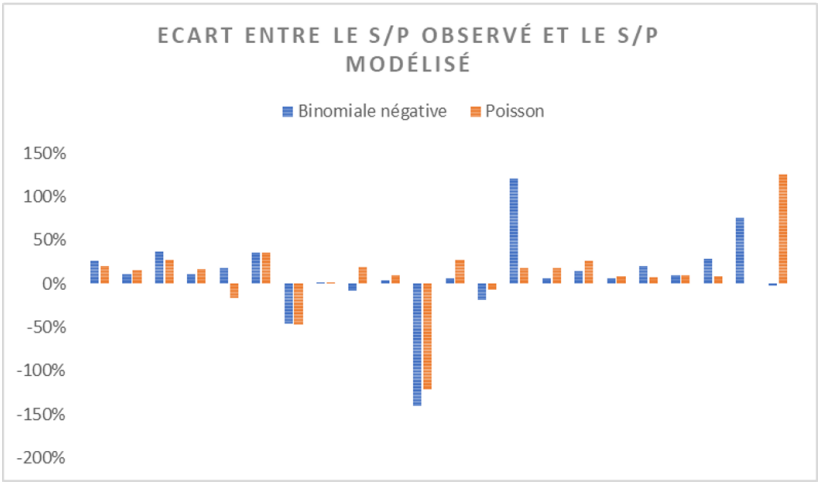


FIGURE 4.3 – Ecart entre le ratio de sinistralité modélisé et observé pour la modélisation des nombres par GLM

Nous observons que d'une année à l'autre le ratio d'un partenaire peut baisser ou augmenter drastiquement. Le modèle a du mal à capter ce changement. Maintenant, nous allons regarder les résultats pour :

1. Une approche bayésienne pour la fréquence
2. Une approche par Monte Carlo pour les coûts graves

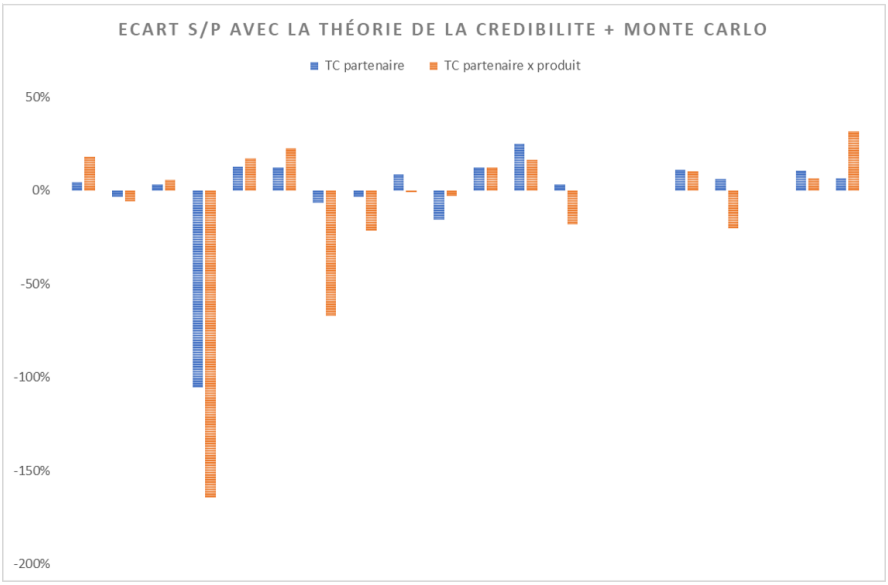


FIGURE 4.4 – Ecart entre le ratio de sinistralité modélisé et observé pour la modélisation des nombres par le bayésien

4.2 Modélisation avec une approche purement bayésienne

Dans cette partie, nous allons observer les résultats obtenus sur le coût des sinistres par une modélisation bayésienne. Dans un premier temps, on a réalisé une modélisation du coût annuelle des sinistres graves par l'approche Bühlmann-Straub en définissant la prime acquise par le partenaire comme poids puis une modélisation par l'approche de Jewell en réalisant une hiérarchisation par produit commercialisé.

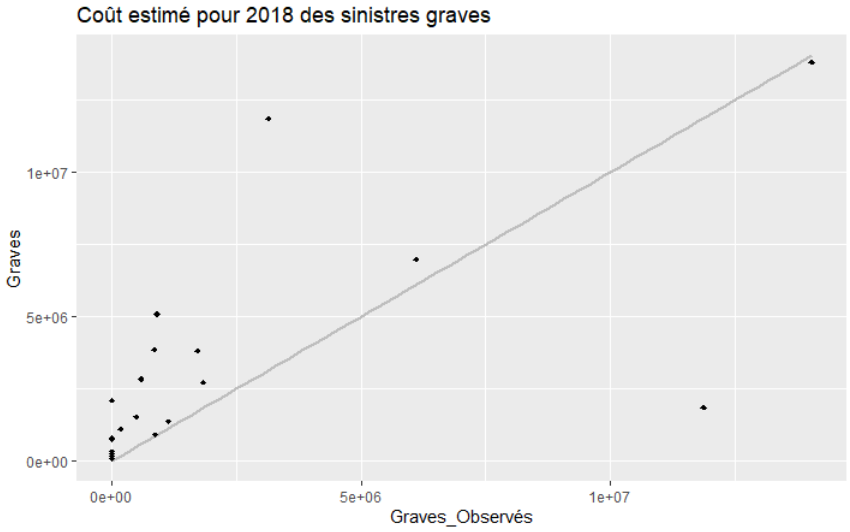


FIGURE 4.5 – Résultat avec un modèle de Bühlmann-Straub, en prenant en compte la prime acquise par partenaire

Nous constatons que pour le modèle on surestime le coût annuel des sinistres graves mais les résultats semblent intéressant à exploiter, puisque nous ne nous considérons pas à l'ultime de nos valeurs observées. En comparant avec l'approche de Jewell, dont la maille de modélisation est beaucoup plus fine puisque les estimations résultent d'un approche par partenaire x produit, on peut voir une trop grande instabilité :

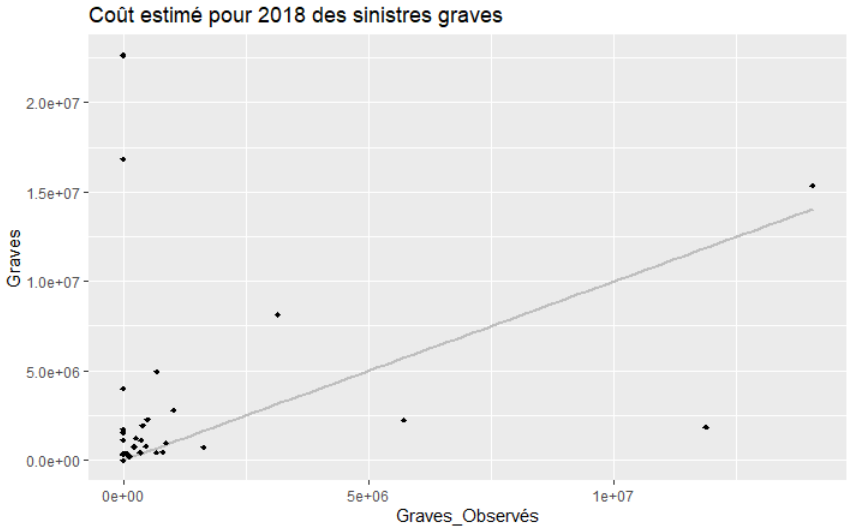


FIGURE 4.6 – Résultat avec un modèle de Jewell, en prenant en compte la prime acquise par partenaire x produit

En effet, le modèle est moins bien précis à cette maille pour plusieurs raisons :

- La maille étant beaucoup trop fine, on a du mal à obtenir une stabilité pour un partenaire x produit et nous perdons en robustesse.
- certains produits peuvent arrêter d'être commercialisé et donc passer en run off, donc il reste encore de la prime acquise mais le ratio est modélisé sur une dynamique différente.
- Certains produits vont être transférés vers un autre et nous détecterons mal la sinistralité.
- En comparant pour un partenaire avec une grande expérience globale présent sur les trois produits mais une grande hétérogénéïté entre ces derniers, nous obtenons une sous évaluation de 3% avec ce modèle contre une surévaluation de 7% avec le premier modèle.

Dans le graphique ci-dessous, nous comparons les écarts de S/P pour chaque partenaire entre le ratio modélisé et le ratio observé, pour chaque type de modélisation.

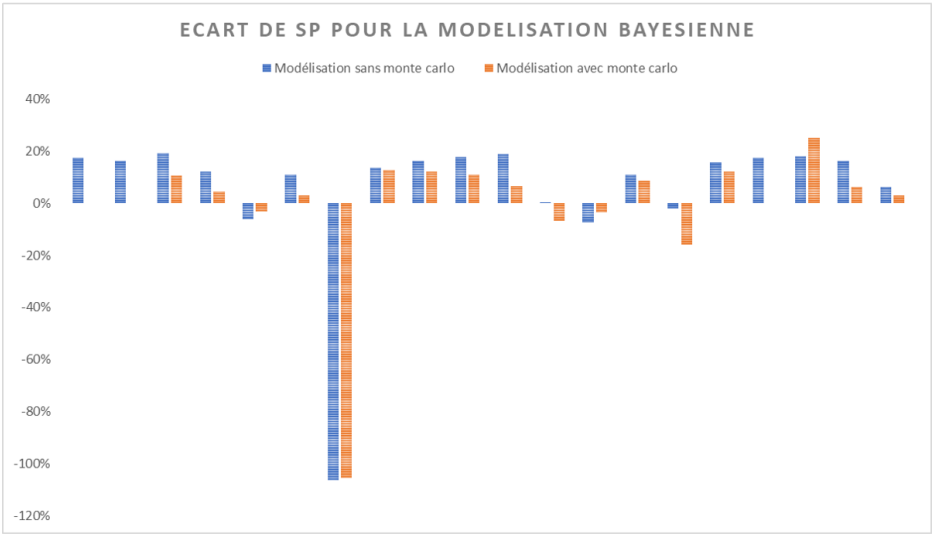


FIGURE 4.7 – Ecart entre le ratio de sinistralité modélisé

4.3 Application opérationnelle

Quand un nouveau partenariat est négocié avec un courtier, un tarificateur propre à ses besoins et à sa demande est construit par l'équipe. Il pourra ainsi intégrer à son portefeuille de nouveaux contrats. Cependant, chaque partenaire possède un profil d'assurés différent, une cartographie différente et une politique de distribution différente. Ainsi, il existe une très grande hétérogénéité parmi les partenaires pour un même type de produit.
Un moyen simple d'évaluer la rentabilité annuelle d'un partenaire est de calculer son ratio de sinistres sur primes, acquises lors de l'exercice. Le ratio S/P étant très volatile, il est peu acceptable d'évaluer les performances en cours à partir du seul ratio de l'année précédente.
Les résultats obtenus à l'ultime pour l'année 2018 vont être comparés au S/P observé par partenaires à fin 2018. Pour les partenaires regroupés, nous allons comparer chaque ratio observé avec le S/P obtenu pour le groupe.
Ainsi,

1. si $S/P_{retenue} \leq$ S/P$_{reel}$: Cela signifie que l'historique dudit partenaire montre que ce dernier a fait preuve d'une sinistralité qui était sous-estimée a priori. La prise en considération de cet historique pourra être traduite par une majoration de la prime après analyse des cas observés par modalité de contrat. Nous pouvons illustrer ce cas pour un partenaire en particulier où nous observons un écart de plus de 100% du ratio. Nous sommes confrontés à une véritable problématique car l'impact des graves sur le ratio de sinistralité n'a été observé qu'un an après sa survenance. Ce cas de sinistralité n'est pas isolé, il peut être intéressant, par la suite, de regrouper les éventuelles informations complémentaires contenues dans les bases contrats de nos partenaires et de répondre aussi aux questions suivantes :
   — A quel coût s'élève chaque sinistre en moyenne et le comparer à l'historique du partenaire.

— Quels coefficients ajustés dans ma tarification pour absorber ces pertes financières ? Plutôt que d'appliquer ces majorations à tout mon portefeuille, au risque de subir une vague de résiliation, il vaudrait mieux sélectionner le type de contrat impacté par la sinistralité ?

2. Si $S/P_{retenue} \geq$ S/P$_{reel}$: Cela signifie que l'historique dudit partenaire montre que ce dernier a fait preuve d'une sinistralité qui était surestimée a priori. La prise en considération de cet historique pourra être traduite par une réduction de prime après analyse des cas observés par modalité de contrat. Ce cas est majoritairement observé dans nos modélisations et il peut poser problème si :
   — La réduction est telle que le partenaire va voir son ratio de S/P se dégrader.
   — Nous observons une hausse inattendue de la sinistralité et que nous nous retrouvons à devoir réappliquer une majoration de prime.

Conclusion

Dans un marché très concurrentiel tel que celui de l'assurance automobile, la relation commerciale entre le partenaire et l'assureur est d'une importance majeure. Les résultats du partenaire font l'objet de négociations tarifaires chaque année. Il est donc important de mesurer l'impact de la sinistralité grave sur ces résultats et donc apporter une stratégie de commercialisation performante pour les années qui suivent. Cet impact doit être le plus précis possible pour ne pas sous-évaluer le risque futur. Une sur évaluation peut être risquée puisqu'elle peut amener, au pire des cas, à l'arrêt du partenariat ou de la commercialisation du produit.
Pour traiter cette problématique, il est intéressant de mettre à l'épreuve le seuil de sinistralité grave déjà en place à l'Équité. Par le biais de la théorie des valeurs extrêmes, l'étude réalisée nous a permis de déterminer un seuil et une loi sous-jacente à nos données.
Dans un premier temps, nous cherchons à modéliser ces données par une méthode fréquence x coût moyen. Nous appliquons différentes approches pour modéliser la fréquence de nos sinistres. Dans un deuxième temps, nous comparons cette modélisation avec une approche purement bayésienne.
Quelle que soit la méthode, la charge de sinistres graves est souvent surestimée par le modèle. En allant plus en détail sur ces résultats, cette surestimation peut être due aux points suivants :

1. le délai de consolidation peut prendre plus de temps et nous n'observons pas, à vision ultime, nos sinistres ;
2. l'année de référence, 2018, est considérée comme une " bonne" année dans l'ensemble contrairement à l'année 2017 qui a connu un pique de sinistralité important.

Ces deux points énumérés, ci-dessus, posent question sur les limites de nos modèles. L'intérêt d'une méthode séparée pour la fréquence et le coût moyen est que nous ne pouvons pas déterminer d'où peut provenir cette baisse de sinistralité, est-elle due à une baisse de la fréquence ou de la charge individuelle. Dans les deux cas, à différentes visions, il faudra rester prudent sur la survenance tardive des sinistres graves, due à une ouverture de dossier ou d'une réévaluation à la hausse du coût.

Bibliographie

[M1] Y. LUO. *Amelioration de la modélisation de sinistres grave à l'aide d'une approche d'apprentissage*, ISFA, mémoire présenté le 26 Avril 2016.

[M2] M. VEGNI. *Modélisation du coût des sinistres extrêmes en assurance automobile*, ISUP, mémoire présenté en 2011.

[M3] Y. XU. *Modélisation des sinistres graves dans l'assurance santé internationale*, ESSEC-ISUP, mémoire présenté en 2017.

[M4] T. DURAND. *Évaluation et optimisation de la rentabilité dun portefeuille automobile*, EURIA, mémoire présenté en 2016.

[M5] P. RINDER. *Modélisation des sinistres corporels en automobile*, ENSAE, mémoire présenté en 2016.

[A1] A. TRIVIERE. *Modélisation des sinistres corporels pour la tarification dun traité de Réassurance Automobile en Excédent de Sinistres*. Présentation 2018.

[A2] D. V. HINKLEY. *Bootstrap Methods and Their Application*. Article in Technimetrics, January 1997.

[C1] V. GOULET. *Mathématiques actuarielles IARD II, (Théorie de la crédibilité)*. Université Laval, 2010.

[M6] V. DESERT. *Rentabilité attendue des intermédiaires en assurance automobile* , CEA, mémoire présenté en 2015.

[M7] E. BONIN. *Méthodes de projection du risque santé : intérêt des séries temporelles* , IFSA, mémoire présenté en 2010.

[C2] M. PINHEIRO and R. GROTJAHN. *An Introduction to Extreme Value Statistics*.

[C3] V.MONBET. *Modèles linéaires généralisés*, IRMAR Université de Rennes 1.

[C4] A. BORCHANI. *Statistique des valeurs extrêmes dans le cas de lois discrètes.* 2010, 71, XV P. HAL00572559 .

[C5] Institut de veille sanitaire. *Série temporelle et modèles de régression* Département santé environnement.

Annexe A

Loi du maximum et Fischer-Tippett

A.1 Distribution des extrema

On définit le maximum d'un n-échantillon d'une variable aléatoire réelle X, dont les X_i sont considérées comme indépendantes et identiquement distribuées, par :

$$M_n = \max(X_i)_{1 \leq i \leq n}$$

Le but étant de déterminer la loi limite normalisée que suit M_n, on note F_X la fonction de répartition de la variable X.
Ainsi :

$$F_{M_n}(x) = P(M_n < x) = P\left(\bigcap_{i=1}^{n} X_i < x\right) = P(X_i < x)^n = F_X(x)^n$$

On s'intéresse à la distribution asymptotique du maximum, on obtient :

$$\lim_{n \to +\infty} F_{M_n}(x) = \lim_{n \to +\infty} F_X(x)^n = \begin{cases} 0 & \text{si } x < x_F \\ 1 & \text{sinon.} \end{cases}$$

avec $x_F = \sup(x \in \mathbb{R} : F_X(x) < 1)$, le point terminal à droite de la fontion F_X.
La distribution asymptotique du maximum donne une loi dégénérée, une masse de Dirac en x_F. Cela ne nous fournit pas assez d'information, d'où l'idée d'utiliser une transformation par standardisation.

A.2 Théorème de Fischer-Tippet

Théorème 4. *Supposons n variables aléatoires iid $X_1, ..., X_n$ de loi de distribution F. S'il existe deux suites réelles $(a_n)_{n \geq 1}$ et $(b_n)_{n \geq 1}$ avec $b_n \geq 0$, et une fonction de repartition dégénérée G telle que, $\frac{M_n - a_n}{b_n} \to G$ lorque n tend vers l'infini, alors G est du même type que l'une des trois lois suivantes :*

— *Loi de Fréchet,*

$$\Phi_\alpha(x) = \begin{cases} \exp(-x^{-\alpha}) & \textit{si } x>0 \\ 0 & \textit{sinon.} \end{cases}, \alpha>0$$

— *Loi de Weibull,*

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & \textit{si } x<0 \\ 1 & \textit{sinon.} \end{cases}, \alpha<0$$

— *Loi de Gumbel,*

$$\Upsilon(x) = \exp(-\exp(-x))$$

On dit alors que F_X appartient au domaine d'attraction maximum de G et on note $X \in MDA(G)$. Ces trois lois sont des cas particuliers de la forme la plus générale de la distribution des valeurs extrêmes notée GEV *Generalized Extreme Value Distribution*, en introduisant les paramètres de localisation μ et de dispersion σ :

$$G_{\mu,\sigma,\xi}(x) = \exp - \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}$$

avec ξ le paramètre de forme, ou l'indice de queue qui ne dépend que de la loi de X.
La connaissance de ce paramètre permet de caractériser à un changement d'échelle près, le comportement asymptotique du maximum normalisé.
Nous avons les correspondances suivantes :

— Fréchet, $\xi = \alpha^{-1} > 0$
— Gumbel, $\xi = 0$
— Weibull, $\xi = \alpha - 1 < 0$

Annexe B

Classification

B.1 Classification par k-means

Principe

Le K-means est un algorithme non supervisé de clustering non hiérarchique dont le but est de regrouper les observations similaires en K clusters différents. Chaque observation ne peut appartenir qu'à un seul cluster. L'ensemble des données peut être représenté sous forme de matrice où chaque ligne représente une observation :

$$\begin{matrix} x_{(1,1)} & x_{(1,2)} & x_{(1,3)} & \dots & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,3)} & \dots & x_{(2,n)} \\ & & \vdots & & \\ x_{(m,1)} & x_{(m,2)} & x_{(m,3)} & \dots & x_{(m,n)} \end{matrix}$$

L'algorithme commence par prendre K individus comme centre initiaux, la partition initiale étant obtenue par affectation des individus au centre le plus proche puis nous itérons :

- une étape de représentation qui consiste à définire un centroïd pour chaque classe, le centre de gravité
- une étape d'affectation ou l'on affecte chaque individu à la classe dont le centroide est le plus proche, ici au sens de la distance euclidienne.

Cet algorithme converge vers une partition réalisant un minimum local de l'intertie intra-classe, ou le plus grand pourcentage d'inertie expliqué.

Choix du nombre de clusters

La première étape consiste à choisir un bon nombre K de clusters. Pour cela, on lance l'algorithme plusieurs fois avec des valeurs de K différentes et on calcule la variance des clusters qui correspond à la distance entre les observations contenues dans le cluster et son centre appelé "centroïd" :

$$V = \sum_j \sum_i D(c_j, x_i^j)^2$$

où c_j représente le cluster du centroïd j et x_i représente la i-ème observation du cluster j. Graphiquement, le k optimal pour l'algorithme est la valeur située au "coude". Cette méthode est appelée "méthode du coude".

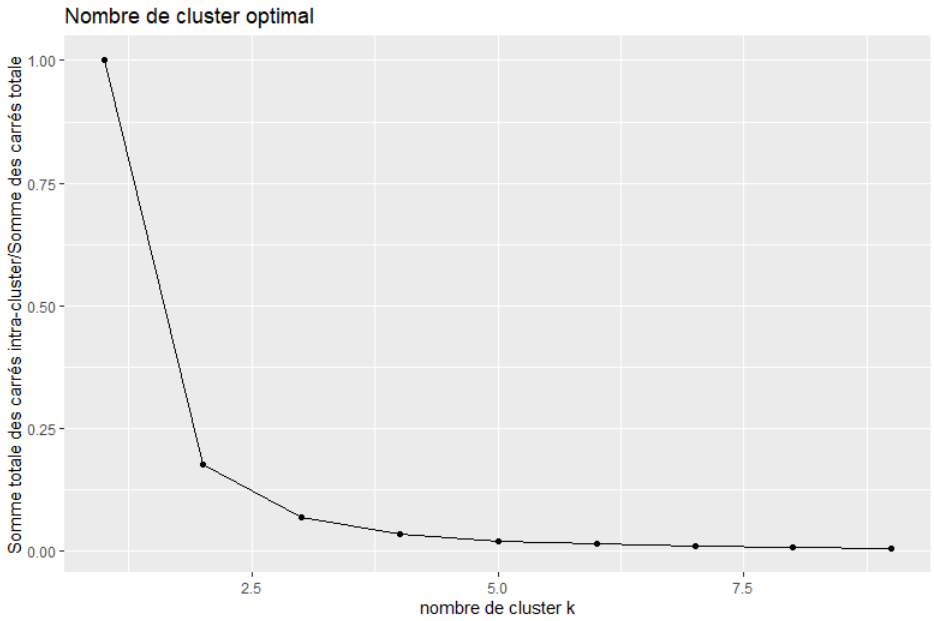


FIGURE B.1 – Méthode du coude, choix du k optimal application pour le produit deux roues

Le graphique ci-dessus représente le ratio entre la variance au sein des clusters sur la variance totale. Ce ratio diminue au fur et à mesure que le nombre de clusters augmente. Le coude que l'on observe à k=3 indique que les clusters supplémentaires ont peu de valeurs. Ainsi on ne présentera que 3 clusters dans nos applications.

Application à nos données

L'idée de base derrière ce clustering serait d'homogénéïser le portefeuille de partenaires avec lesquels on rencontre différentes problématiques :

- des partenaires possèdant un chiffre d'affaires croissant chaque année, les performances historiques ne reflèteront pas les performances futures ;
- des courtiers dont le partenariat est très jeune, donc pas assez d'informations passées ;
- des systèmes de distributions différents, parfois totalement digitalisés ;
- un cadre de souscription parfois différentes également ;
- des objectifs de rentabilité différente, voire propre à chaque partenaire.

Les informations utilisées sont :

- le chiffre d'affaires annuel ;
- le nombre d'assurés dans le portefeuille de chaque partenaire ;

— le nombre de sinistres en RC corporelle annuel du portefeuille ;
— le taux de sinistralité global du portefeuille ;
— le nombre annuel d'affaires nouvelles du partenaire ;
— le nombre annuel de résiliations ;
— la prime moyenne du partenaire ;

Les variables dont on dispose pour ce clustering sont proches des variables utilisées opérationnellement pour déterminer les objectifs du partenariat. De plus, on déterminera un clustering différent pour chaque produit commercialisé par les partenaires.

On visualise la matrice de corrélation des variables explicatives pour la classification K-means. On oberve une très forte corrélation entre le montant de primes acquises annuel, le nombre de sinistres et la taille du portefeuille du partenaire.

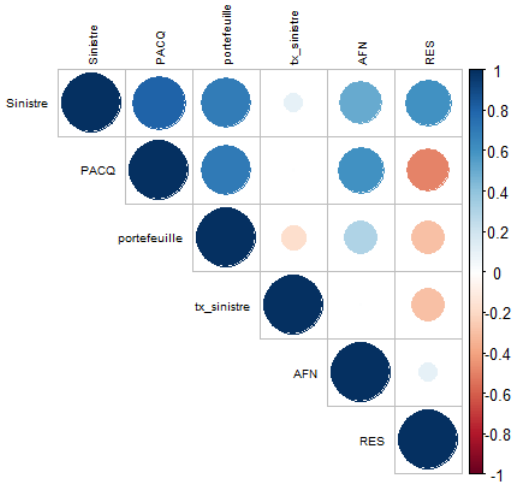


FIGURE B.2 – Matrice de correlation des variables explicatives

Afin de visualiser les données et les resultats de l'algorithme de K-means, on réduit le nombre de dimensions à l'aide d'une ACP, analyse en composantes principales qui génère deux variables à partir des initiales.

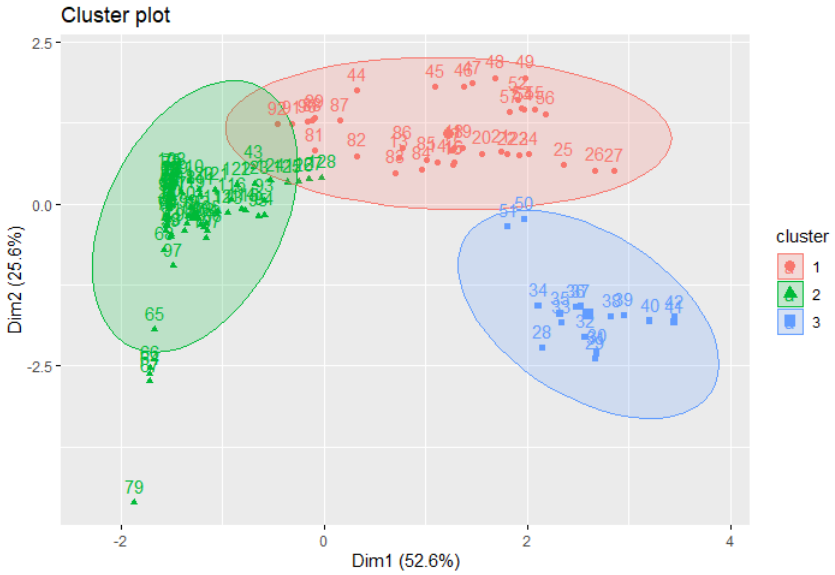


FIGURE B.3 – Resultat pour le deux roues de l'algorithme K-means après réduction de dimension

On distingue bien les trois groupes formés par l'algorithme des centres mobiles. Cependant, on voit bien que pour le groupe 2 les points sont très concentrés vers le centre de l'ellipse, ce sont en majorité les plus petits partenaires qui sont représentés dans ce groupe.

Catégorie	S/P en % 2018 vu en 2020	S/P en % historique vu en 2020
auto standard	**82**	**92**
1	85	96
2	63	81
3	53	75
auto malussée	**87**	**75**
1	73	68
2	126	76
3	69	78
Deux roues	**59**	**69**
1	60	61
2	76	69
3	55	73
GENERAL	78	80

FIGURE B.4 – Statistiques de performance par groupe, classification par l'algorithme de K-means

B.2 Classification Hiérarchique sur Composantes Principales

Principe

L'approche HCPC nous permet de combiner :

1. Les méthode des composantes principales ;
2. La classification ascendante hiérarchique ;

3. le partitionnement en k-means.

L'algorithme est le suivant :

1. On effectue une ACP puis on choisit le nombre de dimensions à retenir ;
2. On applique ensuite une classification hierarchique sur le résultat de l'ACP ;
3. On choisit le nombre de groupes en fonction du dendogramme obtenu après la classification. Un premier partitionnement est effectué ;
4. Ce partitionnement est perfectionné grâce à l'algorithme du k-means.

Application

Analyse en composante principale L'analyse en composantes principales permet d'analyser et de visualiser notre jeu de données. Chaque variable peut être considérée comme une dimension différente. Avec plus de 3 variables dans notre jeu de données, il est diffcile de les visualiser. Cette technique va permettre de synthétiser l'infomation en seulement quelques nouvelles variables, les composantes principales. Elles correspondent à une combinaison linéaire des variables originelles. Pour choisir le nombre de dimensions, on a tendance à regarder les premiers axes principaux afin de trouver des profils intéressants. Ici, on trace le graphique des valeurs propres, on remarque que les trois premières composantes principales expliquent 84% de la variation. Le nombre d'axes est déterminé par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables, c'est-à-dire trois axes.

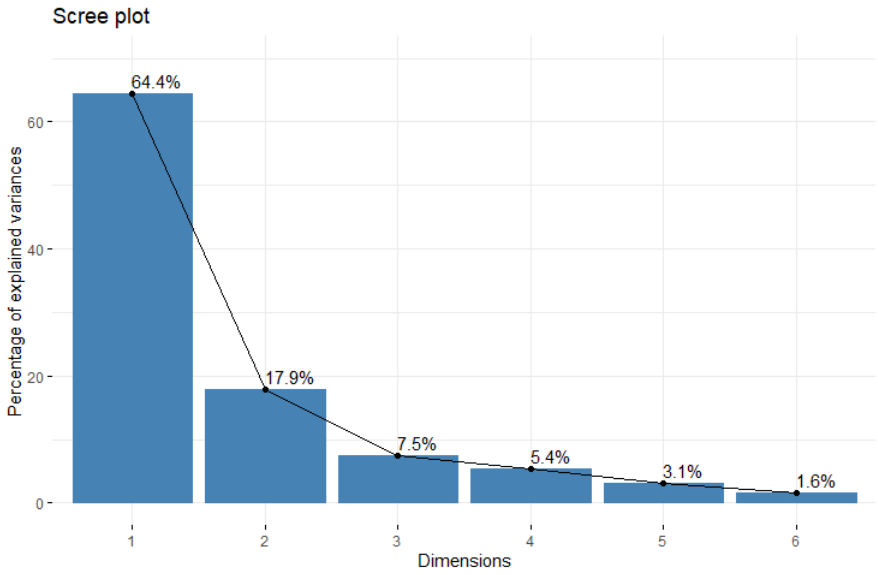


FIGURE B.5 – Première étape de la HCPC sur le produit deux roues, l'ACP

Le graphique ci-dessous est également connu sous le nom de graphique de corrélation des variables. Il montre les relations entre toutes les variables.

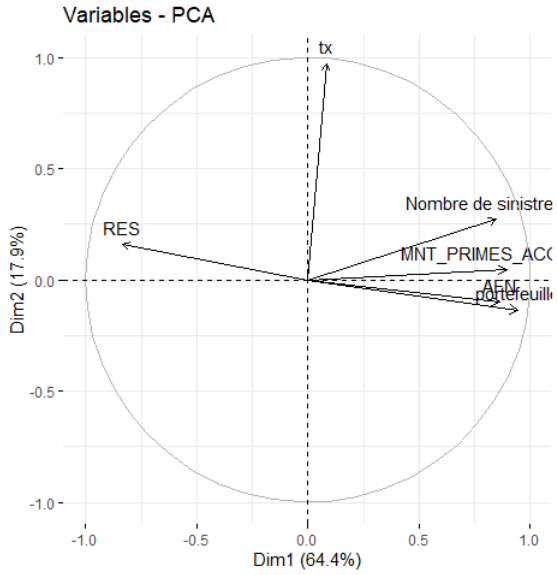


FIGURE B.6 – Résultat de l'ACP, produit deux roues

Les variables Nombre de sinistres, PACQ, AFN et portefeuille sont positivement correlées, et regroupées entre elles par rapport au premier axe. On voit que la qualité de représentation de toutes les variables est très bonne, car la distance entre ces variables et l'origine est importante.

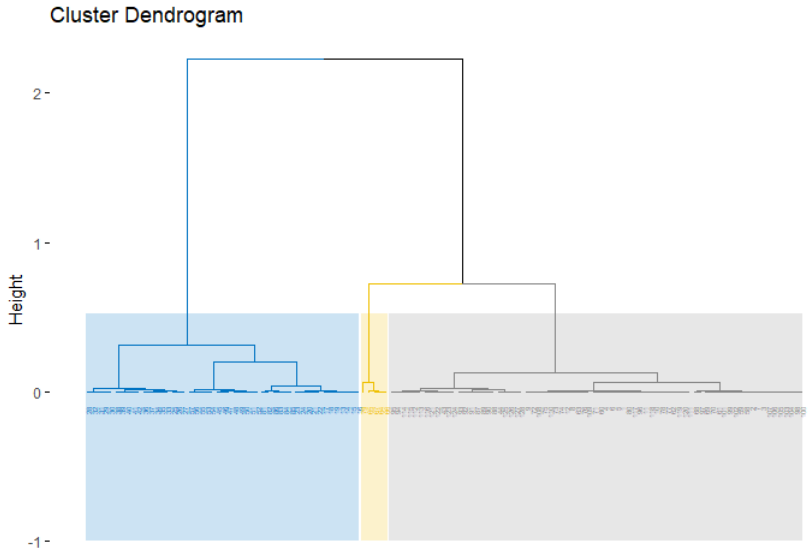


FIGURE B.7 – Dendogramme après classification, produit deux roues

Dendogramme et nombre de groupes Le dendogramme nous suggère une solution à 3 groupes. Cependant les groupes sont de tailles très différentes. On va visualiser les groupes suggérés.

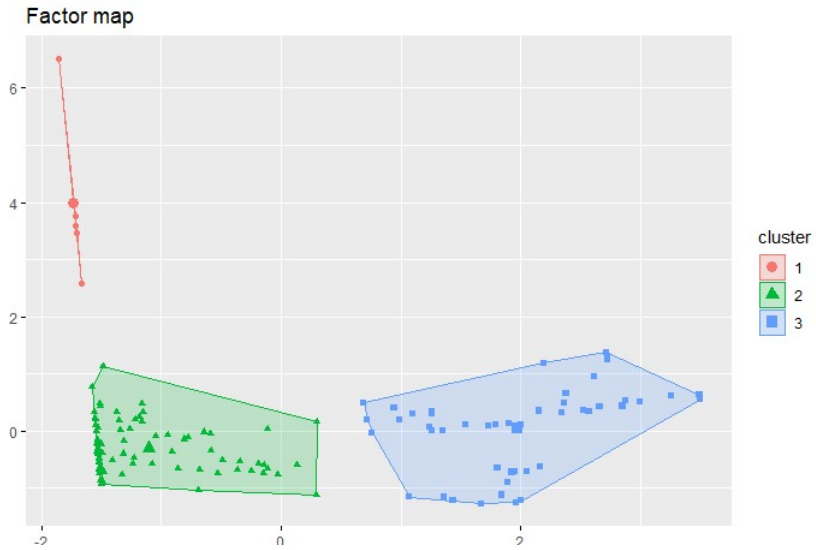


FIGURE B.8 – Resultat de l'HCPC, produit deux roues

résultat Les resultats de la classification ne semblent pas avoir répondu à notre problématique. On voit que l'un des groupes n'est pas optimal à une modélisation. On va comparer les performances des groupes formés et choisir la classification la plus performante pour la modélisation mais également pour notre stratégie de revalorisation.

Catégorie	S/P en % 2018 vu en 2020	S/P en % historique vu en 2020
auto standard	**82**	**92**
1	79	96
2		44
3	92	77
auto malussée	**87**	**75**
1		224
2	105	75
3	69	71
Deux roues	**59**	**69**
1	65	65
2		88
3	55	70
GENERAL	78	80

FIGURE B.9 – Statistiques de performance par groupe, classification par l'algorithme de classification hiérarchique

Annexe C

Rappel et Notions

C.1 Maximum de vraisemblance

Loi discrète

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires iid de loi discrète $L(\theta)$. La vraisemblance de l'échantillon $(x_1, ..., x_n)$ pour la loi L est donnée par la fonction

$$L_n : (x_1, \ldots, x_n; \theta) \mapsto L_n (x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \mathbb{P}_\theta (\{X_i = x_i\}) .$$

Loi continue

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires iid de loi continue $L(\theta)$ et de densité f_θ. La vraisemblance de l'échantillon $(x_1, ..., x_n)$ pour la loi L est donnée par la fonction

$$L_n (x_1, \ldots, x_n; \theta) := \prod_{i=1}^{n} f_\theta (x_i) .$$

Estimateur

L'estimateur du maximum de vraisemblance du paramètre θ de la loi $L(\theta)$ est donné par :

$$Argmax_\theta \{L_n (x_1, \ldots, x_n; \theta)\}$$

avec $x_i \in \mathbb{N}\ \forall i \geq 1$ pour les lois discrètes et $x_i \in \mathbb{R}\ \forall i \geq 1$ pour les lois continues.

C.2 Algorithme de Newton Raphson

La méthode consiste à introduire une suite (x_n) d'approximations successives de l'équation $f(x) = 0$ Nous allons initialiser la suite à un x_0 au voisinage de la solution. Puis à partir de ce x_0, nous calculons x_1 en traçant la tangente à $\mathcal{C}_f$ en x_0. Enfin nous réitérons ce procédé n fois.

Nous obtenons alors la formule de récurrence suivante :

x_{n+1} est l'abscisse du point d'intersection de la tangente à G_f en x_n avec l'axe des abscisses.
L'équation de la tangente en x_n est

$$y = f'(x_n)(x - x_n) + f(x_n)$$

Cette tangente coupe l'axe des abscisses quand $y = 0$:

$$f'(x_n)(x - x_n) + f(x_n) = 0 \Leftrightarrow f'(x_n)(x - x_n) = -f(x_n)$$

$$x - x_n = -\frac{f(x_n)}{f'(x_n)} \Leftrightarrow x = x_n - \frac{f(x_n)}{f'(x_n)}$$

Nous obtenons donc la relation de récurrence suivante :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Pour que la suite (x_n) existe : la fonction f doit être dérivable en chacun des points considérés. En pratique la fonction doit être dérivable dans un intervalle centré en α contenant x_0. La dérivée ne doit pas s'annuler sur cet intervalle. Pour que la suite (x_n) soit convergente, les conditions dépassent le cours de terminale, mais en pratique, il faut prendre un x_0 assez proche de la valeur α qui annule la fonction. Nous le déterminons à l'aide du théorème des valeurs intermédiaires.

On définit un critère d'arrêt pour notre algorithme avec une précision p tel que :

$$\frac{f(x_n)}{f'(x_n)} < 10^{-p}$$

C.3 Rappel sur les tests

Table de Kolmogorov Smirnov

n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

FIGURE C.1 – Table de Kolmogorov Smirnov

Test d'ajustement du χ^2

Définition

Soit $p = (p_1, ..., p_k)$ une mesure de probabilité sur l'ensemble $\{1, ..., k\}$ et $q = (q_1, ..., q_k)$ une autre mesure de probabilité. La distance du χ^2 de q à p est donnée par :

$$\chi^2(p, q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{p_i}.$$

Le χ^2 d'ajustement entre la loi p et la loi empirique $\bar{p}_n$ est la variable aléatoire :

$$\chi_n^2(p, \bar{p}_n) = n\chi^2(p, \bar{p}_n) = \sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i}.$$

Il peut être écrit de cette façon :

$$\chi_n^2(p, \bar{p}_n) = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}$$

où $e_i = np_i = E[N_i]$ est la valeur espérée de N_i Si n est grand, $\chi_n^2(p, \bar{p}_n)$ suit la loi du $\chi^2(k-1)$

Procédure du test

Soit $(X_n)_{n1}$ une suite de variables aléatoires iid de loi discrète sur l'ensemble $\{1, ..., k\}$. On veut tester l'hypothèse :

1. H_0 "la loi de l'échantillon est égale à p"
2. H_1 " la loi de l'échantillon est différente de p"

Sous l'hypothèse H_0, et si n est assez grand, on sait que la variable aléatoire $\chi_n^2(p, \bar{p}_n)$ suit la loi $\chi^2(k-1)$, donc

$$\mathbb{P}_{H_0}(rejeterH_0) = \mathbb{P}_{H_0}\left(\chi_n^2(p, \bar{p}_n) > \chi_{k-1,\alpha}^2\right) \approx \alpha$$

ce qui montre que le test est justifié.

C.4 Test de Shapiro Wilk

Le test de Shapiro-Wilk va tester l'hypothèse nulle selon laquelle un échantillon $x_1, ..., x_n$ est issu d'une population normalement distribuée. Par conséquent, si la p-value du test est significative, l'échantillon ne suit pas une loi normale. La statistique du test W est donnée par :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

les coefficients a_i sont donnés par $\left(\frac{(m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}\right)_i$ où les m_i sont les espérances des statistiques d'ordres x_i et V est la matrice de variance-covariance.

C.5 KPSS

Le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) permet de déterminer si une série temporelle est stationnaire autour d'une moyenne ou possède une tendance linéaire, ou si elle est non stationnaire. On veut accepter l'hypothèse nulle qui affirme que nos données sont stationnaires et rejeter l'hypothèse alternative.
Le test est basé sur de la régression linéaire. Il sépare la série temporelle en 3 parties : une tendance déterministe (ω_t), une marche aléatoire t et une erreur stationnaire (ϵ_t). On obtient l'équation de régression suivante :
$\begin{cases} y_t = \omega_t + \lambda t + u_t & u_t \sim iid\aleph\left(0, \sigma_u^2\right) \\ \omega_t = \omega_{t-1} + v_t & v_t \sim iid\aleph\left(0, \sigma_v^2\right) \end{cases}$ On teste l'hypothèse nulle $\sigma_v^2 = 0$ qui, sous l'hypothèse de stationnarité des u_t, assure la stationnarité de y_t autour d'une tendance (ou autour d'une constante s'il est possible d'établir que λ est non significativement différent de 0).

C.6 Durbin Watson

Le test de Durbin-Watson est un test statistique qui permet de tester l'autocorrélation des résidus dans un modèle de régression linéaire. Il permet de vérifier la significativité de ρ dans la formule

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

où ϵ_t est le résidu estimé et u_t est un bruit blanc. L'hypothèse nulle H0 stipule qu'il y a non auto-corrélation donc $\rho = 0$. L'hypothèse alternative H1 stipule qu'il y a auto-corrélation donc ρ différent de 0 avec toujours $|\rho| < 1$. La statistique de Durbin-Watson est définie par :

$$DW = \frac{\sum_{t=2}^{n} \left(\varepsilon_t - \varepsilon_{t-1}\right)^2}{\sum_{t=1}^{n} \varepsilon_t^2}$$

C.7 Modèle linéaire gaussien

Le modèle linéaire gausien s'écrit sous la forme :

$$Y = X\beta + E$$

avec $E_n(0, \sigma^2 I_n)$ On note ϵ_i les observations du vecteur aléatoire E, indépendamment et identiquement distribuées selon une loi $N(0, \sigma^2)$. Par conséquent, on note y_i les observations du vecteur Y indépendamment distribuées selon $N((X\beta)_i, \sigma^2)$.
Il en découle la normalité de Y, distribuée selon la loi $N_n(X\beta, \sigma^2 I_n)$.

Estimation

On peut estimer nos paramètres par la méthode du maximum de vraisemblance. En effet, la vraisemblance associée au modèle s'écrit :

$$L(X, \beta, \sigma|Y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)(Y - X\beta)^t\right)$$

Nous obtenons comme estimateur de β :

$$X^tX\hat{\beta} - X^tY = 0 \Leftrightarrow \hat{\beta} = \left(X^tX\right)^{-1} X^tY$$

L'estimateur biaisé de σ s'écrit :

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}\hat{\epsilon}^t}{n}$$

Un deuxième estimateur de σ non biaisé peut s'écrire :

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}\hat{\epsilon}^t}{n-p-1}$$

Enfin les estimations de la variable réponse Y s'écrivent :

$$\hat{Y} = X\hat{\beta} = X\left(X^tX\right)^{-1} X^tY$$

Lien avec la famille exponentielle

Dans ce mémoire, il nous est présenté la notion de famille exponentielle, dont la loi normale fait partie. Nous allons donner des exemples rapides de lois appartenant à la famille exponentielle :

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$	$c(y_i, \phi)$
Normale $(\mu_i; \sigma^2)$	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right\}$
Poisson (μ_i)	$\ln(\mu_i)$	1	ϕ	$\exp(\theta_i)$	$-\ln y!$
Binomiale $\frac{1}{m_i}(m_i; \mu_i)$	$\ln\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	ϕ	$\ln(1+\exp\theta_i)$	$\ln\binom{m_i}{m_i y_i}$
$Gamma(\mu_i; \alpha)$	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-\log(-\theta)$	$\alpha\ln(\alpha y) - \ln y - \ln\Gamma(\alpha)$
Inverse Gaussienne $(\mu_i; \sigma^2)$	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left\{\ln(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$

Annexe D

Outil Excel

Dans ce mémoire, nous représentons quelques graphiques venant directement de l'outil mis en place. Cet outil va permettre d'obtenir, lors des revalorisations de primes, un indicateur de la tendance de sinistralité grave que l'on peut esperer obtenir par partenaire mais également d'un point de vue global.
Cet outil est composé de plusieurs onglets permettant soit l'implémentation des inputs, le calcul et la visualisation des résultats :

1. Paramètre des données : Permet de rentrer les paramètres liés aux données ;
2. Implémentation des données ;
3. Impact de l'inflation : Le choix de la prise en compte de l'inflation revient à l'utilisateur ;
4. Paramètre du modèle : Permet de choisir le seuil à tester et résume les résultats ainsi que certaines indications ;
5. Calcul - Détermination du seuil : Calcul le seuil par méthode de Hill et de Pickands ;
6. Calcul - Détermination des estimateurs : Utilise la méthode du MLE pour obtenir les paramètres ;
7. Modèle Fréquence sur l'ensemble du portefeuille : Détermine le nombre de sinistres à partir du seuil et va modéliser une loi de poisson ou binomial négative ;
8. Modèle basé sur la crédbilité : Estime le nombre de sinistres par modélisation bayésienne pour chaque partenaire ;
9. Monte Carlo : Utilise les estimations données plus tôt pour modéliser la charge de sinistre potentiel au global et par partenaire.

Pour des raisons de confidentialité, nous ne montrerons que l'onglet qui permet de paramétrer les données et le modèle de l'outil en question, on utilise des macros VBA pour faciliter et automatiser la procédure :

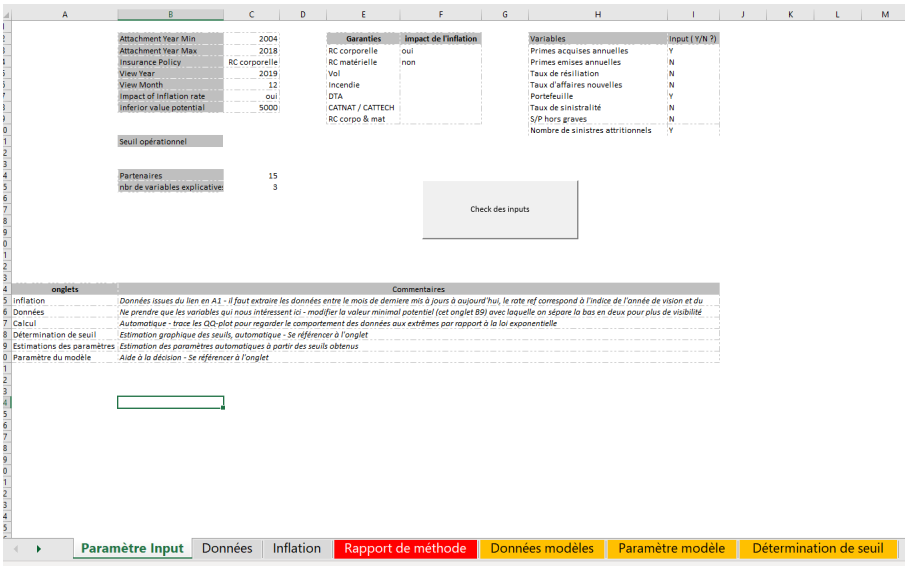


FIGURE D.1 – Onglet des paramètres des données

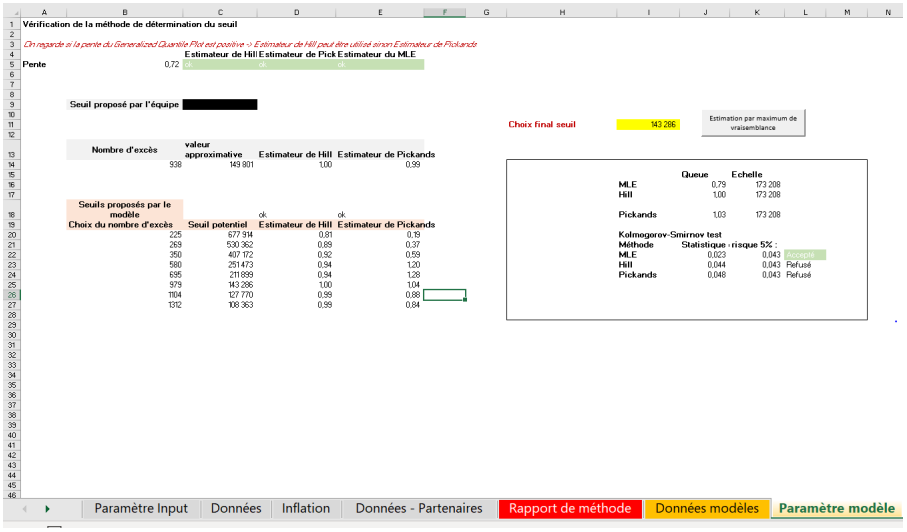


FIGURE D.2 – Onglet des paramètres du modèle