Optimisation de la prise en compte de la sinistralité dans la tarification Automoteur Agricole

Zyad Osseni

Mémoire présenté devant

l'UFR de Mathématique et d'Informatique

pour l'obtention du Diplôme Universitaire d'Actuaire de Strasbourg

et l'admission à l'Institut des Actuaires

Le 24 Janvier 2014

Par : Zyad OSSENI

Titre: Optimisation de la prise en compte de la sinistralité dans la tarification Automoteur Agricole

Membres du jury de l'Institut des Actuaires

Membres du jury de l'UdS :

M. Jacques Franchi
Mme Sandrine Spaeter-Loehrer

M. Karl-Théodor Eisele
M. Jean-Lucien Netzer
Mme Myriam Maumy-Bertrand
M. Bernard Heinkel
M. Patrick Rondé
M. Jean Bérard

Invités :

M. Philippe Artzner
M. Pierre Devolder

M. Jean Modry
Mme Magali Kelle-Vignon
M. Alexandre You

Secrétariat : Mme Maire-Lantz
Mme Fidelin

Bibliothèque : Mme Christine Disdier

Entreprise :

Nom : CA PACIFICA

Directeur de mémoire en entreprise :
Nom : Laura CANDAS et Yann MERCUZOT

Invité :
Nom : Christophe DUTANG

CONFIDENTIALITE : 5 ans

Signature du responsable entreprise

Signature du candidat

AVERTISSEMENT :

Pour des raisons de confidentialité, certaines techniques et valeurs numériques ne seront pas présentées. Egalement, certains graphiques n'auront pas d'échelle de valeur dans le but de préserver les résultats de nos analyses.

Sommaire

Résumé

L'assurance Automoteur Agricole est un produit qui couvre des dégâts liés à des engins agricoles tels que les tracteurs ou les moissonneuses batteuses. Ce produit possède des similitudes avec un produit classique d'assurance automobile, à travers les garanties proposées dans les contrats.

D'un point de vue technique cependant, le produit Automoteur Agricole n'a pas encore atteint sa maturité tarifaire chez PACIFICA. L'ajustement du tarif à la sinistralité individuelle constitue une question centrale dans le cadre des évolutions tarifaires apportées au produit Automoteur Agricole depuis quelque temps.

Ce mémoire décrit un processus de prise en compte de la sinistralité individuelle des contrats dans leur tarification. Dans un premier temps, la démarche appliquée consiste à modéliser séparément la fréquence des sinistres et le coût moyen des sinistres à l'aide des modèles linéaires généralisés (GLM). L'approche dite « Fréquence x Coût moyen » permet d'analyser la pertinence et le rôle des principaux critères tarifaires qui constituent l'équation tarifaire du produit Automoteur Agricole. Elle permet aussi de vérifier que la sinistralité individuelle passée est significativement prédictive de la fréquence de sinistres future.

Ce document propose ensuite, d'étudier la théorie de la crédibilité linéaire. Les modèles de Bühlmann-Straub et de Jewell sont successivement appliqués sur un jeu de données représentatif du portefeuille pour estimer a posteriori la fréquence de sinistre. Enfin, plusieurs scénarios de sinistralités sont présentés pour analyser l'impact de la prise en compte de la sinistralité individuelle dans l'estimation de la fréquence de sinistres attendue.

Mots clés : l'assurance Automoteur Agricole – évolutions tarifaires - sinistralité individuelle – équation tarifaire – modèles linéaires généralisés – théorie de la crédibilité linéaire – approche « Fréquence x Coût moyen ».

Summary

The insurance Agricultural Engine is a product which covers damages bound to agricultural machines such as tractors or combine-harvesters. This product has similarities with a classic product of car insurance, through proposed guarantees proposed in contracts.

From a technical point of view however, the Agricultural Engine product has not reached its maturity tariff at PACIFICA yet. The adjustment of the rate with the individual loss ratio constitutes a central question within the framework of the tariff evolutions made to the Agricultural Engine product for some time.

This essay describes a process of taking into consideration of the individual loss ratio of contracts in their pricing. At first, the approach applied consists in modeling separately the frequency of the disasters and their average cost using the generalized linear models (GLM). The so called “Frequency x average Cost” approach allows to analyze the relevance and the role of the main tariff criteria which establish the tariff equation of the Agricultural Engine product. This also allows to verify that the individual loss ratio is significantly predictable with regard to the expected frequency of disasters.

This document proposes then, to study the theory of the linear credibility. The models of Bühlmann-Straub and Jewell are successively applied to a representative database of the portfolio to estimate a posteriori the frequency of disaster. Finally, several scenarios of loss ratios are presented to analyze the impact of the consideration of the individual loss ratio in the estimation of the expected frequency of disasters.

Key words: The insurance Agricultural Engine - tariff evolutions - individual loss ratio - tariff equation - generalized linear models - theory of the linear credibility – approach “Frequency x average Cost".

Remerciements

Ce mémoire vient clôturer la réalisation d'un stage de fin d'études actuarielles au sein du service actuariat de la direction du marché du particuliers (DPART) et qui débouche sur l'obtention du titre d'actuaire. Plusieurs personnes ont contribué directement et indirectement pour que cet aboutissement soit effectif.

Je tiens tout d'abord à remercier Yann MERCUZOT, responsable du service actuariat, Laura CANDAS et Christophe DUTANG, mon tuteur académique, pour leur encadrement exceptionnel, leur patience et la confiance qu'ils m'ont chacun accordée.

Je remercie l'ensemble des collaborateurs du service actuariat pour m'avoir accueilli chaleureusement dans leur équipe durant ces 6 mois. Un grand merci plus généralement aux différents collaborateurs de la direction du marché des particuliers et de celle du marché agri-pro (DMAP) avec qui j'ai pu échanger de près ou de loin sur mon sujet de mémoire.

Mes remerciements au personnel enseignant et administratif de l'Université de Strasbourg et plus particulièrement au master actuariat pour la formation académique transmise.

Enfin, je dédicace ce mémoire à ma famille, qui m'a toujours soutenu et aidé à concrétiser mes rêves.

Liste des abréviations

- AA (Automoteur Agricole)
- AGIRA (Assurance pour la Gestion des Informations sur le risque en Assurance)
- AIC (Akaike Informative Criterion)
- BDG (Bris de Glace)
- BDM (Bris de matériels)
- BIC (Bayesian Informative Criterion)
- DOM (Dommages Accidentels)
- FFSA (Fédération Française des Sociétés d'Assurance)
- GLM (General Linear Models)
- GPD (General Pareto Distribution)
- HT (Hors Taxe)
- IARD (Incendie Accidents et Risques Divers)
- OAT (Outil d'Analyse Tarifaire)
- POT (Peaks Over Threshold)
- RC (Responsabilité Civile)
- VE (Valeurs Extrêmes)

Introduction

Lors de la conception d'un produit d'assurance, via des méthodes statistiques, l'actuaire se sert de facteurs de risque dits observables pour déterminer la prime d'assurance. Les facteurs de risque sont issus d'informations récoltées auprès de la personne assurée. Dans le domaine de l'assurance IARD (Incendie, Accidents et Risques Divers) par exemple, ces informations peuvent porter à la fois sur la personne assurée, le bien assuré et le lieu d'utilisation du bien. Grâce à ces facteurs, l'actuaire segmente le portefeuille de polices en classes de risques les plus homogènes[1] possibles. De ce fait, les clients appartenant à la même classe de risque devront payer la même prime de risque. Ce processus permet à la fois de respecter le principe fondamental de mutualisation de l'assurance et de répondre au problème d'anti-sélection inhérent au marché de l'assurance, en particulier en assurance automoteur agricole. Cette dernière est une couverture en dommages et responsabilités, destinée aux propriétaires d'automoteurs agricoles.

Sur ce marché de l'assurance automoteur agricole, la concurrence entre compagnies d'assurance est de plus en plus accrue. Il est donc essentiel pour un assureur tel que PACIFICA, de proposer un tarif qui soit le plus possible en adéquation avec le risque automoteur agricole. Une solution logique consisterait donc à optimiser le processus de segmentation du risque sous-jacent au portefeuille des contrats automoteurs agricoles.

Sur le sujet, le mémoire présenté se penche sur l'intérêt d'intégrer la sinistralité individuelle dans la tarification du produit automoteur agricole. Nous articulons ce mémoire autour de trois chapitres.

Le premier chapitre présente le cadre opérationnel de l'étude, à savoir le marché de l'assurance Automoteur Agricole (AA) et décrit plus en détail la problématique du mémoire. Le second chapitre est consacré à la modélisation de la sinistralité d'un portefeuille AA à l'aide des **modèles linéaires généralisés**. Enfin, le troisième chapitre est consacré à l'étude de la prise en compte de la sinistralité passée à l'aide de la **théorie de la crédibilité**.

[1] *Les facteurs de risques utilisés pour définir le tarif sont alors communément appelés* ***critères tarifaires****.*

Chapitre 1. L'assurance automoteur et sa tarification

1.1 L'assurance automobile agricole

1.1.1 Définition d'un automoteur agricole

Les engins agricoles, sont des instruments mécaniques utilisés par des exploitants agricoles pour mener leurs activités. Ces engins sont une substitution ou un complément à la main-d'œuvre dans le cadre d'une production agricole. Parmi ces engins agricoles, nous citons :

- **Les tracteurs agricoles** : ce sont des automoteurs équipés de 4 roues de tailles souvent importantes ou de chenilles. Un tracteur permet de tracter des remorques ou d'autres machines agricoles de par sa structure (charrues, épandeurs à fumier, machines agricoles comportant des pièces rotatives…).

- **Les moissonneuses-batteuses** : ce sont des automoteurs destinés à la récolte des plantes à graines en une seule opération. Elles permettent de réaliser simultanément la moisson et le battage. Coûteuses, leur utilisation est ponctuelle mais intensive.

- **Les ensileuses** : elles servent à récolter du fourrage vert (herbe, maïs ou céréales immatures) ou du fourrage pré-fané pour faire de l'ensilage.

La souscription à un contrat d'assurance Automoteur Agricole est indispensable pour deux raisons. D'une part, les automoteurs agricoles sont essentiels au bon fonctionnement d'une exploitation agricole. Ils représentent un investissement élevé pour leurs propriétaires tant leur achat et le prix de la maintenance sont élevés. Les conséquences d'une immobilisation ou d'un sinistre peuvent peser sur l'activité de l'exploitant. D'autre part, il y a une obligation juridique[2] à la souscription d'un contrat d'assurance automoteur.

[2] **Article L211-1** du Code des assurances en annexe A

1.1.2 Les garanties chez *PACIFICA*

Le produit d'assurance *Automoteur Agricole* (AA) contient plusieurs garanties dont la souscription se fait par type de formule. La **Figure 1** présente le contenu de chaque formule.

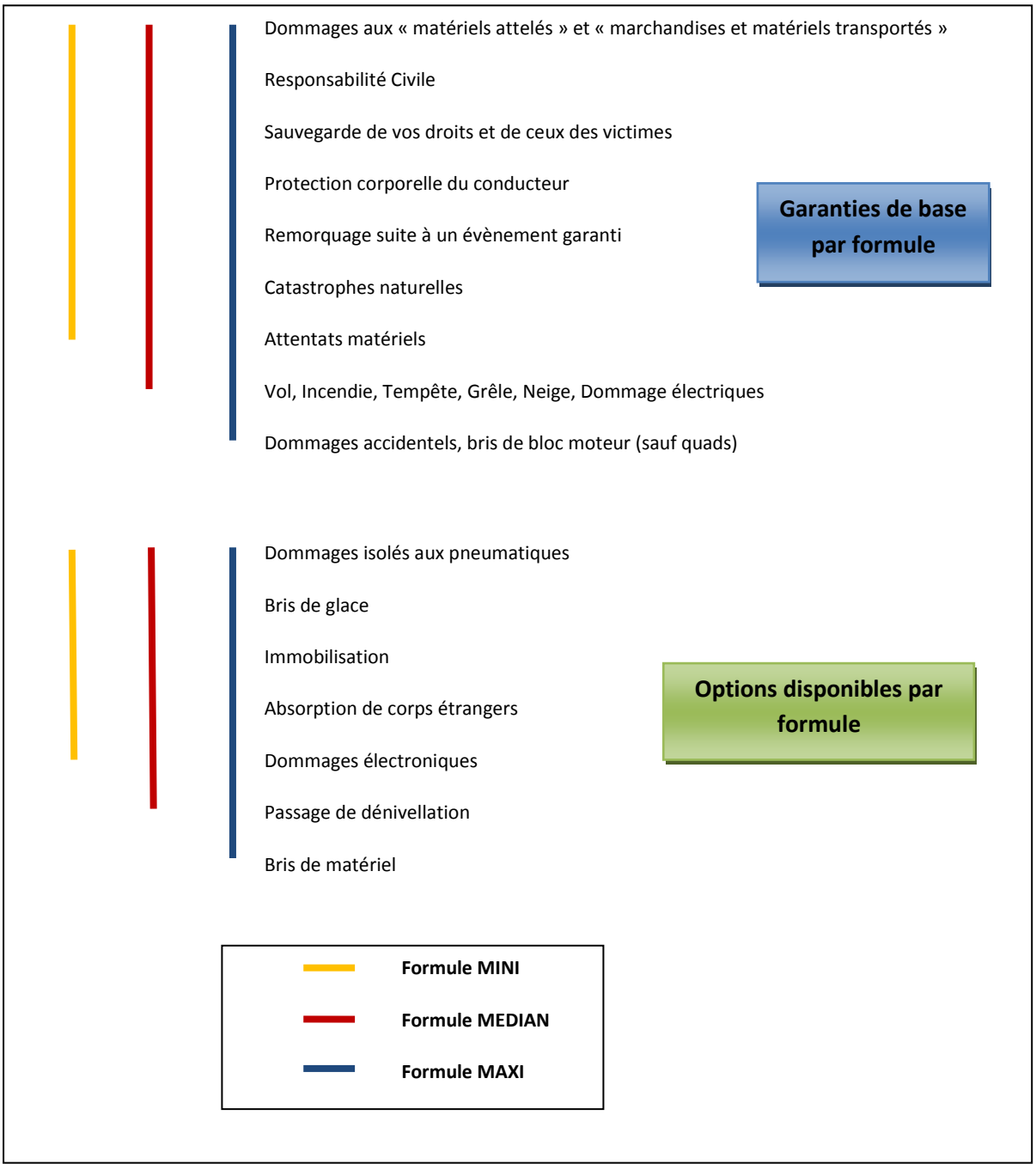


Figure 1 : Liste des garanties par Formule

Les formules constituées d'un mixte de garanties seront analysées plus en détail dans le chapitre 2. Le **Tableau** ci-dessus ci-après présente un descriptif d'une liste non exhaustive des garanties AA. Ces garanties ont un intérêt particulier dans la tarification du risque AA du fait de leur sinistralité.

Garantie	Description
RC (Responsabilité Civile)	*C'est la principale garantie du produit AA, étant obligatoire d'un point de vue juridique. Elle est souscrite pour faire face aux conséquences pécuniaires de la responsabilité civile pour les dommages causés à autrui. Elle est dite **corporelle** lorsqu'il s'agit de l'atteinte à l'intégrité physique d'une tiers personne et **matérielle** lorsqu'il s'agit d'un dommage causé sur le bien appartenant à une tiers personne. La distinction RC matérielle/corporelle est importante en termes de tarification. En effet un sinistre corporel est caractérisé par une faible fréquence des sinistres et un coût d'indemnisation important. A l'inverse, un sinistre matériel est caractérisé par une fréquence des sinistres élevée et un coût moindre.*
IV (Incendie, Vol)	*Elle couvre les dommages subis par l'automoteur suite à un incendie ou à un acte de vandalisme de la part d'autrui.*
BDG (Bris de glace)	*Elle est souscrite pour le remboursement à l'assuré des frais engagés suite à un bris accidentel des éléments :* ❖ *Les glaces* ❖ *Les blocs optiques* ❖ *Les feux de signalisation*
BDM (Bris de matériels)	*Elle sert à couvrir :* ❖ *Toute destruction ou détérioration accidentelle de l'automoteur assuré ainsi que ses accessoires* ❖ *S'applique également pendant les opérations de démontage, de remontage et de vérifications nécessitées par l'entretien ou la révision de l'automoteur assuré*
DOM (Dommages Accidentels)	*Elle sert à couvrir l'automoteur contre tout dommage accidentel direct provenant :* ❖ *D'un choc contre un corps fixe ou mobile* ❖ *D'une immersion* ❖ *D'un choc entre les composants d'un même attelage*

Tableau 1 : Description des principales garanties AA

1.1.3 Les spécificités du risque automobile agricole

Toutes garanties confondues, la fréquence des sinistres observée chez les automoteurs agricoles est semblable à celle observée chez les automoteurs urbains (entre **16% et 20% d'après les données PACIFICA**). Cependant les contrats sinistrés d'une année à l'autre sont très concentrés en comparaison avec la sinistralité d'un portefeuille d'assurance automobile urbain. Ceci est une source d'hétérogénéité résiduelle du portefeuille AA qui nous incitera dans le cadre du mémoire à préconiser une tarification intégrant la sinistralité individuelle passée.

Par ailleurs, le risque *Automoteur Agricole* est caractérisé par une **corrélation forte entre l'intensité de la sinistralité et la période d'exploitation agricole**. Faible en début d'année, la fréquence des sinistres connait des pics pendant la période de moisson (entre Juin et Septembre).

En termes de sinistralité, le produit AA met environ 2 ans à vieillir et à stabiliser ses résultats techniques. Il y a environ 10 % de malis entre la vision d'un exercice de survenance **N** et celle de l'exercice **N+1**.

1.2 La tarification actuelle de l'assurance automobile agricole chez PACIFICA

La tarification d'un contrat d'assurance *Automoteur Agricole* s'effectue en plusieurs étapes comme le montre la **Figure** ci- après.

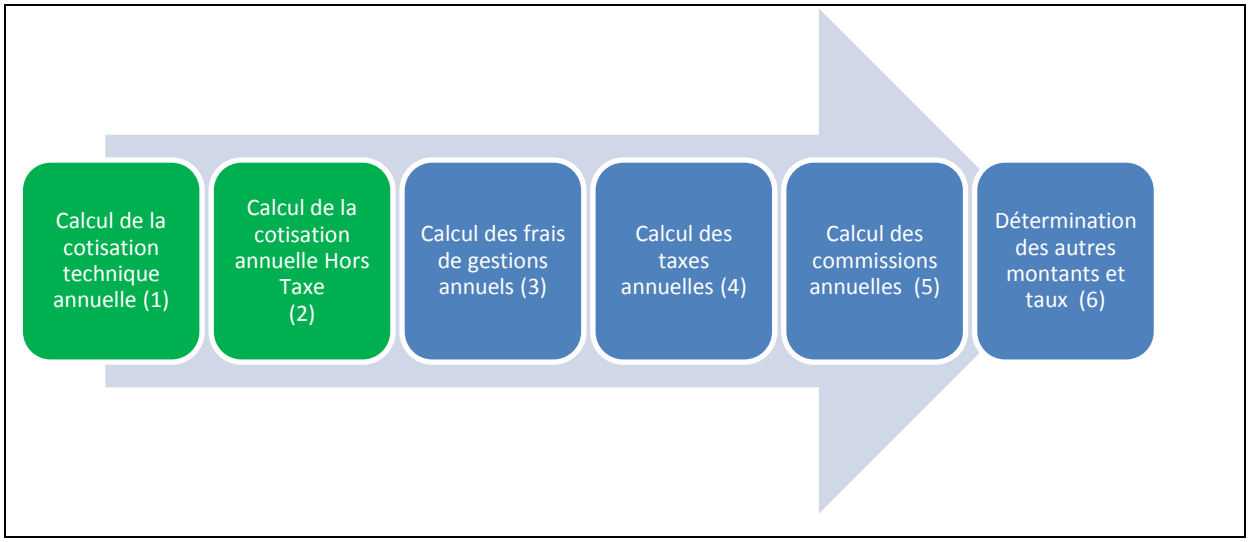


Figure 2 : Le processus de tarification d'un contrat Automoteur Agricole

Les étapes en bleu (3), (4), (5) et (6) sont pilotées sans tenir compte du risque intrinsèque AA. Les étapes (1) et (2) conduisent à la détermination d'une prime pure du contrat AA. Ces dernières étapes prenant en compte le risque intrinsèque AA, constituent l'ossature du mémoire. Toutefois nous ne présentons que l'étape (1) car l'étape (2) étant soumise à la règle de confidentialité.

1.2.1 Calcul de la cotisation technique annuelle

Comme pour l'assurance automobile des particuliers, l'évaluation du risque automobile agricole passe par sa segmentation, en fonction de plusieurs critères tarifaires. Ces derniers définissent des **facteurs explicatifs** de la sinistralité. L'évaluation statistique de ces facteurs permet de déterminer le coût du sinistre moyen le plus probable. Les critères tarifaires dépendent logiquement des caractéristiques sur :

- **Le souscripteur d'assurance** : son âge, son activité, son expérience,…etc.
- **L'automoteur** : la famille d'engin, la puissance de l'engin, l'âge de l'engin,…etc.
- **Le contrat** : le package des garanties, durée du contrat,…etc.

L'inventaire de ces critères tarifaires permet d'obtenir **une équation tarifaire** qui est une modélisation du risque couvert. Le tarif proposé au souscripteur d'un contrat AA est donc basé sur des informations provenant de sa personne, de l'automoteur assuré et des spécificités du contrat. L'encadré suivant, présente la structure de l'équation tarifaire du produit AA.

Nous avons,

$$\boldsymbol{P = B \times V_1 \times V_2 \times \ldots,}$$

Où $\boldsymbol{P}$ désigne la cotisation technique annuelle ou *prime pure a priori*, $\boldsymbol{B}$ désigne une *prime de base* et $\boldsymbol{V_i(i = 1, 2, \ldots)}$ désigne un *coefficient multiplicateur* pour le $i^{ème}$ critère tarifaire retenu.

1.2.2 Critique de la méthode actuellement appliquée

1.2.2.1 Maturité tarifaire non atteinte

Peu de compagnies proposent des couvertures du risque automoteur agricole. Les principales compagnies, dont fait partie PACIFICA, proposant des produits d'assurance Automoteur Agricole représentent 80 % du marché national. Le marché de l'assurance Automoteur Agricole est donc peu concurrentiel. Or la concurrence est un moteur d'innovation. Cette faible concurrence explique en partie le peu d'études actuarielles existantes dans le cadre de la tarification d'un contrat automoteur agricole.

1.2.2.2 Une tarification a posteriori perfectible chez PACIFICA

L'assurance *Automoteur Agricole* fait partie du marché de l'assurance des biens agricoles [1]. Toutefois, elle présente des caractéristiques tarifaires proches de celles de l'assurance automobile (pour particulier ou professionnel). En effet, la **section 1.1.2** a présenté des garanties telles que la responsabilité civile (**RC**), le bris de glace (**BDG**) que l'on retrouve également dans un contrat classique d'automobile. Il est donc légitime d'appliquer les mêmes méthodes de tarification.

Cependant, il n'existe pas en assurance Automoteur Agricole un organisme central tel que l'AGIRA[3], qui dispose d'informations (nombre de sinistres, charge de sinistres) sur les clients potentiels (prospects). La présence d'un tel organisme permettrait de mieux connaître le passif de sinistralité d'un souscripteur au produit AA. Le passif d'un client n'est connu qu'à travers son ancienneté dans le portefeuille.

Il n'existe pas non plus, un système réglementaire de bonus-malus. Un tel système permet de réduire le risque d'aléa moral et d'anti-sélection assez important en assurance IARD. Il permet aussi de réduire l'hétérogénéité résiduelle des modèles de tarification [2].

[3] Assurance pour la Gestion des Informations sur le risque en Assurance

1.2.2.3 Proposition de méthodes de tarifications actuarielles

La ressemblance constatée entre l'assurance Automoteur Agricole et l'assurance automobile, permet d'appliquer les méthodes statistiques utilisées pour cette dernière.

L'objet du **Chapitre 2** sera de modéliser la sinistralité AA en intégrant les variables tarifaires incluses dans l'équation tarifaire et d'étudier in fine l'opportunité d'ajouter la sinistralité individuelle antérieure. Il s'agira de modéliser d'une part la fréquence des sinistres et d'autre part le coût moyen des sinistres via des **modèles de régression** sur la base d'un historique de sinistralité AA de 9 exercices successifs.

Sur la question de la prise en compte tarifaire de la sinistralité individuelle, le service actuariat a mis en place un processus de majoration de la prime pure depuis 2009. Ce processus est basé sur l'expérience de sinistralité et la responsabilisation du contrat. Pour des raisons de confidentialité, nous nous permettons de ne pas rentrer plus en détail dans la définition de ces coefficients de majoration. Du fait de la problématique de ce mémoire cependant, nous pouvons affirmer que ces coefficients sont encore perfectibles. L'objet du chapitre 3 sera donc d'étudier l'apport potentiel de la crédibilité pour exploiter la sinistralité individuelle.

Chapitre 2. Optimisation tarifaire à l'aide des GLMs

Le chapitre 1 a présenté l'équation tarifaire du produit AA. Cette équation est définie par un montant de base qui est ajusté en fonction de différents critères tarifaires. L'objet du chapitre 2 est de vérifier la pertinence des principaux critères tarifaires qui définissent cette équation à l'aide d'un outil statistique rigoureux que sont les modèles linéaires généralisés[4]. Il s'agira ensuite de déterminer la pertinence de la sinistralité individuelle dans la prédiction de la prime pure.

Mais avant, il est essentiel d'effectuer une étude « à plat » des caractéristiques du risque AA, sur la base d'un jeu de données d'étude. Ceci constitue l'objet de la section suivante.

2.1 Construction de la base de données et Etudes à «plat»

Toute modélisation est précédée par la construction d'une base de données appropriée. C'est un processus long qui a été effectué via le logiciel **SAS**[5]. Nous nous intéressons à l'observation des contrats AA assurés au moins 1 jour entre le 01/01/2004 et le 31/12/2012.

2.1.1 La procédure de construction de la base de données d'étude

Au sein du service Actuariat Produit de PACIFICA, les actuaires disposent d'une base de données infocentre. Dans cette bibliothèque, plusieurs types de tables sont stockés et sont utilisés pour réaliser diverses études notamment de tarification. C'est à l'aide de ces tables que la base de données d'étude va être constituée.

2.1.1.1 Les informations sur le contrat

La table **base_contrat** contient les informations relatives aux contrats en portefeuille. C'est-à-dire qu'on retrouve toutes les variables qui déterminent le profil de risque d'un contrat. Le **Tableau** ci-dessous présente les variables qui ont un intérêt dans la suite du mémoire.

[4] *Dans la suite du mémoire, nous emploierons souvent le terme GLMs, qui définit l'acronyme en anglais des modèles linéaires généralisés :* ***General linear Models****.*

[5] ***SAS*** *est un logiciel d'analyse statistique. (Source : http://www.sas.com/offices/NA/canada/fr/technologies/analytics/statistics/stat/)*

Nom du champ	Description
contrat	*Le numéro de contrat de l'assuré*
image	*La période sur laquelle son profil de risque reste stable*
dtdeb	*Date de début d'image*
dtfin	*Date de fin d'image*
annee	*Exercice*
nbjaa	*Nombre de jours assurés au cours de l'exercice*
activite	*L'activité de l'assuré*
fameng	*Famille d'engins*
gamme	*Structure tarifaire*
formule	*Formule du produit choisie par l'assuré*
CR	*Code de la caisse régionale gérant le contrat*
puiss	*Puissance de l'automoteur*
dpt	*Département où se situe de l'automoteur*

Tableau 2 : Les données contrats

L'**image** du contrat qui a une date de début et une date de fin, est la période sur laquelle la vision du profil de risque du contrat est stable ou inchangée. Dès que le contrat est modifié (reconduction du contrat, avenant,…), une nouvelle image est créée.

Une ligne correspond à l'image d'un contrat. Pour chacune de ces observations, va correspondre :

- Un profil de risque
- Une cotisation acquise
- Le nombre de jours assurés sur la période de l'image
- Des informations sur la sinistralité vieillies d'un an (charge, nombre de sinistres)

2.1.1.2 Les informations sur le sinistre AA

La table **base_sinistre** contient les informations relatives aux sinistres, toutes garanties confondues, survenus et déclarés dans la période d'observation. Des retraitements sont effectués sur la charge brute[6] et le nombre de sinistres pour pourvoir obtenir des informations comparables lors des études statistiques effectuées par la suite.

A. Vieillissement de la sinistralité

D'autre part, un montant de sinistres en assurance de dommages et de responsabilité n'est pas toujours entièrement connu au moment de la déclaration du sinistre : il existe des délais liés notamment à l'expertise du sinistre, à l'éventuelle contestation en justice. On note également la déclaration tardive des sinistres (la date de déclaration du sinistre est postérieure de plusieurs mois à la date de survenance du sinistre). Nous procédons donc à un vieillissement de l'exercice de la manière suivante :

- Prise en compte les bonis et les malis ⇒ Modification sur la charge de sinistres brute.
- Prise en compte des déclarations **tardives** ⇒ Modification sur la charge de sinistres brute et sur le nombre de sinistres.

Pour obtenir une base d'observations homogènes pour les études statistiques, chaque exercice est « vieilli » d'un an (sauf 2012 vu à fin 11/2013 car il y a un impact marginal par rapport au 12/2013, la charge étant quasiment stabilisée)

Pour illustration, nous prenons l'exemple d'un assuré (ou contrat k) ayant été victime d'un sinistre corporel en 2009. A la fin de cet exercice un montant de 10000 € est enregistré pour l'indemnisation du contrat k. Cependant l'année suivante, l'état de santé de l'assuré s'est dégradé, ce qui s'est illustré par une revalorisation de 5000 € de la charge brute, avant de se stabiliser ensuite. A la fin de l'exercice 2012 le montant de charge brute associé au contrat k et vieilli d'un an est de 15000 €.

B. Actualisation de la charge de sinistre

La base contient des charges de sinistres qui datent aussi bien de 2004 que de 2012 par exemple. Chaque montant de sinistres est revalorisé à fin décembre 2012, à un taux

[6] Charge brute de réassurance

d'atualisation constant. Ainsi à chaque montant de sinistres Z « vieilli », de l'exercice N on associe le montant de sinistres actualisé suivant:

$$\tilde{Z} = Z \times (1 + 3\%)^{2012-N}.\text{Où } N = 2004, 2005, \dots, 2012$$

Nous considérons à nouveau le contrat k, suite au vieillissement et à la capitalisation, le montant de charge brute retenu pour l'étude est :

$$\tilde{Z} = 15000 \times (1 + 3\%)^{2012-2009}$$

$$\tilde{Z} = 16390\ €$$

Le **Tableau** ci-dessous présente les variables qui ont un intérêt dans la suite du mémoire.

Nom du champ	Description
Contrat	*Le numéro de contrat de l'assuré*
Sin	*Numéro de sinistre*
Charge	*Montant du sinistre déclaré en euro, brute de réassurance et retraité*[7]
DTSURV	*Date de survenance du sinistre*

Tableau 3 : Les données des sinistres survenus et déclarés dans la période d'observation

2.1.1.3 Détermination de la base de données d'étude et contrôles

Très logiquement les tables **base_sinistre** et **base_contrat** ont un champ en commun (la variable **contrat**): c'est un champ appartenant à **la clé primaire** de chacune des tables. Une clé primaire permet d'identifier de manière unique un enregistrement dans une table. Elle permet notamment d'effectuer des opérations de jointures entre les tables. La jointure entre les tables **base_sinistre** et **base_contrat** va permettre de rattacher un profil de risque à chaque sinistre via le numéro de contrat et la date de survenance du sinistre. Cette dernière doit être comprise entre la date de début et la date de fin de l'image du contrat. Le nombre de sinistres d'un contrat est alors obtenu en comptant le nombre de fois que le champ sin est renseigné au cours d'un exercice.

[7] Actualisation et vieillissement

Nous effectuons ensuite les contrôles suivants pour valider la base finale :

- *L'homogénéité* : On vérifie que les données sont toutes liées au produit AA uniquement.
- *La vérification des sources des données* : elle se traduit par l'analyse des tables d'où elles proviennent.
- *Examen des données manquantes* : la base de données constituée n'a pas présenté de données manquantes. Si cela avait été le cas, il aurait fallu rajouter une catégorie aux variables catégorielles indiquant lorsque l'information est manquante ou plus simplement ignorer l'information lorsqu'elle ne constitue pas un élément majeur.

2.1.2 Analyses descriptives des données

Nous effectuons une analyse descriptive de la base de données en amont de la modélisation. En effet, cette étape permet de mieux connaître le portefeuille et d'avoir du recul dans l'interprétation des résultats de la modélisation par la suite.

2.1.2.1 Le nombre et le montant des sinistres

A. Le nombre de sinistres

Il s'agit du nombre de sinistres déclarés par l'assuré à la compagnie d'assurance. Il est représenté dans la base de données par la variable **nbsin**.

a. La fréquence des sinistres

La fréquence annuelle des sinistres pour chaque contrat est ensuite déduite par la formule suivante :

$$f_t = \frac{N_t}{Nombre\ d'années\ assurance}, \text{ pour } t = 2004, \dots, 2012$$

Une **année assurance** est la durée de présence d'un contrat au cours d'un exercice. L'année assurance du contrat k lors de l'exercice a est donnée par :

$$(Année\ assurance)_{k\,a} = \frac{(Nombre\ de\ jours\ assurés)_{k\,a}}{366}.$$

Le **nombre d'années assurance** se déduit en sommant les **années assurance**. C'est une mesure qui correspond mieux à la définition de l'exposition au risque du portefeuille, que le

nombre même de contrats présents dans ce portefeuille. Par exemple, nous supposons que pour l'exercice n, il y ait 5 sinistres pour 10 contrats et 7 années assurance alors :

- La fréquence des sinistres selon le nombre de contrats est de $\frac{5}{10}$
- La fréquence des sinistres selon le nombre d'années assurance est de $\frac{5}{7} \approx 71\ \%$

Il est logique de considérer plus risqué un **individu A** avec 6 mois de présence dans l'année et qui a un sinistre ($fréquence_A = 2$), qu'un **individu B** avec 12 mois de présence et qui a un sinistre ($fréquence_A > fréquence_B = 1$).

b. Analyse de la fréquence des sinistres du portefeuille

Dans la **Figure** ci-dessous nous représentons la fréquence des sinistres de la base de données en fonction des différents exercices.

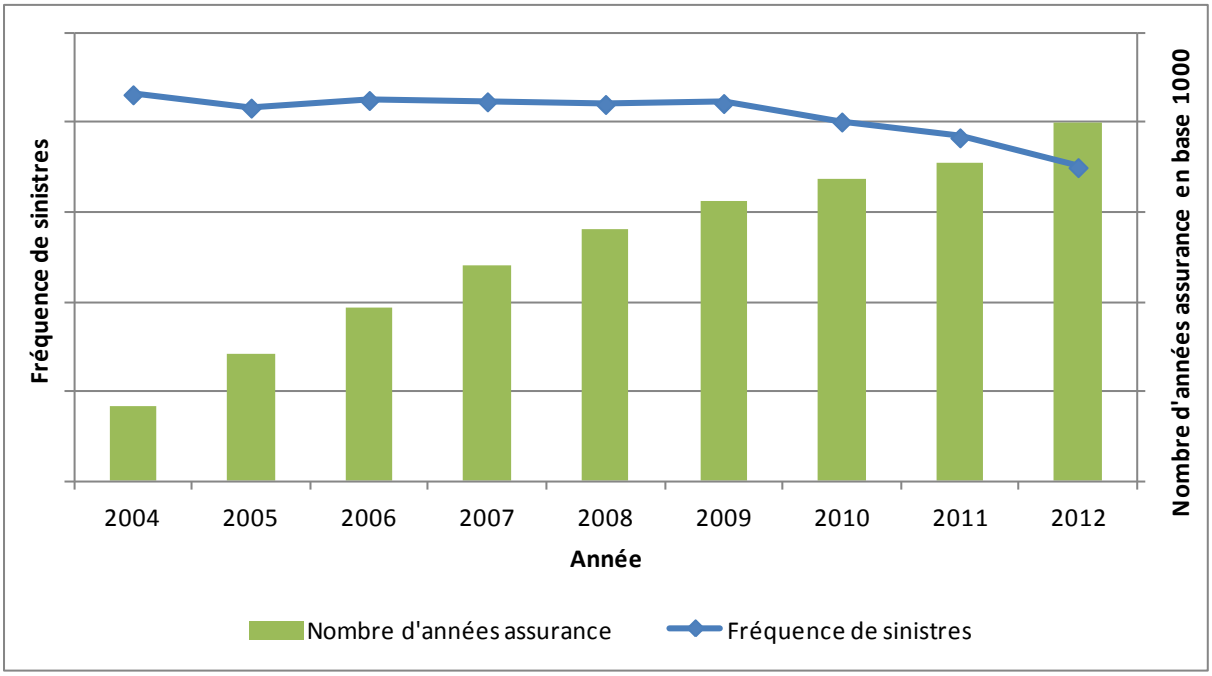


Figure 3 : Evolution de la fréquence des sinistres du portefeuille AA entre 2004 et 2012

La courbe en bleu donne la fréquence des sinistres de chaque exercice dans le portefeuille AA étudié. L'histogramme en vert correspond au **nombre d'années assurance** du portefeuille pour chaque exercice. Chaque bâtonnet de l'histogramme traduit le poids de l'exercice sur la période d'étude. La croissance du **nombre d'années assurance** signifie qu'il y a davantage d'affaires nouvelles (souscriptions nouvelles) que de résiliations dans le

portefeuille. Cette évolution traduit le gain de part de marché de PACIFICA qui commercialise le produit AA depuis 2003. D'après la **Figure** ci-dessus, la fréquence des sinistres était stable entre 2005 et 2009. Par contre, elle décroit régulièrement depuis 2010 et ce malgré l'augmentation du nombre de contrats. Cette évolution coïncide avec l'intégration à cette date (2009) de coefficients d'ajustement techniques censés prendre en compte la sinistralité individuelle.

B. Le montant des sinistres

Le montant des sinistres est l'autre élément principal de la sinistralité d'un contrat. Cette section présente de façon succincte la différence faite en pratique entre les sinistres ordinaires (ou *attritional claims* en Anglais) et les sinistres graves (*atypical claims*). Cela amènera ensuite, sur la détermination d'un seuil de distinction entre sinistres ordinaires et sinistres graves pour le portefeuille étudié. Enfin, une analyse descriptive du coût moyen du portefeuille est examinée dans un troisième paragraphe.

a. Sinistres ordinaires vs sinistres graves

L'analyse mathématique de la sévérité d'un sinistre conduit à distinguer les sinistres **ordinaires** des sinistres **graves**. Les sinistres ordinaires sont ceux qui surviennent le plus souvent dans le portefeuille avec un montant de charge ne dépassant pas un certain seuil. Tandis que les sinistres graves surviennent avec une faible fréquence, mais ont une sévérité très importante. Suivant le type de sinistralité étudiée, les outils statistiques sont différents. L'évaluation des sinistres graves fait notamment appel à la **Théorie des Valeurs Extrêmes** [3]. Le présent mémoire porte sur la sinistralité « ordinaire » du produit AA, il ne sera donc pas question de modéliser la sinistralité au-delà d'un seuil. En effet, l'objectif final étant d'étudier l'impact de la sinistralité individuelle, un postulat de départ consiste à considérer qu'un individu est responsable en fréquence mais pas en coût. Nous ne chercherons donc pas à identifier l'impact des sinistres graves sur la sinistralité future. Toutefois, la modélisation des sinistres ordinaires nécessite la détermination d'un seuil de sinistres graves.

i. Choix du seuil des sinistres graves sur la base de données AA

Un seuil de **50 K€** est choisi par le service actuariat produit chez PACIFICA, pour distinguer les sinistres ordinaires des sinistres graves. Ce seuil est appliqué sur différents produits dont le produit AA. Le choix de ce seuil a été fait en comparant l'ensemble des facteurs de risque discriminants pour les sinistres graves et pour les sinistres ordinaires, suivant différents seuils. Par cette méthode fastidieuse mais pratique, il s'est avéré que le choix d'un seuil de 50 K€ pour distinguer les sinistres ordinaires des graves était convenable. Nous proposons dans ce mémoire différentes méthodes alternatives pour vérifier que le seuil de 50 K€ peut être conservé dans notre étude.

Historique des sinistres AA de 2004 à 2012

Le jeu de données est constitué de l'ensemble des charges de sinistres AA entre 2004 et 2012. Nous disposons ainsi 170 000 observations de charge brute de sinistres.

Minimum	Moyenne	Maximum
0,01	1656,795	3940420

Tableau 4 : La charge de sinistres AA

Nous observons sur l'historique des montants de sinistres qu'il y a des écarts importants entre le minimum, la moyenne et le maximum des montants de sinistres. Nous étudions ensuite la distribution de la charge de sinistres à travers la représentation de la fonction de répartition.

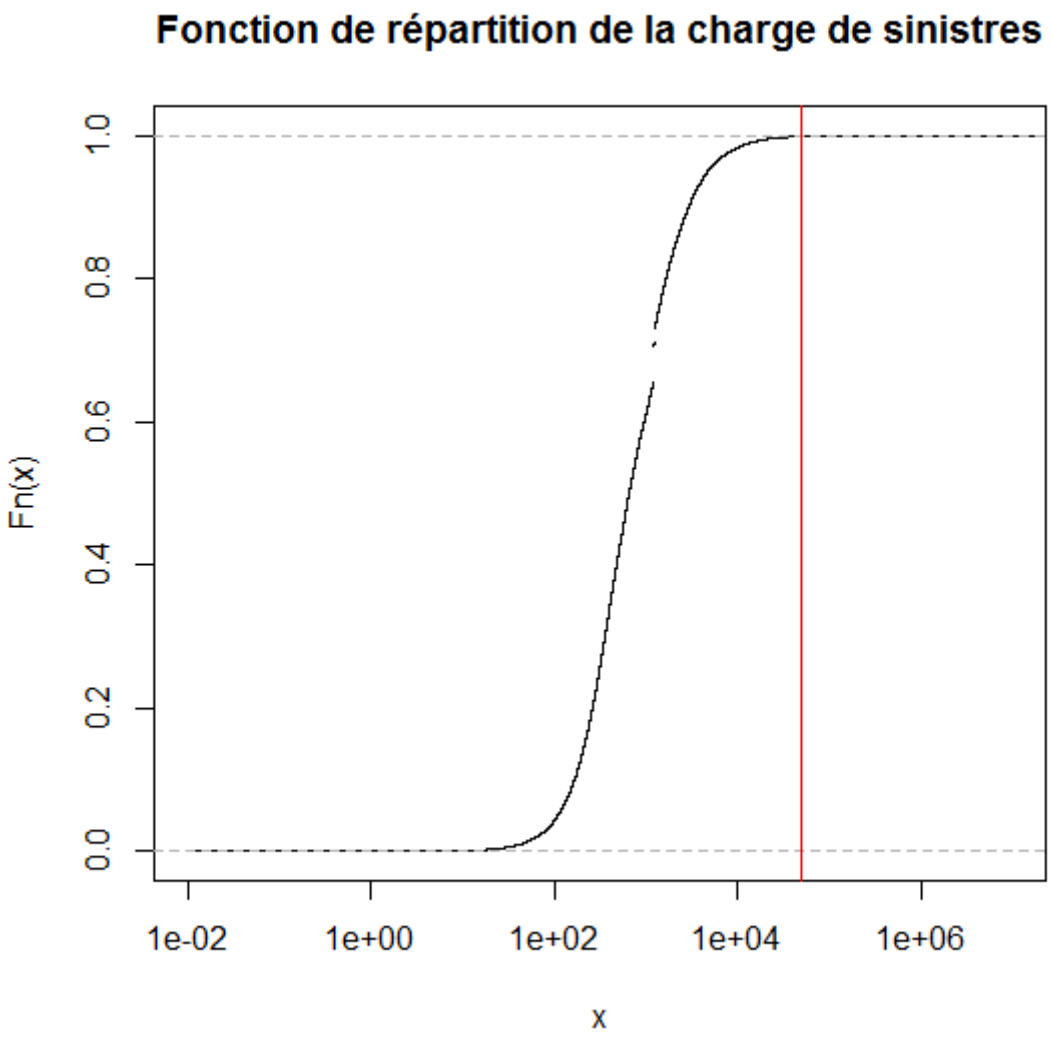


Figure 4 : Etudes de la distribution de la charge de sinistres AA

La masse des sinistres AA est concentrée sur des montants bien inférieurs à 1 million, d'après la fonction de densité. Sur la fonction de répartition de la charge de sinistres, nous observons que la charge de sinistres à 50 K€ est proche d'un quantile à 99%.

Le **Tableau** suivant donne le nombre d'observations supérieures à différents seuils.

Seuil	10000	20000	30000	40000	50000	60000	70000	80000	90000
nb	2888	998	575	356	255	185	156	127	107
Fréquence relative	1,72%	0,59%	0,34%	0,21%	0,15%	0,11%	0,09%	0,08%	0,06%

Tableau 5 : Exemples de seuils

ii. Application de la théorie des valeurs extrêmes

En assurance la modélisation des gros sinistres se fait dans le domaine d'attraction de Fréchet. Or dans ce domaine, la loi asymptotique des excès au-delà d'un seuil donné, est une loi de Pareto généralisée indexée par un paramètre γ et un deuxième paramètre σ. Ces paramètres, appelés respectivement indice de valeur extrême (ou paramètre de forme de la loi de Pareto généralisée) et paramètre d'échelle, apportent une information sur la forme de la queue de distribution des excès. De plus, la valeur du paramètre γ définit le domaine d'attraction :

- Domaine de Fréchet ($\gamma > 0$).
- Domaine de Weibull ($\gamma = 0$).
- Domaine de Gumbel ($\gamma < 0$).

Soit $X_1, \dots, X_n$ la liste des observations représentant l'historique des montants de sinistres AA sur la période de 2004 à 2012. La suite $\left(X_{n-k+1,n}\right)_{j=1,\dots,n}$désigne la statistique d'ordre des observations.

- ***Estimation de Hill***

Nous utilisons la méthode d'estimation de l'indice V.E par Hill pour déterminer le seuil des sinistres graves AA.

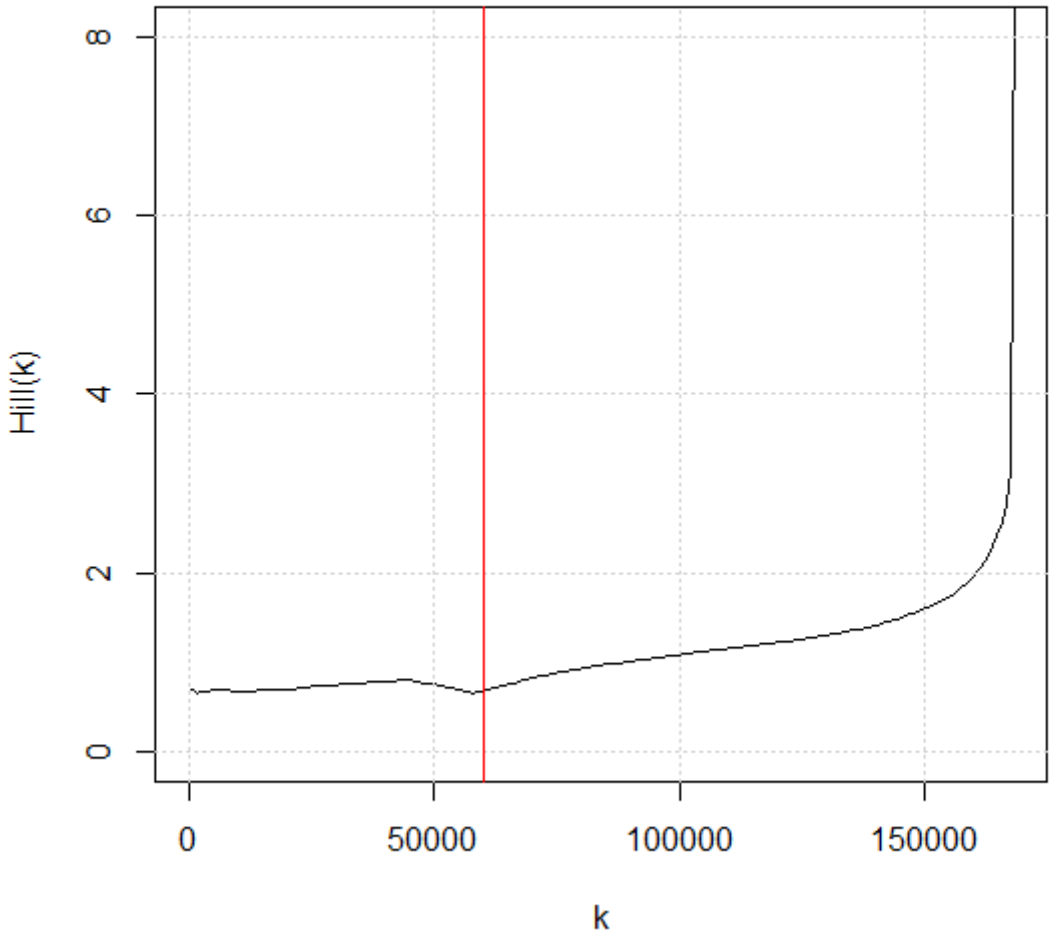


Figure 5 : Graphique de l'estimateur de Hill en fonction du seuil k

La représentation ci-dessus, nous donne les différentes valeurs de l'estimation de γ par la méthode de Hill. Nous observons que l'estimation est relativement stable jusqu'à un seuil de 60 K€. Nous en déduisons d'après cette méthode que l'on peut prendre un seuil de graves jusqu'à 60 K€.

- ***Le Pareto Quantile Plot***

C'est une représentation très utile pour visualiser graphiquement si les observations sont distribuées selon une loi du domaine de Fréchet. Dans ce domaine, le graphe $\left(\log\frac{n+1}{k}, \log X_{n-k+1,n}\right)$ serait approximativement linéaire avec une pente de γ, pour les

petites valeurs de k, c'est-à-dire les points extrêmes. Nous construisons ce graphe sur la base des sinistres AA.

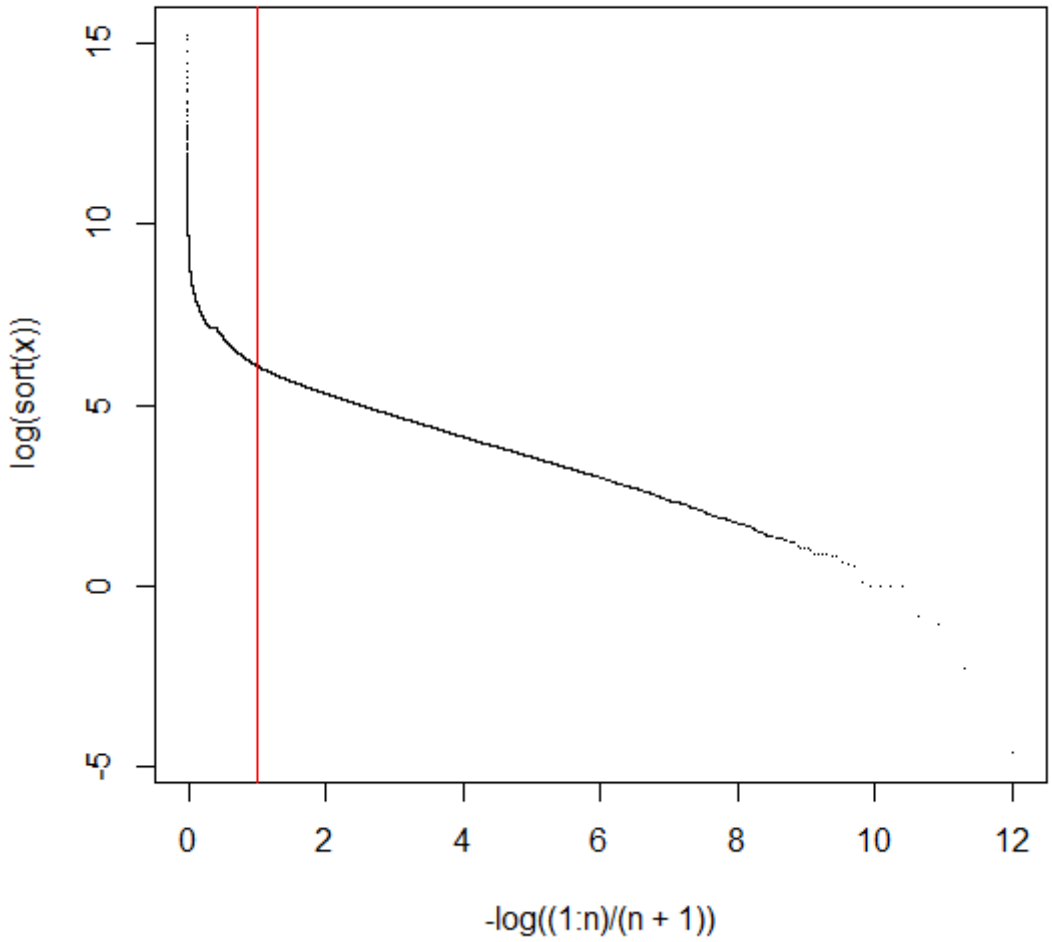


Figure 6 : Le graphique du Pareto quantile plot

D'après la représentation ci-dessus, nous déduisons que le seuil se situe au tour de 1 en échelle $\log(.)$, ce qui correspond à un montant proche de 50K€.

- ***Méthode POT (Peaks Over Threshold)***

On cherche à nouveau à déterminer le seuil optimal à l'aide de la méthode POT. Celle-ci estime les paramètres précédemment cités (γ et σ) selon différents seuils. Ainsi on choisira le seuil pour lequel l'estimation des paramètres est stable.

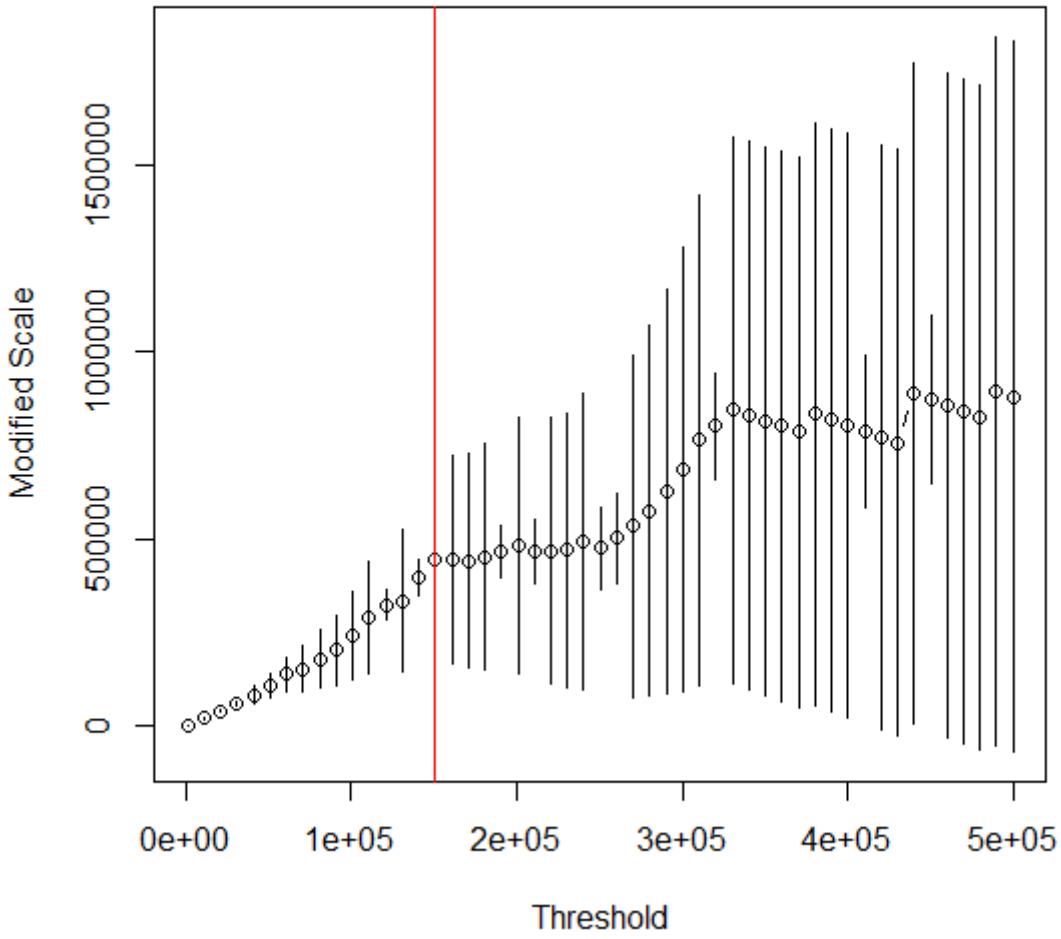


Figure 7 : Estimation du paramètre d'échelle

On observe sur la **Figure** ci-dessus que le seuil se situe à **150K €.** C'est en effet la première valeur à partir de laquelle, l'estimation du paramètre d'échelle est stable.

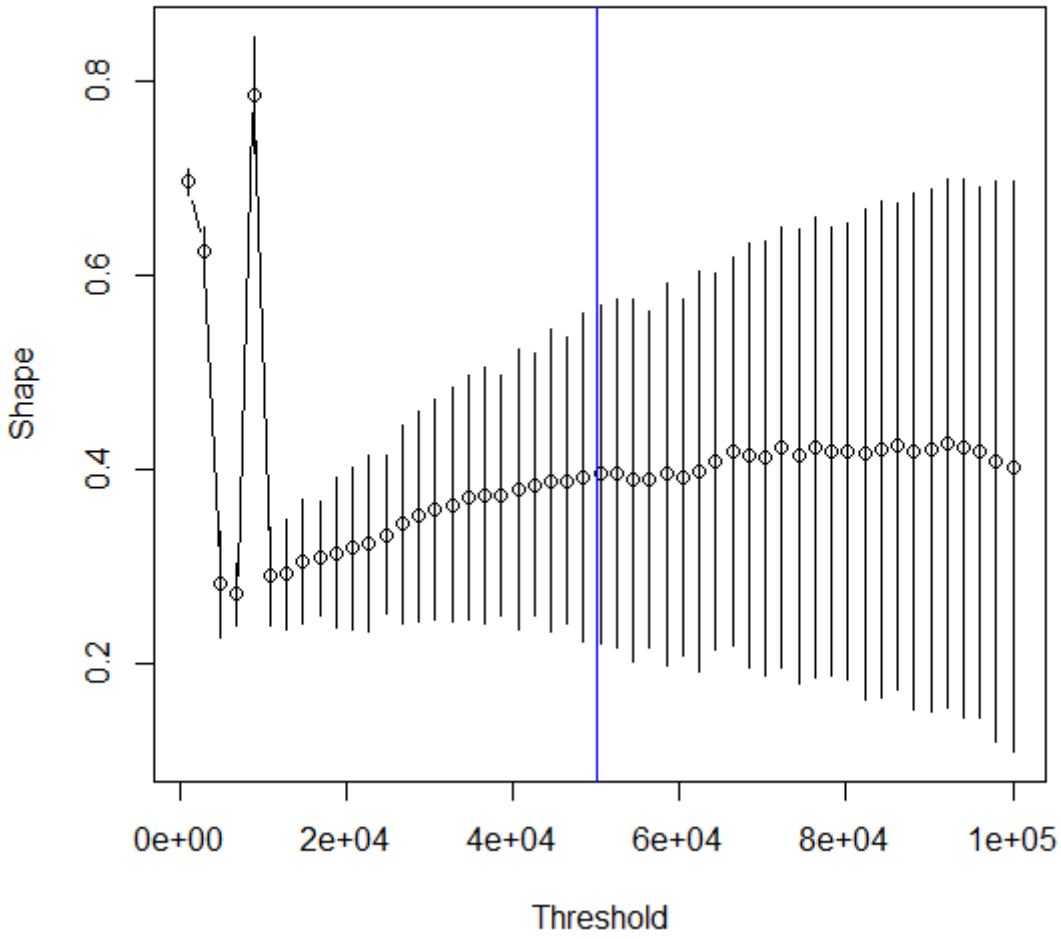


Figure 8 : Estimation du paramètre de forme de la GPD[8]

Sur la Figure ci-dessus, le seuil à partir du quel l'estimation du paramètre de forme est stable se situe à **50 K€**.

- ***Synthèse des méthodologies de choix du seuil***

Nous avons utilisé plusieurs approches pour estimer le seuil optimal permettant de distinguer les sinistres graves et les sinistres ordinaires. Chaque approche propose un intervalle de seuils acceptables. Le montant de 50K€ étant à chaque fois compris dans chacun des intervalles de définition de seuils, nous pouvons de ce fait, le conserver comme niveau de seuil de graves dans le cadre de l'étude.

[8] *GPD: General Pareto Distribution*

Notons Z, le montant des sinistres et l la valeur du seuil. Nous cherchons donc à modéliser $\mathbb{E}(Z \wedge l)$ qui définit le coût moyen sous-crête, où $Z \wedge l$ est le minimum entre Z et l.

b. *Analyse de coût moyen sous-crête des sinistres du portefeuille*

Le coût moyen sous-crête est défini empiriquement par

$$C_t = \frac{\sum_{i=1}^{N_t} Z_{i,t} \wedge l}{N_t}.$$

Pour $t = 2004, \ldots, 2012$.

La **Figure** ci-après représente une observation du coût moyen sous-crête d'un sinistre du portefeuille AA, capitalisé à fin décembre 2012.

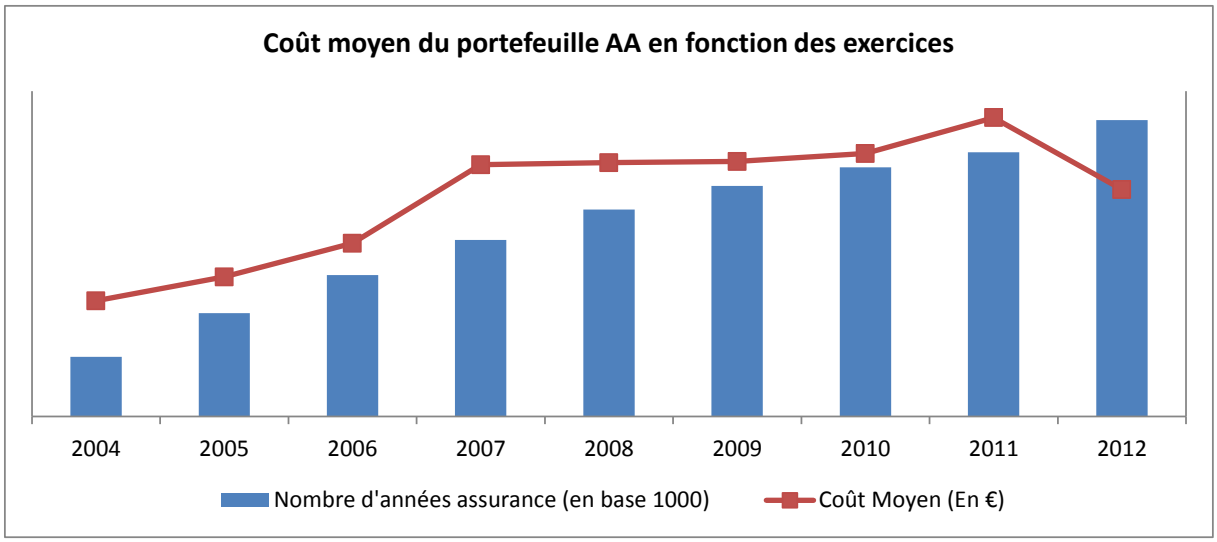


Figure 9 : Evolution du coût moyen des sinistres

De 2004 à 2007, on observe une croissance du coût moyen des sinistres du portefeuille, qui se stabilise par la suite. Cette évolution est liée à l'évolution de la structure du portefeuille AA, qui était ainsi importante dans les premières années puis s'est stabilisée ensuite.

2.1.2.2 Variables discriminantes principales

A. L'activité de l'assuré

L'assurance *Automoteur Agricole* s'adresse aux agriculteurs ainsi qu'aux artisans, commerçants et professions libérales. Le domaine d'activité renseigne sur le mode

d'utilisation de l'automoteur. Les activités recensées dans le portefeuille sont définies dans le **tableau suivant** :

Code variable	Définition
AUTR	*Regroupement d'activités*
CULT	***Cultivateur***
HORT	***Horticulteur***
INTE	*Elevage Intensif agricole*
PLAI	*Plaisancier*
POEL	***Polyculteur Eleveur***
RETR	*Retraité agricole*
VITI	***Viticulteur***

Tableau 6 : Liste des activités des personnes assurées dans le portefeuille AA

La modalité **AUTR** désigne un regroupement d'activités peu représentées dans le portefeuille AA. La plupart des activités sont liées au monde agricole d'après le **Tableau** (modalités en gras). Les principaux souscripteurs sont des agriculteurs bien que l'assurance *Automoteur Agricole* s'adresse à un large rayon de professions. Dans la **Figure** ci-dessous nous analysons le comportement de la sinistralité en fonction des différentes modalités de la variable activité.

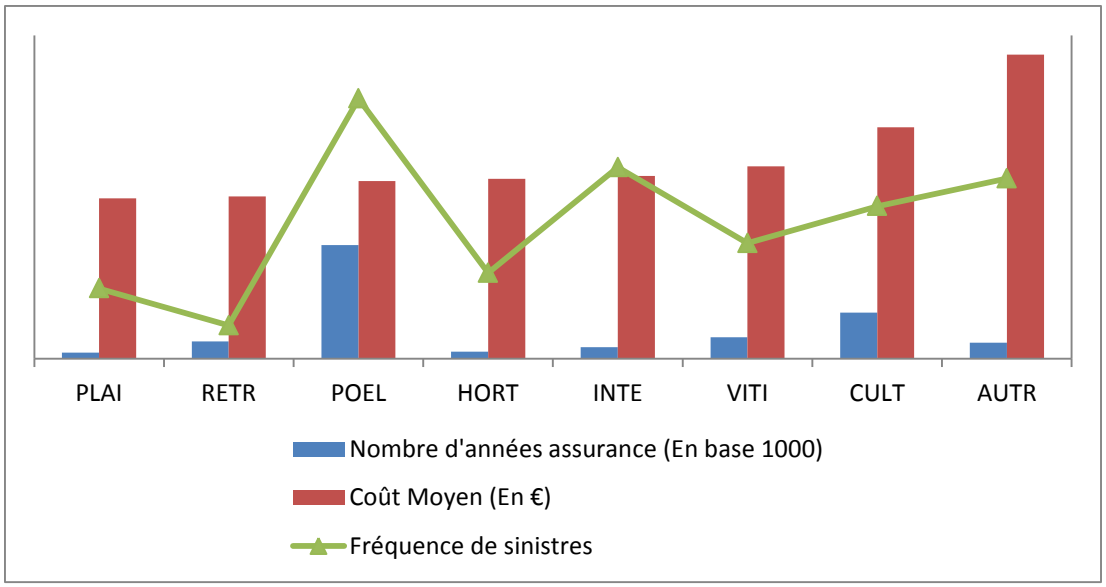


Figure 10 : La sinistralité selon l'activité

Nous observons que les polyculteurs éleveurs (**POEL**) sont les assurés les plus représentés du portefeuille AA étudié, suivis des cultivateurs (**CULT**) et des retraités agricole (**RETR**). Cette observation est faite en analysant le **nombre d'années assurance** par occurrence.

Les polyculteurs éleveurs ont aussi la fréquence des sinistres la plus élevée du portefeuille. Ils sont suivis, sur ce point, des éleveurs intensifs (**INTE**) et des cultivateurs. Sans compter les branches d'activités regroupées dans la variable **AUTR**, les retraités agricoles présentent la fréquence de sinistres la plus faible du portefeuille.

Le coût moyen des sinistres est réparti de façon équilibrée selon les activités, sauf pour les cultivateurs et l'ensemble d'autres activités peu représentées dans le portefeuille (**AUTR**) qui présentent un coût moyen supérieure. L'activité est donc discriminante pour la fréquence des sinistres, mais semble peu discriminante pour le coût moyen des sinistres.

B. La variable formule

Nous rappelons que la couverture AA est un produit multirisque dans la mesure où elle propose plusieurs types de garanties. Ces garanties sont souscrites sous forme de packages représentés par la variable **formule,** dont nous rappelons les modalités décrites dans le chapitre 1.

- **MINI**
- **MEDIAN**

- **MAXI**

La **Figure ci-dessous** représente la sinistralité selon le type de **formule** souscrite dans le portefeuille AA.

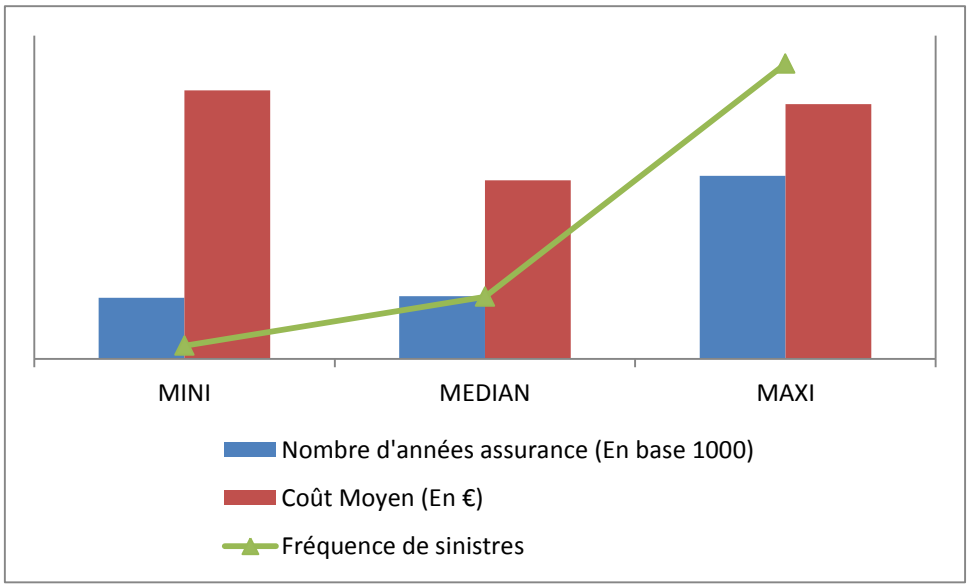


Figure 11 : La sinistralité selon la formule

La formule **MAXI** est la plus représentée du portefeuille. Elle a une fréquence des sinistres supérieure aux autres formules. Le nombre de garanties contenues dans une formule traduit aussi l'exposition au risque du contrat. Il est donc logique que l'exposition au risque d'un contrat ayant choisi la formule **MAXI** soit supérieure à celle ayant choisi la formule **MEDIAN**. On observe également que le coût moyen du portefeuille en **MINI** est plus important que le coût moyen en **MEDIAN**. Cela s'explique également par le fait que l'exposition au risque de la formule **MEDIAN** est supérieure à celle de la formule **MINI**. Cette exposition a donc tendance à « sous-pondérer » le coût moyen de la **MEDIAN** par rapport à la **MINI**.

C. L'âge de l'automoteur

C'est une variable quantitative à valeur entière qui traduit la vieillesse de l'automoteur. L'âge de l'automoteur est catégorisé pour en faciliter l'étude. La variable catégorielle désignée est **agev**. Par exemple, un automoteur ayant 13 ans, appartient à la tranche d'âge des 10 à 15 ans. La **Figure** ci-dessous présente la sinistralité du produit AA en fonction de l'âge de l'automoteur agricole.

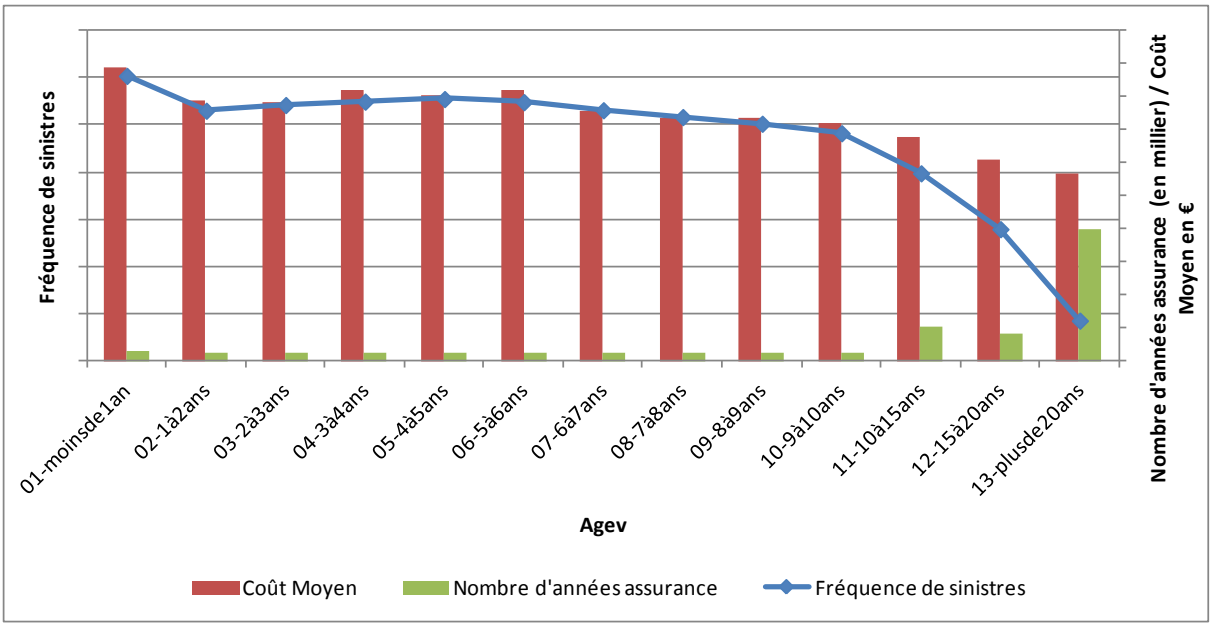


Figure 12 : Sinistralité en fonction de l'âge de l'automoteur

D'après la **Figure** ci-dessus, les automoteurs les plus représentés dans le portefeuille sont ceux de plus de 20 ans. On observe aussi que la fréquence des sinistres décroît avec l'âge de l'automoteur de façon significative pour les automoteurs de plus de 10 ans. Plus un automoteur est âgé moins il est utilisé au profit des automoteurs plus récents. Ces derniers sont donc plus exposés au risque. Le **comportement** du coût moyen par rapport à la vétusté de l'automoteur est semblable à celle de la fréquence. Il serait donc judicieux de regrouper certaines modalités de la variable **agev** qui présentent des caractéristiques de sinistralité semblables.

D. La puissance de l'automoteur

C'est un indicateur de la taille de l'engin agricole. La puissance est mesurée en cheval vapeur (**CV**). Elle est représentée dans la base de données par la variable puissance de l'automoteur dont les modalités correspondent à des classes de puissance. La **Figure** suivante présente comment la sinistralité du portefeuille est répartie en fonction de la variable puissance de l'automoteur.

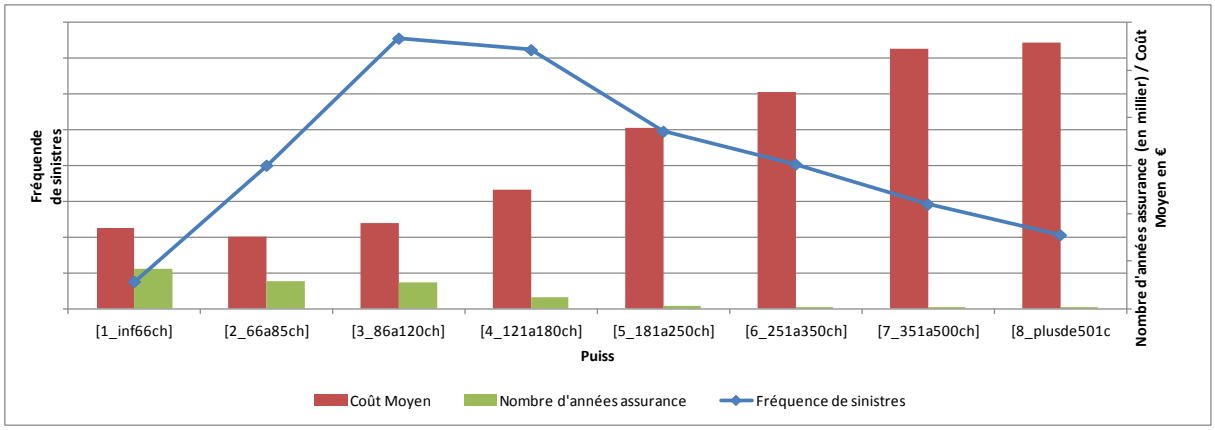


Figure 13 : Sinistralité en fonction de la puissance de l'automoteur

Le portefeuille AA contient peu d'automoteurs de plus de 181 CV. On observe en effet une décroissance du nombre d'années assurance en fonction de la puissance. Le coût moyen des sinistres augmente avec la taille de l'automoteur agricole car ce sont les matériels plus plus chers et les plus sophistiqués. La fréquence des sinistres croît également avec la puissance de l'automoteur, puis décroît pour les automoteurs de plus de 120 CV. Les automoteurs de grande taille sont généralement utilisés pour des tâches plus ponctuelles que les autres automoteurs. De ce fait, on explique l'évolution de la fréquence des sinistres pour les automoteurs de grande taille par le fait de « l'usage ».

E. La famille d'engins

La famille d'engins correspond à la catégorie à laquelle l'automoteur appartient. Les catégories sont établies en tenant compte de plusieurs caractéristiques de l'automoteur comme :

- La vitesse maximale
- La masse à vide
- La garde au sol
- L'usage

Par exemple, la **Figure suivante** présente trois types de catégories de tracteurs établies par le **code de la route (art. R311-1)**.

LES CATEGORIES DE TRACTEURS → T : tracteur à roues
→ C : tracteur à chenilles

Catégorie	Vitesse maxi par construc.	Voie mini	Masse à vide	Garde au sol
T1 ou C1 tracteurs standards	≤ 40km/h	≥ 1150 mm	> 600 kg	≤ 1000 mm
T2 ou C2 (*) tracteurs à voie étroite	≤ 40km/h	< 1150 mm	> 600 kg	≤ 600 mm
T3 ou C3 micro tracteurs	≤ 40km/h	-	≤ 600 kg	-

(*) Lorsque la valeur de la hauteur du centre de gravité du tracteur (mesurée par rapport au sol), divisée par la moyenne des voies minimales de chaque essieu est supérieure à 0,90, la vitesse maximale par construction est **limitée à 30 km/h** (tracteurs très étroits et/ou à haute garde au sol).

Figure 14 : Exemples de catégories de tracteurs selon le code de la route (art. R311-1)

Dans le portefeuille étudié, les automoteurs sont répartis suivant 15 catégories. La **Figure** suivante présente comment la sinistralité du portefeuille est répartie selon la famille d'engins (représentée par la variable **fameng**).

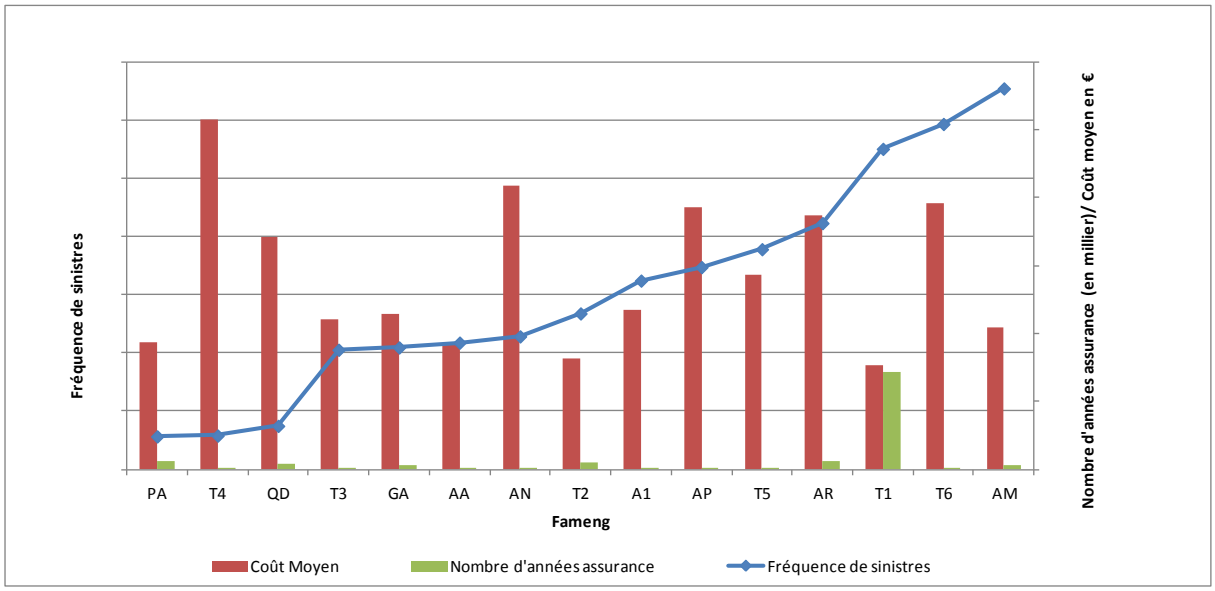


Figure 15 : Sinistralité en fonction de la famille d'engins

Nous observons une forte hétérogénéité en termes de sinistralité entre les différentes familles d'engins agricoles. La variable **fameng** est aussi bien discriminante en termes de coût moyen des sinistres que de fréquence des sinistres. Cependant, le portefeuille AA n'est

constitué en majorité que d'automoteurs de type T1, et de PA, AR, T2 dans une bien moindre mesure.

2.1.2.3 Synthèse

L'analyse descriptive des variables du portefeuille nous a permis d'avoir une première idée des profils de risque AA. Suivant les composantes de la sinistralité (fréquence et coût moyen) les variables telles que l'activité et la formule sont discriminantes pour l'une et pas nécessairement pour l'autre. Ce constat nous conforte dans le choix de modéliser séparément la fréquence des sinistres et le coût moyen des sinistres AA.

2.2 Modélisation de la sinistralité : théorie des GLMs

2.2.1 Présentation de la théorie

2.2.1.1 Origines et définition formelle

A. Histoire des modèles linéaires généralisés

Les modèles linéaires généralisés (*generalized linear models* en anglais) ont été introduits par John NELDER et Robert WEDDERBURN en 1972. L'objectif était d'unifier des méthodes statistiques comprenant **la régression linéaire, la régression logistique et la régression de Poisson**. Les GLMs fournissent un cadre théorique de modèles de régression pour une grande variété de distributions, allant bien au-delà de la seule loi Normale des modèles linéaires.

Les techniques GLM constituent un outil de choix par les actuaires pour l'élaboration d'une tarification en assurance automobile. Elles permettent d'analyser un grand nombre de phénomènes, voir [4] et [5].

B. Limites des méthodes classiques

Les modèles GLM permettent de lever certains obstacles rencontrés dans l'application d'analyse descriptive univariée. En effet, l'analyse à un facteur ne permet pas de prendre en compte l'effet d'autres facteurs sur le phénomène étudié. Par exemple, l'analyse à un facteur de la fréquence des sinistres d'un véhicule en fonction de sa vétusté ne prend pas en compte le fait que les jeunes conducteurs possèdent en majorité des véhicules anciens. Or les jeunes conducteurs sont les assurés les plus risqués selon la classe d'âge. Ainsi le fait que les jeunes conducteurs possèdent majoritairement des véhicules anciens contribuent à accroître la fréquence des sinistres attendue de ces véhicules. De ce fait, l'approche univariée sépare mal les effets marginaux des variables tarifaires lorsqu'elles sont corrélées entre elles.

D'autre part, les hypothèses de base des modèles linéaires gaussiens sont difficilement tenables dans le cadre d'une application en assurance. Pour un phénomène tel que le montant des sinistres, la distribution est à valeur dans $\mathbb{R}_+$ et non $\mathbb{R}$ tout entier. De plus, la loi

est généralement asymétrique, ce qui n'est pas le cas de la loi Normale. Les modèles GLM permettent de lever ces nombreuses limites en reposant sur des hypothèses plus réalistes.

C. Les lois de probabilités possibles

On cherche à modéliser une variable Y à l'aide d'un certain nombre de variables explicatives $X = (X_1, \dots, X_p)^t$. Pour rappelle, la régression linéaire suppose que

$$Y \sim \mathcal{N}(\mu, \sigma^2) \text{ Où } \mu = X^t\beta$$

Cette représentation est généralisée par les modèles GLM de la manière suivante :

$$Y \sim \mathcal{Loi}(\mu, \sigma^2) \text{ Où } \mu = \mathbb{E}(Y) = g^{-1}(X^t\beta)$$

Où g est la *fonction lien* et $\mathcal{Loi}$ désigne une loi appartenant à la **famille exponentielle** [6]. L'application des GLMs pour modéliser la sinistralité du risque AA est précédée dans ce mémoire, par la présentation de plusieurs concepts qui constituent l'ossature de cette technique statistique.

2.2.1.2 La famille exponentielle

Dans les modèles GLM comme dans tout modèle de régression, une hypothèse de loi est faite sur l'aléa sous-jacent à la variable d'intérêt. Il a été présenté précédemment que les modèles linéaires faisaient une hypothèse de **loi Normale**. Les modèles GLM élargissent le spectre des lois d'application, en supposant qu'elles appartiennent à la famille exponentielle. Cette généralisation rend de ce fait plus applicable les modèles GLM notamment dans le domaine assurantiel. La fonction de masse ou probabilité d'une loi appartenant à la famille exponentielle est structurée de la manière suivante:

$$f(y, \theta, \phi, \omega) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi, \omega)\right\}, y \in \mathcal{S}$$

- $\mathcal{S}$ est un sous ensemble de $\mathbb{N}$ ou de $\mathbb{R}$
- $a(\phi) = \frac{\phi}{\omega}$,
- $\phi > 0$: *Paramètre de dispersion*
- $\omega \geq 0$: *Poids connu a priori*[9]

[9] Pondération d'une observation, liée à sa «taille » (volume, durée d'observation, effectifs, …etc.) soit à la crédibilité qu'on souhaite lui accorder.

- b, la *fonction de cumulant*
- θ : Paramètre à valeur dans $\mathbb{R}$ et lié à l'espérance de la loi.

Les moments d'ordre 1 et 2 sont définis respectivement par :

$$\mathbb{E}(Y) = b'(\theta) \text{ Et } \mathbb{V}(Y) = b''(\theta) \times \frac{\phi}{\omega}$$

Pour illustration, le **Tableau** suivant présente quelques lois usuelles ainsi que leur fonction lien canonique associée.

	Normale	Poisson	Gamma	Inverse Gaussien
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{G}(\mu, v)$	$\mathcal{IG}(\mu, \sigma^2)$
Support	$\mathbb{R}$	$\mathbb{N}$	$]0; +\infty[$	$]0; +\infty[$
$\boldsymbol{\phi}$	$\frac{\sigma^2}{2}$	$Exp(\theta)$	$-\log(-\theta)$	$-(-2\theta)^{-\frac{1}{2}}$
$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}(\boldsymbol{\theta})$	θ	$Exp(\theta)$	$\frac{1}{\theta}$	$(-2\theta)^{-\frac{1}{2}}$
$\boldsymbol{g_c}$	Identité	log	$Inverse$	$1/\mu^2$

Tableau 7 : Exemples de lois appartenant à la famille exponentielle et leur fonction lien canonique

En résumé, pour choisir un modèle GLM il faut donc :

- Choisir une loi pour la variable d'intérêt Y dans la famille exponentielle
- Choisir une fonction lien inversible g

Pour utiliser un modèle GLM il faut estimer les paramètres β, ϕ.

2.2.1.3 La fonction lien

La fonction lien g définit la relation entre la variable d'intérêt Y, et les variables explicatives X en modélisant l'effet de ces dernières composantes sur Y. Cette fonction g est supposée **monotone et différentiable**.

A chacune des lois de la famille exponentielle, on associe une fonction lien: c'est la **fonction lien canonique**. Elle a plus un intérêt théorique que pratique. Selon le phénomène que l'on souhaite étudier, certaines fonctions de liens sont plus adaptées que d'autres. Le **Tableau** suivant présente des exemples de fonctions de lien selon les applications.

Nom du lien	Fonction lien	Application
Lien identité	$g(\mu) = \mu$	*Modèle linéaire classique*
Lien log	$g(\mu) = \log(\mu)$	*Modèles multiplicatifs (Fréquence - Coût moyen)*
Lien logit	$g(\mu) = \log\frac{\mu}{1-\mu}$	*Modèle de Scoring*

Tableau 8 : Exemples d'utilisation

2.2.1.4 La fonction de vraisemblance

En pratique, les **coefficients de régression** $\beta_1, \beta_2, \dots \beta_p$ et le **coefficient de dispersion** ϕ ne sont pas connus et doivent donc être estimés sur la base de données. La section suivante porte sur l'estimation des coefficients de régression par la méthode de maximum de vraisemblance.

A. Définition

On considère n individus sur lesquels sont observées $y_1, y_2, \dots, y_n$, les n observations de la variable à expliquer et $X_{11}, X_{21}, \dots, X_{n1}, X_{12}, X_{22}, \dots, X_{1p}, X_{2p}, \dots, X_{np}$ les n observations pour chacune des p variables explicatives. Les observations des variables explicatives sont regroupées dans une matrice X à n lignes et p colonnes. La **vraisemblance** notée $\mathcal{L}$, est la probabilité d'obtenir les observations réalisées au sein du portefeuille dans le modèle considéré,

$$\mathcal{L}(\beta) = \mathcal{L}(y, \theta, \phi, \omega) = \prod_i^n f(y_i, \theta_i, \phi, \omega_i) \text{ Où } y = (y_1, y_2, \dots, y_n)^t$$

B. Programme de maximisation de la fonction de vraisemblance

L'estimation de $\beta = (\beta_1, \beta_2, \dots \beta_p)$ par la **méthode de maximum de vraisemblance** consiste à déterminer $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_p)$ en maximisant $\mathcal{L}(\beta)$. Afin de faciliter l'obtention du maximum de vraisemblance de β nous passons à la fonction de **log-vraisemblance**.

Ainsi la détermination du coefficient β est déduite de la résolution du programme suivant :

$$\begin{cases} \max_{\beta} L(\beta) = \log \mathcal{L}(\beta) \\ \mu_i = b'(\theta_i) \\ g(\theta_i) = \eta_i = \sum_{j=1}^{p} X_{ij}\beta_j \end{cases}$$

La résolution de ce programme se fait numériquement. En pratique nous utilisons **l'algorithme de Newton-Raphson** [7]**.**

C. Estimation du paramètre de dispersion

Sauf dans le cas d'un modèle de régression de Poisson ou Binomial ($\phi = 1$), **le paramètre de dispersion** ϕ est à estimer. En pratique nous utilisons l'estimation du χ^2 de Pearson :

$$\hat{\phi} = \frac{1}{n-p-1}(y-\hat{\mu})^t \mathcal{I}_n (y-\hat{\mu})$$

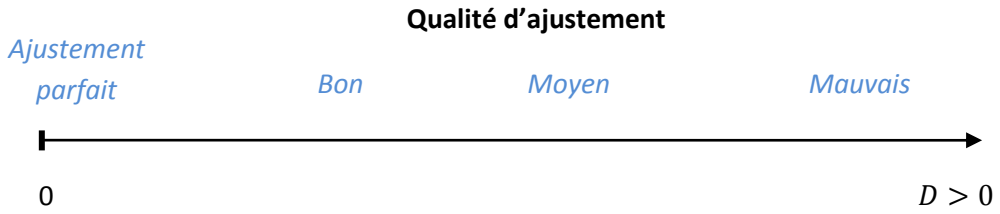
Où $\mathcal{I}_n$ est la **matrice d'information de Fisher**

2.2.2 Outils statistiques

Après avoir fait des hypothèses de loi et fonction lien pour un GLM, il est important de juger de la qualité du modèle ainsi choisi. Cette section présente différents outils statistiques qui permettent d'effectuer des contrôles sur le modèle GLM défini.

2.2.2.1 La déviance

La déviance est un écart en termes de log-vraisemblance entre **le modèle saturé** d'ajustement maximum et le modèle considéré. C'est un moyen de mesurer l'adéquation du modèle avec les données d'étude.



Nous notons $\mathcal{L}(y|y)$ la vraisemblance associée au **modèle saturé** et $\mathcal{L}(\hat{\mu}|y)$ celle associée au modèle étudié. La **déviance** est alors donnée par,

$$D = 2\phi(\log \mathcal{L}(y|y) - \log \mathcal{L}(\hat{\mu}|y)) \geq 0$$

En pratique, nous utilisons plutôt la **déviance standardisée**[10] donnée par :

$$D^* = \frac{D}{\phi}$$

Remarque:

En pratique, le **modèle saturé** n'est pas intéressant car il reproduit les observations sans les résumer, avec autant de paramètres que d'observations.

2.2.2.2 Test de déviance entre deux modèles emboîtés (test de type III)

Soit un modèle 1 (p_1 paramètres) et un modèle 2 dont les variables explicatives constituent un sous-ensemble des variables explicatives du modèle 1 (le modèle 2 a p_2 paramètres).

H_0 : Les 2 modèles sont équivalents

H_1 : les 2 modèles ne sont pas équivalents

L'objectif est de savoir si les variables explicatives supplémentaires du modèle 1 par rapport au modèle 2 constituent une amélioration significative. Pour cela nous analysons la différence de leur déviance :

$$\Delta D^* = D_2^* - D_1^*$$

Cette analyse est décrite de la manière suivante :

1. Nous choisissons un seuil (en général 5% ou 1%)
2. L'observation de ΔD^* est calculée, notons là ${\Delta D^*}_{\text{obs}}$
3. Nous notons que l'approximation de ΔD^* par une loi de χ^2 à $p_1 - p_2$ degrés de liberté est toujours acceptable. Ce qui permet de faire le test suivant :
   - Si ${\Delta D^*}_{\text{obs}} > \text{q}_{1-\alpha}(p_1 - p_2)$ alors H_0 est rejeté au profit de H_1, le modèle 1 est accepté
   - Si ${\Delta D^*}_{\text{obs}} \leq \text{q}_{1-\alpha}(p_1 - p_2)$ alors H_0 est conservé, le modèle 1 n'est pas accepté

[10] Par simplification dans la suite du mémoire, lorsque nous parlerons de déviance, il s'agira en fait de déviance standardisée.

2.2.2.3 Critères de choix

Les critères de choix tels que l'AIC [7] ou le BIC [8] sont une autre manière de comparer des modèles et de choisir le plus adéquat. Ces méthodes s'appliquent sur des modèles qui ne sont pas forcément emboîtés les uns dans les autres.

- Par définition l'**AIC** (*Akaike Informative Criterion*) pour un modèle à p paramètres est défini par:

$$AIC = -2L + 2p$$

Où L est la log-vraisemblance du modèle. Nous notons que plus la vraisemblance est grande, plus grande est donc la log-vraisemblance et meilleur est le modèle. Cependant si l'on met le nombre maximum de paramètres (ce qui est le modèle saturé) alors L sera maximum. Il suffit donc de rajouter des paramètres pour la faire augmenter. Pour obtenir un modèle de taille raisonnable il sera donc bon de la pénaliser par une fonction du nombre de paramètres, ici $2p$.

- Le **BIC** (Bayesian Informative Criterion) est un autre critère de choix défini par

$$BIC = -2L + p \log n$$

➡ Le modèle à choisir est celui qui présentera le critère de choix le plus faible.

2.2.2.4 Les résidus

Le **résidu** représente l'écart entre l'observation y_i et son estimation par le modèle. Dans le processus de contrôle d'un modèle GLM, l'analyse des résidus est une étape importante. Elle permet de :

- Vérifier que le modèle décrit bien les observations y_i
- Vérifier l'hypothèse sur la **fonction de variance** selon la distribution du modèle choisie

  Par exemples:
  - Loi Normale : fonction de variance constante (**homoscédasticité**)
  - Loi de Poisson : fonction de variance égale à la moyenne

A priori dans le cas non gaussien, les résidus bruts $y_i - \mu_i$ n'ont pas de raison d'être gaussien. Ainsi la littérature propose d'autres lois de résidus où une éventuelle normalité est envisageable.

A. Les résidus de Pearson

Les *résidus de Pearson* sont définis par :

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Par définition on standardise les résidus par la variance théorique de yi, fournie par la **fonction de variance** $V(.)$ et donc par le modèle. Ce n'est pas la variance de l'estimation $\hat{\mu}_i$ qui est un estimateur donc aléatoire.

Les *résidus de Pearson Standardisés*

$$\hat{\varepsilon}_i = \frac{y_i - \hat{y}_i}{\sqrt{\left((1 - h_{ii})V(\hat{\mu}_i)\right)}}$$

Où h_{ii} est le i$^{\text{ème}}$ terme diagonal de la matrice **chapeau** (*hat matrix*) :

$$H = D^{\frac{1}{2}}X(X^tDX)^{-1}X^tD^{\frac{1}{2}}$$

Où $X^tDX = \mathcal{I}_n$

$$D = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_n \end{pmatrix} \text{ Et } d_i = V^2(\mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 \frac{\phi}{\omega_i}$$

B. Les résidus de déviance

Les *résidus de déviance* sont définis par

$$d_i = signe(y_i - \hat{\mu}_i)\sqrt{2 \times \left(\mathcal{L}(y_i|y_i) - \mathcal{L}(\hat{\mu}_i|y_i)\right)}$$

Les *résidus de déviance standardisés* sont définis par

$$\hat{\varepsilon}_i^D = signe(y_i - \hat{\mu}_i)\sqrt{\frac{2 \times \left(\mathcal{L}(y_i|y_i) - \mathcal{L}(\hat{\mu}_i|y_i)\right)}{(1 - h_{ii})}}$$

En pratique ce sont les résidus de déviances qui sont généralement utilisés.

2.3 Application des GLMs pour la modélisation de la sinistralité AA

La section précédente a présenté les bases d'un modèle GLM. Celui-ci se décrit comme un groupement d'outils statistiques qui permettent de modéliser et d'expliquer une palette de phénomènes notamment en assurance. Dans cette section, nous appliquons ces méthodes pour modéliser la sinistralité AA en recherchant des facteurs de risque pertinents. La modélisation sera réalisée ici toutes garanties confondues. Cette approche permet en effet de tenir compte de l'hétérogénéité de comportements (ou de l'aversion au risque) des individus selon leur profil et des corrélations entre garanties qui en découlent. Par exemple, le risque RC des individus souscrivant une formule Tous Risques pourrait être différent du risque RC des individus en formule au Tiers. Plutôt que de modéliser chaque garantie et rechercher ensuite les corrélations entre elles, nous prenons le parti de modéliser ici directement la prime pure totale.

Pour déterminer des critères tarifaires, le service actuariat PACIFICA a l'habitude d'effectuer des analyses directes sur la charge de sinistres du produit. Pour étudier la pertinence des critères introduits dans l'équation tarifaire, nous proposons une approche alternative. Cette approche consiste à modéliser séparément la fréquence de sinistres et le coût moyen des sinistres. Ceci dans le but d'obtenir une meilleure compréhension des facteurs de risque qui influencent la sinistralité AA [9]. De plus, nous pourrons en déduire la « composante risque » (la fréquence ou le coût moyen) sur laquelle porte principalement l'anti-sélection.

La modélisation « fréquence x coût moyen » est une approche traditionnellement utilisée pour la tarification d'une police en assurance non vie. Nous donnons ci-dessous le principe mathématique qui lui est sous-jacent.

Soit N une variable aléatoire représentant le nombre de sinistres au cours de l'année et $(Z_i)_{i \in \mathbb{N}}$ un vecteur de variables aléatoires représentant les montants de sinistres. Le coût total vaut alors

$$Y = \sum_{i=1}^{N} Z_i$$

La prime pure définie comme $\mathbb{E}(Y)$ vaut

$$\mathbb{E}(Y) = \mathbb{E}(N) \times \mathbb{E}(Z)$$

Sous les hypothèses que

- *N et $(Z_i)_i$ sont indépendants, $N \perp Z_i$ quelque soit i*
- *Les $(Z_i)_{i \in \mathbb{N}}$ sont indépendants et identiquement distribuées (**iid**), de même loi que Z*

Remarque

Pour assurer une mutualisation et permettre à la loi forte des grands nombres de s'appliquer, l'analyse portera sur des **«** profils de risque » plutôt que sur des « contrats ». Un profil de risque est défini par la combinaison de modalités de variables tarifaires. De ce fait, la base de données qui a été précédemment construite, est résumée suivant des profils de risque.

2.3.1 Modélisation de la fréquence des sinistres

2.3.1.1 Choix du modèle

A. Distribution du nombre de sinistres

Pour modéliser le nombre de sinistres, les lois théoriques les plus fréquemment utilisées sont la loi de *Poisson* et la loi *Binomiale Négative*. La loi de *Poisson* appartient à la famille exponentielle. Elle suppose que la variance est égale à la moyenne, ce qui n'est pas toujours vrai en pratique. Cette hypothèse peut avoir des conséquences importantes sur la qualité de l'ajustement. Si la variance est supérieure à la moyenne des observations, il y a sur-dispersion de la variable d'intérêt. Cela conduit à rejeter trop souvent l'hypothèse nulle de non significativité des variables explicatives du modèle. Cette sur-dispersion est prise en compte par la loi *Binomiale Négative*. Nous utilisons donc cette loi pour la modélisation du nombre de sinistres.

Le **Tableau** suivant donne les principales caractéristiques des deux lois présentées.

	Loi de Poisson $\mathcal{P}(\lambda)$	Loi Binomiale Négative $\mathcal{NB}(r,p)$
$\mathbb{P}(\boldsymbol{N=k})$	$e^{-\lambda}\frac{\lambda^k}{k!}$	$\binom{r+k-1}{k}p^r(1-p)^k$
$\mathbb{E}(\boldsymbol{N})$	λ	$\frac{r(1-p)}{p}$
$\mathbb{V}(\boldsymbol{N})$	λ	$\frac{r(1-p)}{p^2}$

Tableau 9 : Présentation de la loi de Poisson et de la loi Binomiale Négative

En **annexe E**, nous testons l'adéquation aux lois théoriques citées des nombres de sinistres observés dans la base de données d'étude.

B. Fonction lien logarithmique

Nous optons pour la **fonction lien logarithmique.** En effet, l'équation tarifaire AA correspondant à la cotisation technique, présentée dans le chapitre 1, a une structure multiplicative. Le modèle multiplicatif établi, simplifie l'interprétation de l'influence des critères tarifaires sur la fréquence des sinistres. En effet, les coefficients β correspondent dès lors à des coefficients multiplicateurs.

C. Le modèle Binomiale Négative

Dans la base de données d'étude, soit N_i le nombre de sinistres déclarés par la $i^{ième}$ observation et X_i son profil de risque associé. Le modèle de régression Binomiale Négative est donné par :

$$\mathbb{P}(N_i = n_i | X_i) = \frac{\Gamma(n_i+\upsilon)}{\Gamma(n_i+1)\times\Gamma(\upsilon)} \times \left(\frac{\upsilon}{\upsilon+\lambda_i}\right)^{\upsilon} \times \left(\frac{\lambda_i}{\upsilon+\lambda_i}\right)^{\lambda_i}, \text{ pour } n_i \in \mathbb{N}$$

En posant $\alpha = 1/\upsilon$, l'espérance et la variance s'expriment par :

$$\mathbb{E}(N_i | X_i) = \lambda_i = \omega_i e^{{X_i}^t \beta} \text{ Et } \mathbb{V}(N_i | X_i) = \lambda_i (1 + \alpha\lambda_i)$$

Si $\alpha = 0$, le modèle Binomiale Négative correspond au modèle de Poisson. Cette astuce permet d'utiliser l'approche GLM pour la loi Binomiale Négative bien que cette dernière n'appartienne pas à la famille exponentielle [4].

Le **poids a priori** ω_i représente la variable **Nombre d'années assurance**.

Nous utilisons les logiciels **SAS** et **Emblem** [10] pour calibrer les GLMs.

2.3.1.2 Sélection des variables explicatives significatives

Après avoir choisi la loi de l'aléa et la fonction lien, il est primordial d'analyser les variables explicatives. Un modèle de régression est d'autant meilleur que les variables explicatives considérées sont significatives. Cela signifie qu'elles expliquent le phénomène étudié (**la variable à expliquer**) de façon pertinente. Dans un ensemble de variables explicatives

considérées en amont, nous omettrons donc les variables qui ne sont pas significatives. De telles variables ont un impact aléatoire sur la variable à expliquer.

Dans ce mémoire, les variables significatives seront sélectionnées parmi un ensemble de variables considérées par les OAT (Outil d'Analyse Tarifaire). Ces derniers sont des suivis Excel construits pour analyser l'ensemble des critères tarifaires impactant les résultats techniques d'un produit. Ces suivis Excel contiennent donc des critères qui sont déjà inclus dans l'équation tarifaire AA, mais également d'autres qui peuvent potentiellement l'être.
Concrètement, la sélection des variables s'effectue via une procédure ***Stepwise Regression*** dont le principe est rappelé dans le paragraphe suivant. Puis nous présentons l'ensemble des variables retenues pour la modélisation de la fréquence des sinistres.

A. Stepwise Regression

L'objectif de la régression Stepwise est de choisir un sous-ensemble des variables tarifaires étape par étape dans un modèle qui est simple, significatif et donne une bonne prédiction. Il existe trois types de procédure *Stepwise Regression*.

a. Forward selection

❖ **Le processus**

On dispose au départ de l facteurs de risque dans la base de données d'étude.

Etape 1 : Modèle à une variable
Dans le **modèle de départ** qui ne contient aucune variable explicative, on ajoute la première variable qui minimise la statistique AIC, lorsqu'elle est rajoutée dans le modèle de départ.

⋮

Etape p : Modèle à p variables
Une $p^{ème}$ variable est rajoutée au modèle à $p-1$ variables de la même manière que les étapes précédentes.

❖ **Arrêt du processus**

Le processus s'arrête :

- Lorsqu'il n'y a plus de variable explicative
- Lorsqu'aucune variable n'améliore significativement le critère AIC

b. *Backward selection*

Contrairement à la méthode *forward selection*, nous considérons initialement un modèle avec toutes les variables. A chacune des étapes suivantes, la variable la moins significative est retirée du modèle selon le test considéré (ici AIC). Le processus s'arrête lorsque toutes les variables explicatives restantes sont significatives.

c. *Bidirectional elimination*

Elle est semblable à la procédure *forward selection* sauf que l'on peut éliminer des variables déjà introduites. En effet, il peut arriver que des variables introduites en début ne soient plus significatives après introduction de nouvelles variables.

d. *Liste des variables retenues*

Suite à l'utilisation de la procédure *forward selection*, nous retenons dans le modèle de la fréquence des sinistres sept variables qui sont présentées dans le **Tableau** suivant :

Code variable	***Définition***
formule	*La formule choisie par l'assuré*
puiss	*Puissance de l'automoteur*
agev	*Age du véhicule*
fameng	*Famille d'engins*
dpt	*Département*
FRCHBDG	*Franchise Bris de glace*
activit	*activité*

Tableau 10 : Liste des variables retenues pour la régression

Dans le **Tableau**, les variables sont ordonnées par ordre de significativité. Ainsi la fréquence de sinistres apparaît comme étant la plus significative.

B. Etude de la corrélation entre variables explicatives

Nous complétons la sélection automatique des critères tarifaires en observant parallèlement les corrélations existant entre les différents variables. La corrélation entre deux critères tarifaires signifie que l'information apportée par l'une d'entre elles peut se déduire de celle apportée par l'autre. Par souci de simplicité, il est préférable de ne garder qu'un critère tarifaire parmi ceux qui sont fortement corrélés. L'étude des corrélations entre les différents

critères tarifaires se fait à l'aide du V de Cramer présenté en **annexe B**. C'est un indicateur qui permet de mesurer la corrélation entre les différents critères sous forme de tableau de contingence. Nous choisissons un seuil de 65 % pour définir les corrélations fortes dans l'étude. Différents niveaux de seuils de corrélations sont présentés ci-dessous.

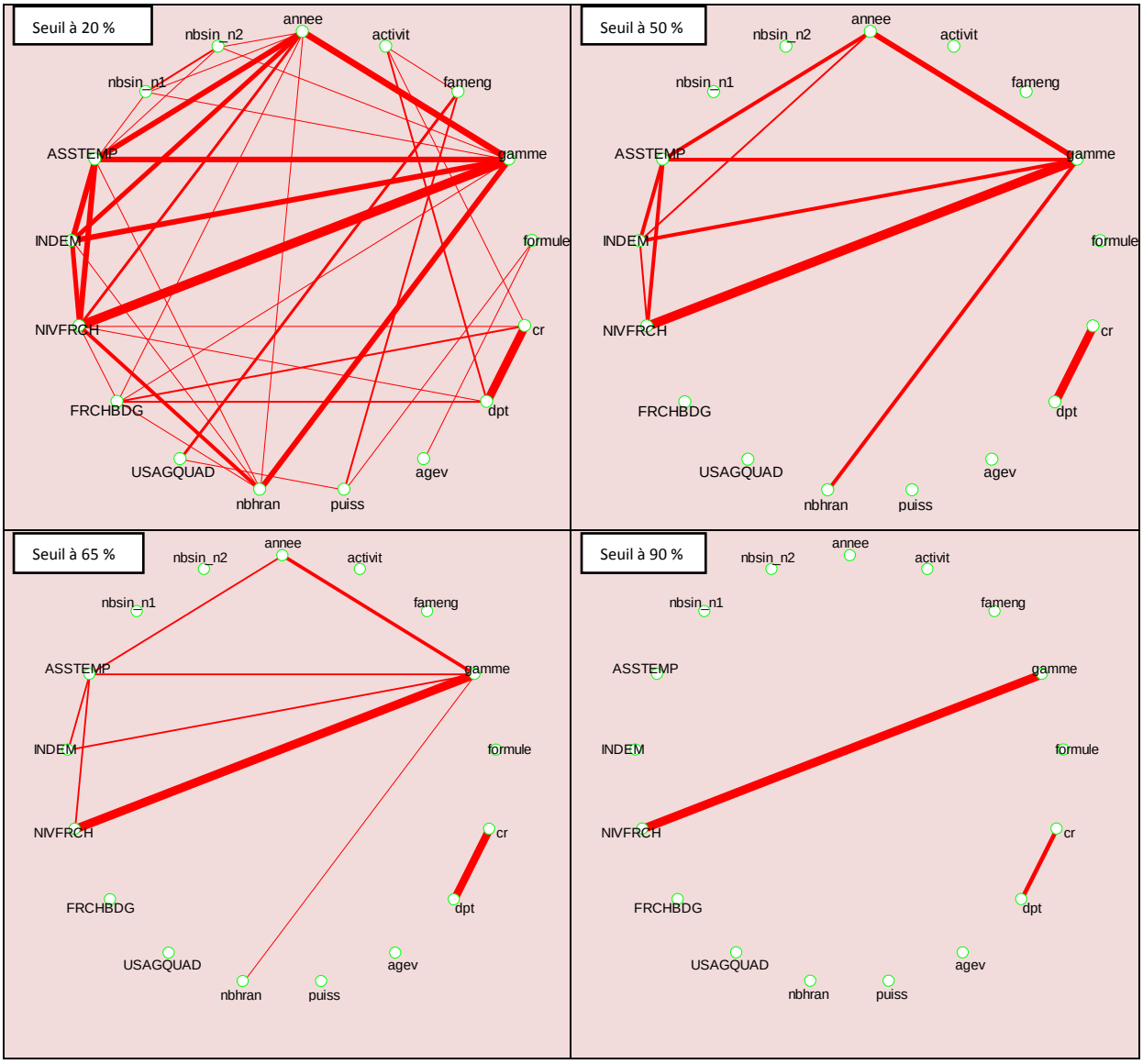


Figure 16 : Exemples de seuil de corrélations entre les différents facteurs potentiels du risque AA

Sur chaque figure les variables testées sont représentées tout au long du cercle de corrélation. Le trait rouge reliant deux variables indique que leur corrélation est au moins égal au seuil indiqué. Le trait est d'autant plus épais que la corrélation entre les deux variables est supérieure au seuil fixé.

Pour le seuil de corrélation fixé, on n'observe pas de corrélation significative entre les variables sélectionnées dans le *Stepwise*.

C. Adéquation du modèle aux observations

Nous vérifions à présent l'adéquation du modèle avec les observations en étudiant son niveau de déviance. Pour cela, nous comparons le niveau de déviance D du modèle rapporté au nombre d'observations n qui est réduit du nombre de paramètres p.

- Si $\frac{D}{n-p} \approx 1$ alors le modèle est adéquat aux données d'observation
- Sinon, le modèle n'est pas adéquat

Ainsi, le résumé du modèle établi avec la déviance minimisée est présenté dans le Tableau suivant :

Model Label	**Modèle fréquence Initial**
Error Structure	**Negative Binomial**
Link Function	**Log**
Model Description	Mean + activit + fameng + formule + dpt + agev + puiss + FRCHBDG
Offset Description	
Observations	1 438 108
Zero Weighted	0
Parameters	137
Fitted Parameters	137
Deviance	591 316,5
Scale Parameter	(Deviance) 0,4112159
Chi Squared Percentage	
AIC	1 114 751,0
BIC	1 116 419,0
Fitting Result	Converged OK

Tableau 11 : Résumé du modèle fréquence

D'après le **Tableau** ci-dessus, le modèle obtenu est convergent (*Fitting Result*). Ceci signifie que l'algorithme numérique de Newton Raphson utilisé pour l'estimation des coefficients β a été mené à terme. L'indicateur de performance du modèle, le *Scale Paramater Deviance* n'est pas tout à fait proche de 1 ce qui signifie que le modèle est encore perfectible. Cette amélioration peut être apportée par le retraitement des modalités des variables sélectionnées.

2.3.1.3 *Traitement des modalités des variables tarifaires significatives*

Certaines variables significatives peuvent présenter des modalités qui ne le sont pas. Cela s'explique par le fait que ces modalités sont peu représentées dans la base de données d'étude. En conséquence, l'estimation des coefficients β associés à ces modalités peut ne pas être très fiable.

D'autres part, la proximité des modalités en termes de comportement de sinistralité (cas de l'âge de l'automoteur) amène à recoder les modalités d'une variable. Autrement dit, nous regroupons les modalités qui présentent une similitude en termes d'impact sur la fréquence des sinistres, tout en conservant la cohérence commerciale du tarif.

- **Cas de l'activité**

Dans le graphique suivant, nous présentons une illustration de l'estimation de la fréquence des sinistres AA selon les modalités de la variable **activit**.

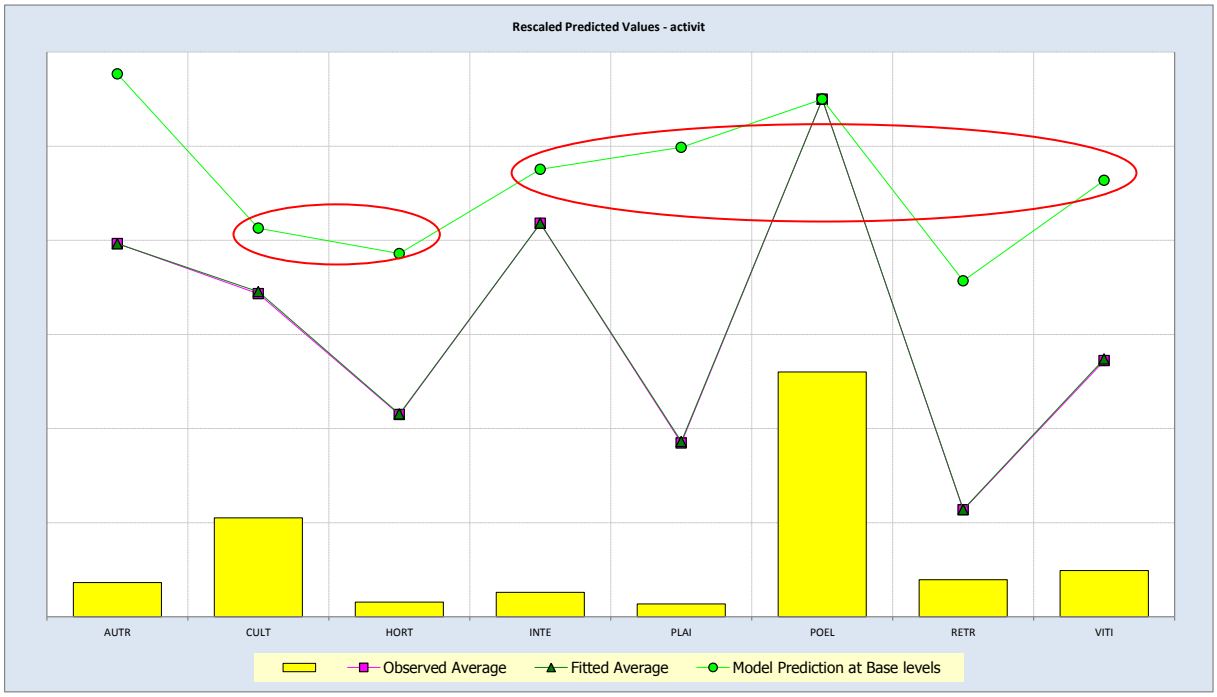


Figure 17 : La fréquence des sinistres en fonction de l'activité

L'histogramme en jaune correspond au nombre d'années assurances associé à chaque modalité. Nous rappelons que dans le modèle fréquence établi, le nombre d'années assurance définit le poids de la variable à expliquer. La courbe rose correspond à la fréquence moyenne des sinistres observée dans le portefeuille. La courbe verte foncée

correspond à la prédiction du modèle. La courbe verte claire correspond à l'effet pur de chaque modalité de l'activité, estimé par le modèle.

Pour faciliter les interprétations, l'effet pur est représenté en base 1. Pour cela, nous établissons une **modalité de référence** – celle qui a le plus de poids- puis on rapporte l'effet pur des autres modalités à celle de cette modalité de référence. Ainsi, l'effet pur de la variable d'occurrence vaut 1 et celle des autres correspond à un coefficient multiplicatif de majoration ou minoration.

On observe que les modalités **INTE**, **PLAI** et **VITI** ont des effets purs relativement proches il en est de même pour les modalités CULT et HORT. Nous décidons donc de regrouper entre elles, les modalités influençant de la même manière la fréquence des sinistres. Ce qui conduit au résultat donné par la **Figure** suivante où l'activité i ($i = 1, 2, 3$) représente un regroupement de profession des assurés.

- ***Activité 1*** *: POEL + AUTR*
- ***Activité 2*** *: CULT + HORT + RETR*
- ***Activité 3*** *: INTE + PLAI + VITI*

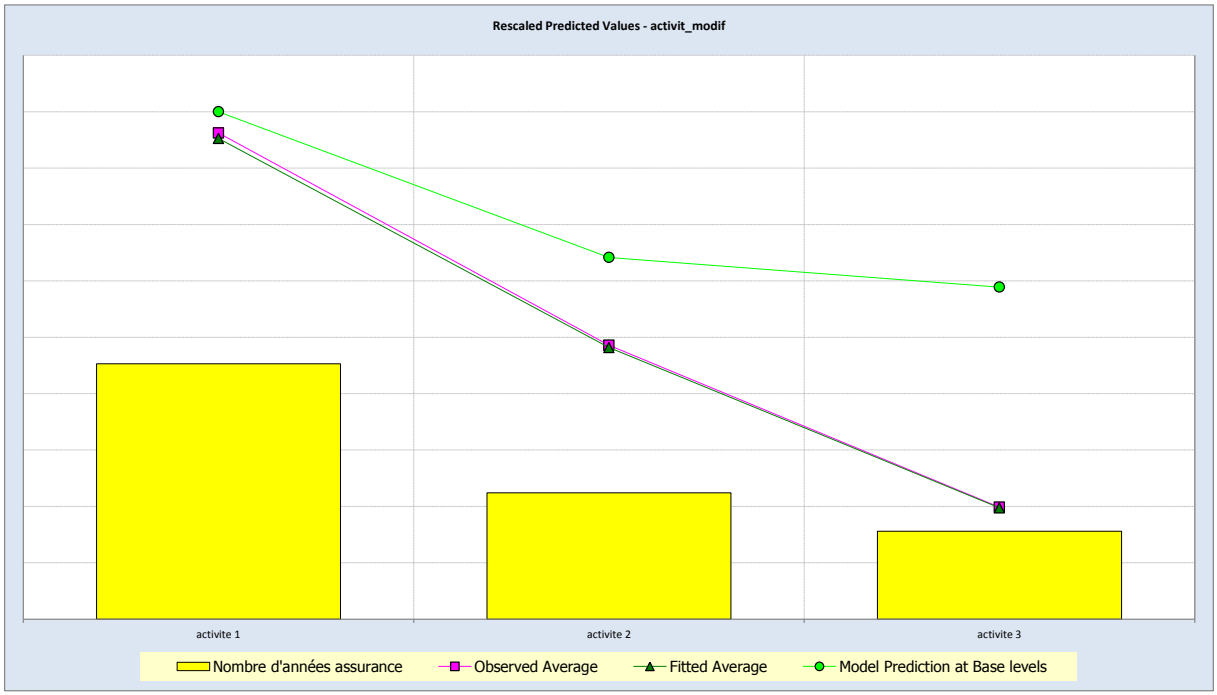


Figure 18 : La fréquence des sinistres en fonction de l'activité après regroupement

On observe que malgré le regroupement, le modèle s'ajuste toujours aussi bien à l'observé.

➢ **Cas du département**

Le traitement de la variable département est aussi très important. Bien que ce soit une variable significative pour le modèle, le très grand nombre de modalités de la variable fait qu'elle est difficilement interprétable comme on peut le voir dans la figure ci-dessous.

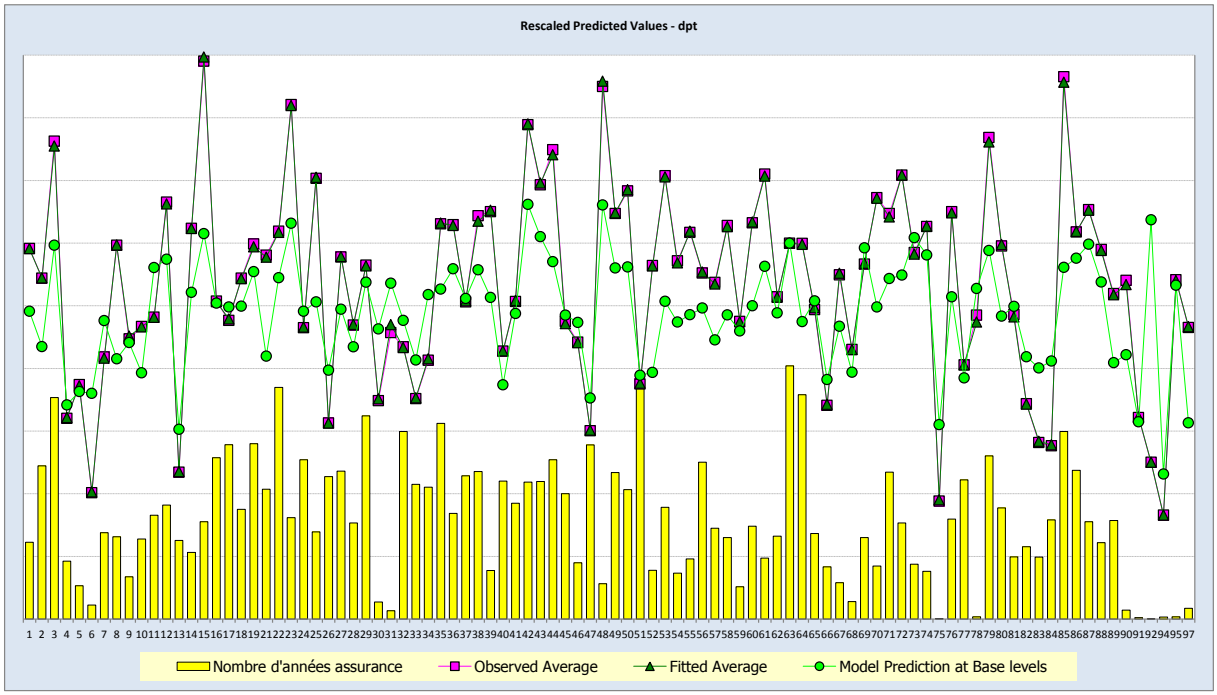


Figure 19 : La fréquence des sinistres en fonction du département

Pour traiter ce problème, nous avons effectué une **classification ascendante hiérarchique** (**CAH**[11]) des coefficients $\hat{\beta}$ associés à chaque département. Nous obtenons alors un regroupement de départements suivant quatre nouvelles modalités :

- **Zone 1**: c'est un regroupement de départements dont le coefficient $\hat{\beta} \in [0{,}33; 0{,}61]$
- **Zone 2**: c'est un regroupement de départements dont le coefficient $\hat{\beta} \in [0{,}62; 0{,}71]$
- **Zone 3**: c'est un regroupement de départements dont le coefficient $\hat{\beta} \in [0{,}73; 0{,}82]$
- **Zone 4**: c'est un regroupement de départements dont le coefficient $\hat{\beta} \in [0{,}83; 1{,}09]$

[11] *La CAH est une méthode d'analyse de données qui consiste à agréger de proche en proche des individus entre eux en utilisant un indice de* ***distance*** *(Exemple – distance euclidienne, distance du Chi2), puis des classes d'individus entre elles en utilisant un indice d'****agrégation*** *(Exemple – indice du lien minimum, indice du lien maximum), jusqu'à obtenir une classe englobant l'ensemble des individus.*

2.3.1.4 Validation du modèle

Nous jugeons enfin la qualité du modèle en effectuant une analyse des résidus. En pratique, nous nous intéressons aux **résidus de déviance standardisés**. La base «Crunched 2500 » est appropriée pour l'étude de la fréquence de sinistres. Il y a deux groupes distincts de polices : ceux qui ont des sinistres et ceux qui n'en ont pas. Si nous ne regroupons pas les polices en groupes, il y aura des «grappes» de résidus.

La **Figure** ci-dessous présente sur un plan, un nuage de points correspondant à un regroupement de résidus de déviance.

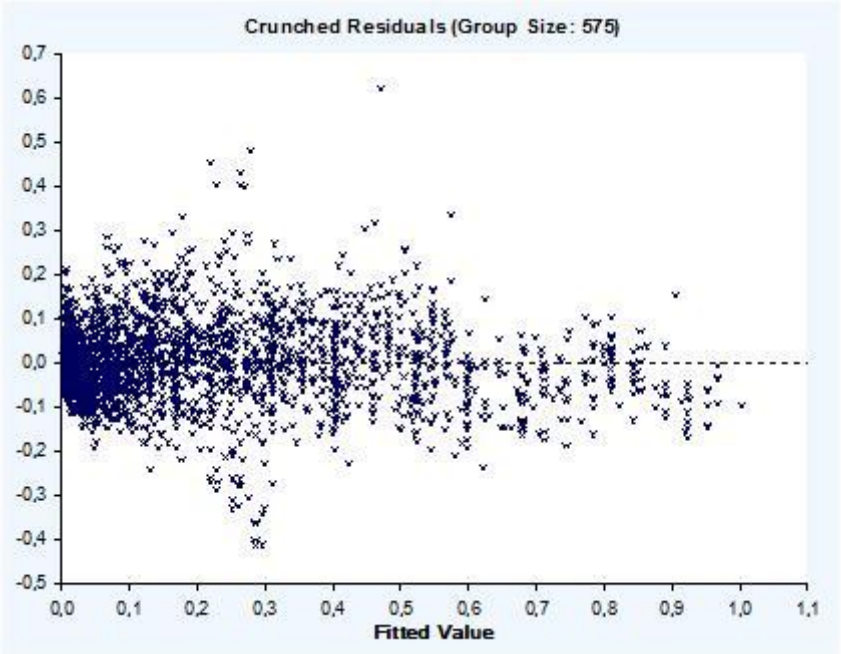


Figure 20 : Les résidus de déviance standardisés

L'axe des abscisses correspond à la fréquence des sinistres prédite sur la base des observations AA. L'axe des ordonnées correspond aux différentes valeurs des résidus de déviance standardisés selon la fréquence des sinistres prédite. L'axe des abscisses traverse de façon symétrique le nuage de points. Nous en déduisons que le modèle construit est de bonne qualité. Ce qui crédibilise l'interprétation de la fréquence des sinistres en fonction des critères tarifaires dans la section suivante. En aval, nous analysons les résidus avec une loi de Poisson. Le résultat ne présente pas une de différences importantes (voir **annexe E**).

2.3.1.5 *Résultat de la modélisation de la fréquence des sinistres*

Ci-dessous, nous présentons le résultat de la modélisation de la fréquence des sinistres après avoir effectué des retraitements sur les modalités pour chaque variable significative retenue. L'explication de chaque champ est donnée après la représentation suivante.

Name	*Value*	*Standard Error*	*Standard Error (%)*	*Exp(Value)*
Mean	-1,18725261	0,00619	0,3	0,3050582
formule (1)	-1,18521959	0,01205	0,6	0,305679
formule (2)	-0,53210319	0,00591	0,7	0,5873683
formule (3)				1
zoneAA (1)	-0,31904251	0,02403	3,1	0,7268447
zoneAA (2)	-0,35283072	0,00439	0,9	0,7026961
zoneAA (3)	-0,11703087	0,00343	1,7	0,8895577
zoneAA (4)				1
FRCHBDG (N)				1
FRCHBDG (O)	0,18515463	0,00313	0,7	1,2034045
New activit (1)				1
New activit (2)	-0,09511768	0,00379	1,2	0,9092659
New activit (3)	-0,33113926	0,00571	1,5	0,7181052
New fameng (1)	-0,67542441	0,00973	0,8	0,5089404
New fameng (2)	-0,09535952	0,00759	5,6	0,9090461
New fameng (3)	-0,29999534	0,025	3,7	0,7408217
New fameng (4)	-0,22341283	0,00732	2,3	0,7997846
New fameng (5)				1
New agev (1)	0,35690387	0,00394	0,8	1,4288985
New agev (2)	0,43484352	0,00397	0,8	1,5447213
New agev (3)				1
New puiss (1)				1
New puiss (2)	0,48032886	0,00631	0,8	1,6166059
New puiss (3)	0,50318575	0,00626	0,6	1,6539821
New puiss (4)	0,71307075	0,00939	0,9	2,0402467

Tableau 12 : Modèle Binomiale Négative pour la modélisation de la fréquence des sinistres dans le portefeuille AA

- **Descriptif de la table des paramètres β estimés**

La première colonne, à partir de la gauche, liste les variables et modalités retenues dans le modèle final. La colonne $value$ donne la valeur des paramètres $\hat{\beta}$ en base 1. Les lignes non renseignées correspondent aux modalités de référence. Cela signifie que la valeur du paramètre $\hat{\beta}$ est nulle. Le « $standard\ error\ pourcentage$» exprimé en pourcentage est un indicateur du résultat du test de Wald (test de nullité du paramètre β associé à chaque modalité). Le fait que les valeurs du champ « $standard\ error\ pourcentage$» soit inférieure à 50 % signifie que la modalité en question est statistiquement discriminante.

Nous rappelons que les coefficients multiplicateurs qui établissent l'équation tarifaire pour la fréquence des sinistres AA sont déduits de l'estimation des coefficients $\hat{\beta}$ par la relation suivante :

$$\gamma_j^F = e^{\beta_j^F}$$

Où le coefficient β_j^F est l'effet pur associé à la $j^{ième}$ modalité/occurrence, sur la fréquence des sinistres, et γ_j^F est le coefficient multiplicateur associé.

L'équation tarifaire pour la fréquence des sinistres est alors définie par :

$$\hat{\mu}^F = e^{\beta_0^F} \times e^{\beta_1^F} \times \ldots \times e^{\beta_p^F}$$

D'après cette équation tarifaire, nous disposons de p critères tarifaires. Le coefficient β_0^F est tel que le montant $e^{\beta_0^F}$ (**fréquence de base**) est la fréquence des sinistres tarifée au contrat représentatif du portefeuille. C'est-à-dire que pour chaque critère tarifaire, il est caractérisé par les modalités de référence.

Exemple :

La fréquence des sinistres tarifée à un assuré ayant les caractéristiques suivantes :

- Formule 1
- ZoneAA 1
- Franchise bris de glace : O
- RETR
- Famille d'engins : T1
- Age du véhicule : 10 ans
- Puissance de l'automoteur : 60 CV

Est de

$$\boldsymbol{F = 0,305 \times 30,5\% \times 73\% \times 120\% \times 91\% \times 100\% \times 100\% \times 100\% = 7.42\%}$$[12]

Grâce à ces différents coefficients multiplicateurs, nous définissons plusieurs segments de risque ou **cases tarifaires.** Dans chacune de ces cases tarifaires, la prime attribuée, sur la base de la fréquence, à chaque assuré est la même.

[12] Il est à rappeler que pour des raisons de confidentialité, les réelles valeurs ne sont pas communiquées.

2.3.2 Modélisation du coût moyen

2.3.2.1 Choix du modèle

Le montant d'un sinistre est une variable continue. De ce fait, une loi de probabilité continue est plus appropriée qu'une loi de probabilité discrète comme pour le nombre de sinistres. Dans la littérature scientifique, la **loi Gamma** et la **loi Log-Normale** sont fréquemment citées pour la modélisation de ce type de variable. En pratique, l'utilisation de la loi Gamma est plus fréquente. L'utilisation de la loi Log-normale conduit à remplacer le montant du sinistre par son logarithme népérien [4]. Du fait de cette transformation, l'interprétation du coût moyen des sinistres devient moins évidente.

Dans la suite du mémoire, nous modéliserons le montant d'un sinistre à l'aide de la loi Gamma $\Gamma\left(\nu, \frac{\nu}{\mu}\right)$ dont la fonction de densité est donnée par

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} exp\left\{-\frac{\nu}{\mu}y\right\}, y \in \mathbb{R}_+$$

Pour l'obtention d'un modèle multiplicatif, nous choisissons la fonction logarithme comme fonction lien, ce qui nous permet d'avoir la relation suivante :

$$\mu_i = exp\left\{\beta_0 + \sum_j^p \beta_j \times x_{ij}\right\}$$

Nous considérons donc une suite de montants de sinistre $C_1, C_2, \dots, C_n$ indépendants et de loi Gamma de moyenne,

$$\mu_i = \mathbb{E}(C_i|x_i) = n_i exp\{\beta^t x_i\}$$

Et de variance,

$$\mathbb{V}(C_i|x_i) = \frac{exp\{\beta^t x_i\}^2}{\nu}$$

Où le poids défini a priori $n_i$correspond au nombre de sinistres associés à l'observation i.

2.3.2.2 Choix des critères tarifaires significatifs

Pour la modélisation du coût moyen des sinistres, l'ensemble des variables considérées sont également issues des OAT. Elles sont les mêmes que celles utilisées pour l'étude de la fréquence des sinistres.

A. Forward selection

Comme pour la modélisation de la fréquence des sinistres, nous procédons à une sélection pas à pas des variables issues des OAT (***forward selection***). La **Figure** suivante nous présente la liste des variables explicatives retenues suite à l'exécution de la procédure.

Code variable	Définition
Puiss	*Puissance de l'automoteur*
dpt	*Département*
fameng	*Famille d'engins*
agev	*Age de l'automoteur*
FRCHBDG	*Franchise Bris de Glace*

Tableau 13 : Liste des variables retenues dans le modèle coût moyen des sinistres

Les variables sont présentées par ordre de significativité, dans le modèle coût moyen établi. Il y a moins de variables explicatives retenues en comparaison avec la modélisation de la fréquence des sinistres. On peut donc dire que dans la tarification du risque AA, la fréquence de sinistres est la principale composante de sinistralité.

B. Etude de la corrélation entre variables explicatives

L'étude des corrélations entre variables, selon la méthode du V de Cramer, est indépendante du modèle. De ce fait, on retrouve à nouveau les mêmes mesures de corrélations que lors de la modélisation de la fréquence des sinistres.

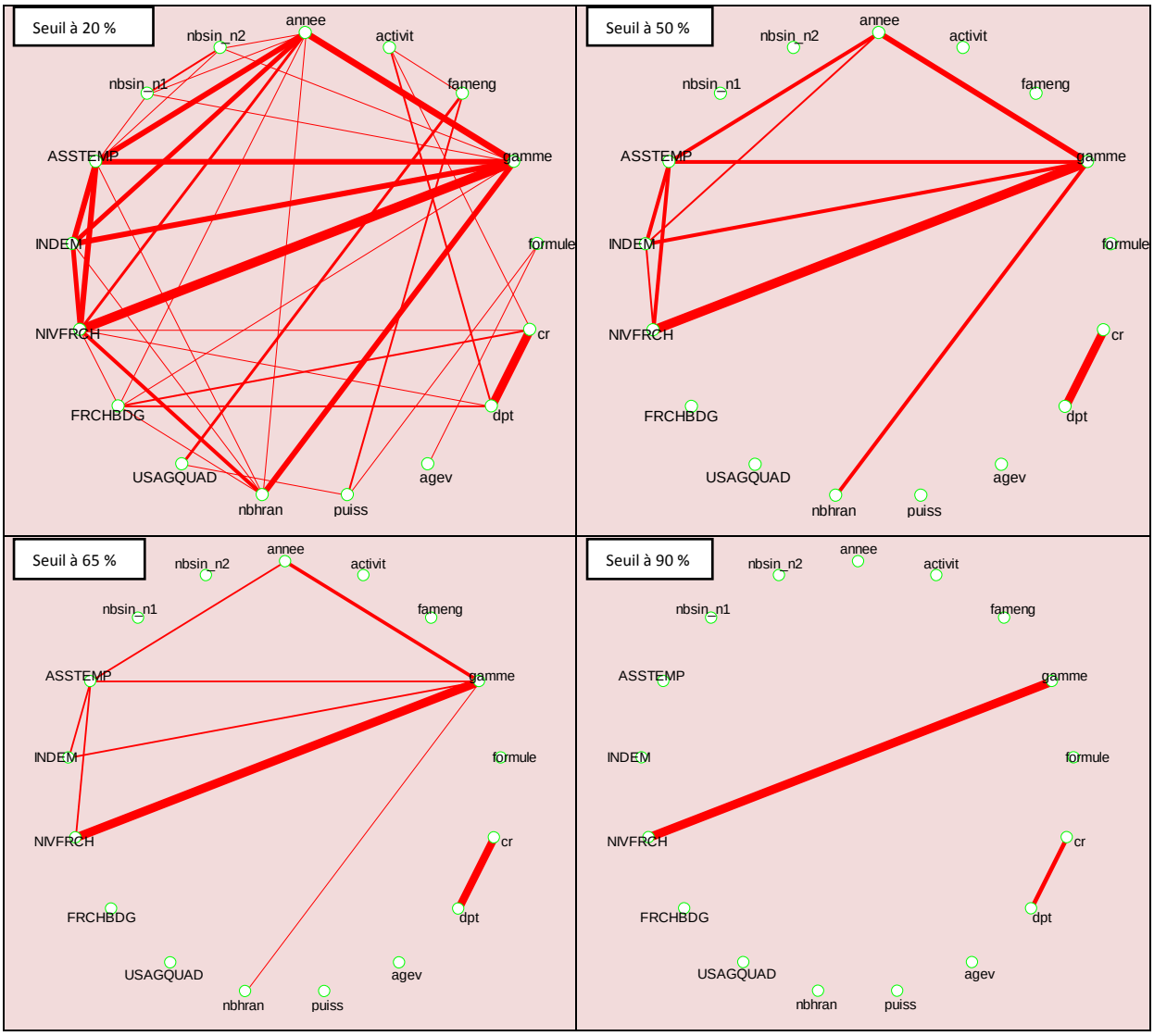


Figure 21 : Exemples de seuil de corrélations entre les différents facteurs potentiels du risque AA

Il n'y a pas de corrélation significative entre les variables retenues pour la modélisation du coût moyen des sinistres. De ce fait, lorsqu'une variable explicative est retraitée, on s'attend à ce qu'il y ait moins d'impact sur les autres variables retenues dans le modèle.

C. Adéquation du modèle aux observations

Après avoir déterminé les variables à inclure dans le modèle – en se basant sur la méthode de sélection automatique et l'étude des corrélations – nous obtenons les renseignements suivants sur le modèle :

Model Label	**Modèle 1**
Error Structure	**Gamma**
Link Function	**Log**
Model Description	Mean + fameng + dpt + agev + puiss + FRCHBDG
Offset Description	
Observations	185 689
Zero Weighted	0
Parameters	128
Fitted Parameters	128
Deviance	285 032,3
Scale Parameter	(Deviance) 1,536057
Chi Squared Percentage	Sub-Model
AIC	2 790 460,0
BIC	2 791 757,0
Fitting Result	Converged OK

Tableau 14 : Informations statistiques sur le modèle du coût moyen

Le modèle obtenu est convergent. Le niveau de déviance minimisé lors de la sélection des variables est de 285032,3. L'indicateur de performance du modèle, le *Scale Paramater Deviance* n'est pas tout à fait proche de 1 ce qui signifie que le modèle est encore perfectible. Nous effectuons à nouveau une étude des modalités de certains facteurs, pour en permettre une meilleure interprétation.

2.3.2.3 Traitement des modalités des variables tarifaires significatives

A nouveau, nous procédons à une simplification du modèle en regroupant dans certaines variables les modalités. Dans la suite nous présentons la démarche effectuée pour le département, la puissance et la famille d'engins.

Cas de la puissance de l'automoteur

La **Figure** suivante montre bien que le coût moyen associé à une police croît avec la taille de l'automoteur. Le modèle retranscrit bien ce phénomène, sauf au niveau de certaines tranches mises en évidence sur la Figure.

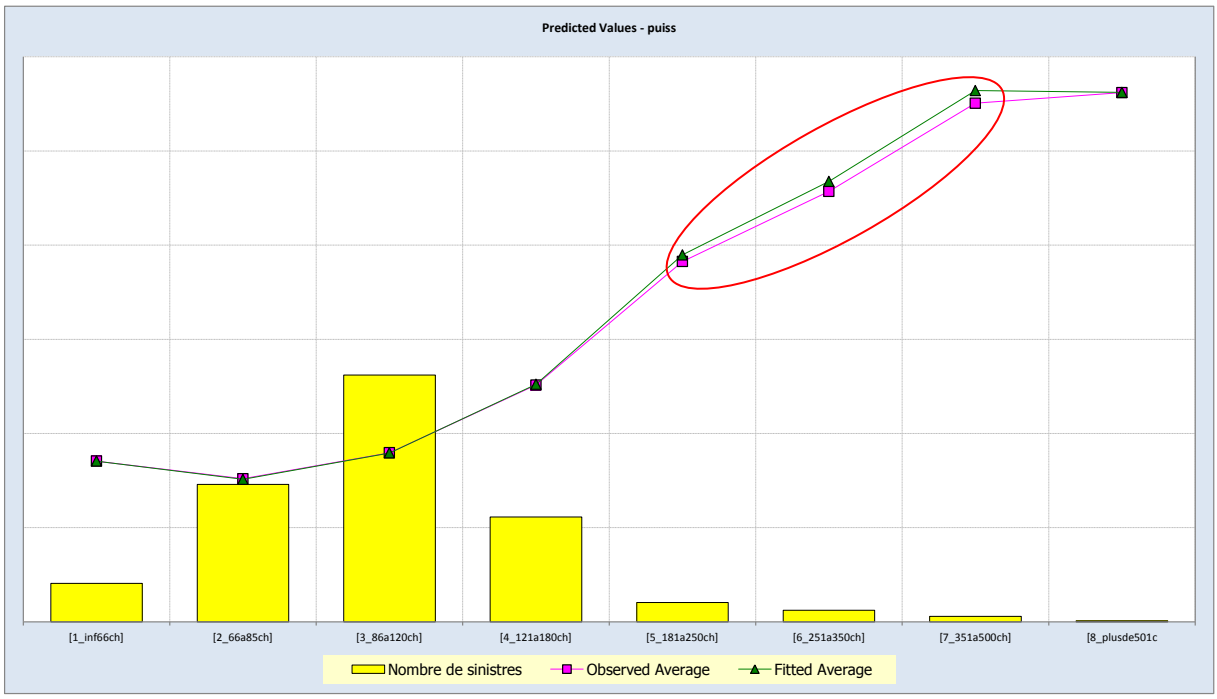


Figure 22 : Prédiction du coût moyen selon la classe de puissance

L'écart observé entre le modèle et l'observation est dû à la faiblesse du nombre de sinistres d'automoteurs de plus de 250 CV dans le portefeuille. Ce qui conduit à un léger biais d'estimation. Nous analysons au plus près ce biais d'estimation dans la **Figure** suivante.

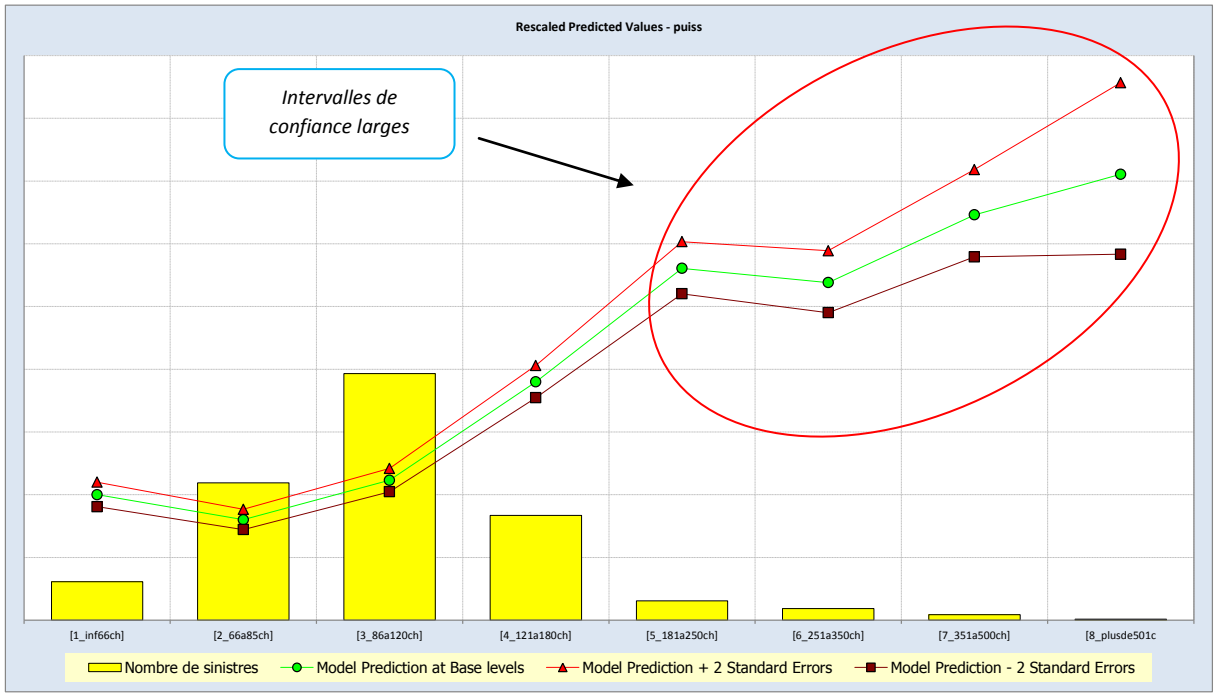


Figure 23 : Intervalle de confiance à 95% de l'estimation des β des modalités de la variable puiss

On déduit de la représentation ci-dessus, que l'estimation de l'effet pur des classes de puissance de plus de 250 CV n'est pas très fiable dans le modèle 1. Nous décidons donc de regrouper les classes de puissance de plus de 250 CV en une seule modalité. Le test de Wald effectué sur les différentes modalités de la puissance, nous conduit à regrouper entre les trois premières modalités (*standard error (%)* > 50 %). Nous présentons en annexe le résultat de l'ajustement.

Cas de la famille d'engins

La représentation de l'ajustement du modèle à l'observation en fonction de la famille d'engins présente localement des biais.

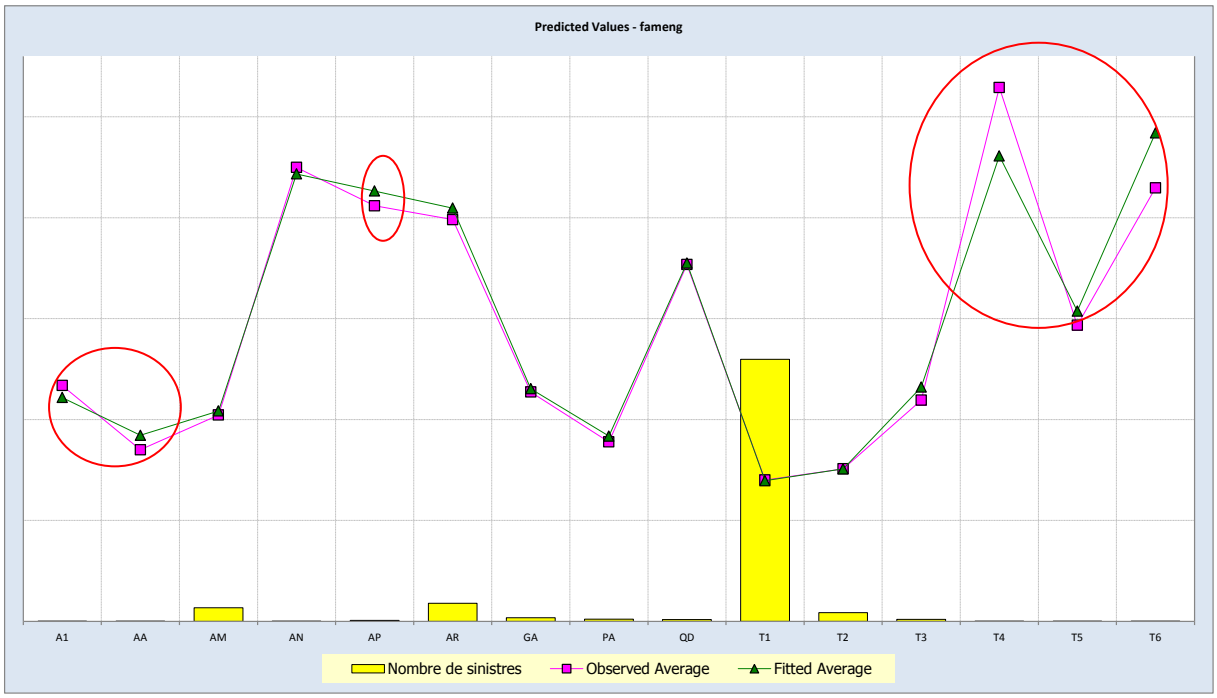


Figure 24 : Prédiction du coût moyen selon la famille d'engin

Il y a très peu de sinistres observés dans la plus part des familles d'engins répertoriées dans la base de données. Nous effectuons donc un test de Wald pour juger la pertinence de chaque modalité.

	fameng (A1)	fameng (AA)	fameng (AM	fameng (AN)	fameng (AP)	fameng (AR)	fameng (GA)	fameng (PA)	fameng (QD)	fameng (T1)	fameng (T2)	fameng (T3)	fameng (T4)	fameng (T5)	fameng (T6)
fameng (A1)															
fameng (AA)	173														
fameng (AM	155	627													
fameng (AN)	29	35	21												
fameng (AP)	37	50	12	54											
fameng (AR)	53	92	9	32	28										
fameng (GA)	72	164	19	27	20	33									
fameng (PA)	333	197	55	19	12	12	20								
fameng (QD)	649	103	28	17	11	10	16	60							
fameng (T1)	136	63	5	15	7	3	6	17	37						
fameng (T2)	632	101	15	17	9	6	10	42	12 940,0	17					
fameng (T3)	84	246	34	25	19	30	114	27	20	9	15				
fameng (T4)	42	55	35	108	233	68	50	31	27	22	26	46			
fameng (T5)	52	80	35	44	101	168	70	29	24	18	23	57	105		
fameng (T6)	36	48	23	84	175	45	33	21	18	14	17	30	2 474,0	83	
	fameng (A1)	fameng (AA)	fameng (AM	fameng (AN)	fameng (AP)	fameng (AR)	fameng (GA)	fameng (PA)	fameng (QD)	fameng (T1)	fameng (T2)	fameng (T3)	fameng (T4)	fameng (T5)	fameng (T6)

Tableau 15 : Test de Wald sur les modalités de la variable puissance de l'automoteur

Les valeurs de la mesure *standard error (%)* qui sont supérieures à 50 %, indiquent que les modalités du croisement ne sont pas statistiquement différentes. D'après ce tableau, nous décidons de ne considérer que trois types de famille d'engins expliquant le coût moyen :

- T1 et autres
- AR
- AM

Le résultat du regroupement est donné en annexe.

Nous retraitons également les modalités des autres variables en fonction de leurs caractéristiques respectives (taille de l'exposition, amplitude des coefficients β, test de Wald, expertise qualitative…).

2.3.2.4 Validation du modèle

Contrairement à la fréquence des sinistres, la modélisation du coût moyen ne porte que sur les contrats ayant eu au moins un sinistre. D'une part le nombre d'observations est ainsi moindre. D'autre part nous n'avons plus besoin de représenter les résidus standardisés de déviance par bloc comme pour la fréquence des sinistres. En pratique nous utilisons le graphique *Contour Plot of Studentized Deviance*, qui permet d'analyser de façon pertinente la structure des résidus dans le cas des variables telles que le coût moyen.

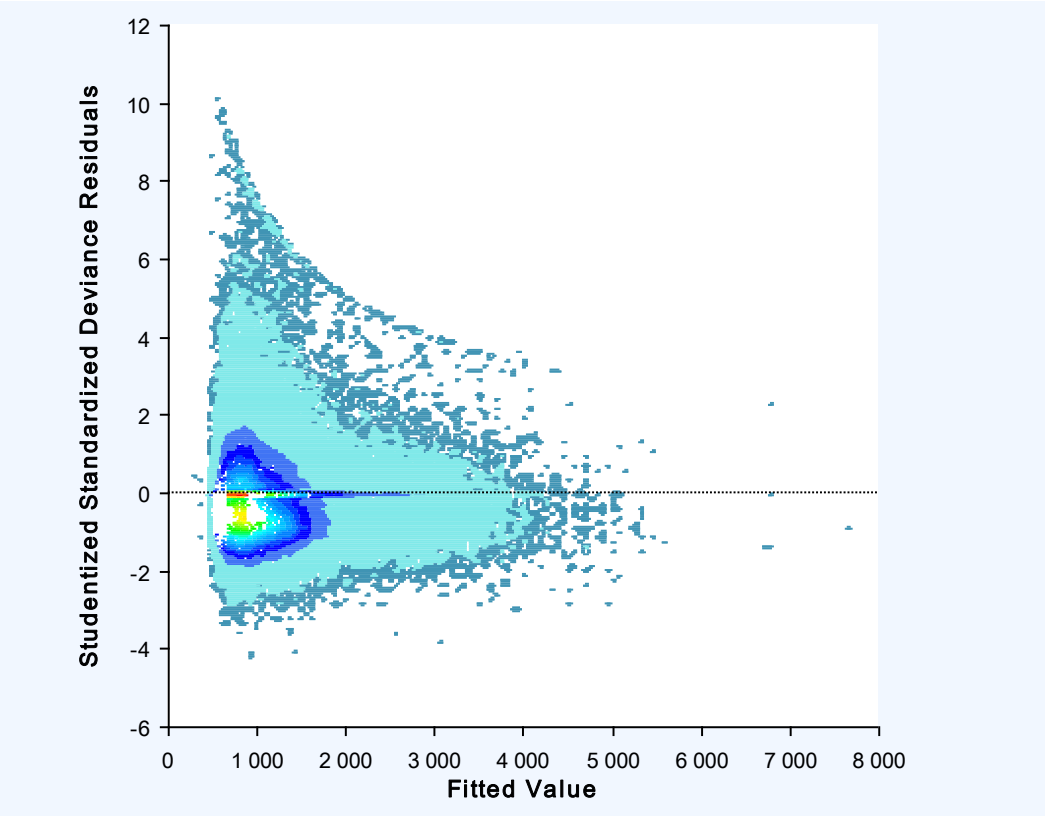


Figure 25 : Résidus de déviance standardisés pour le modèle coût moyen

Nous observons que la distribution des nuages est principalement centrée sur 0 avec néanmoins un volume de sinistres avec des résidus légèrement négatifs. L'ajustement du modèle reste néanmoins de qualité. Nous analysons en **annexe E**, plusieurs distributions usuellement appliquées pour ce type de variable d'intérêt. L'analyse montre que la fonction gamma est celle qui présente une meilleure adéquation. Elle sera donc conservée pour la suite de l'analyse.

2.3.2.5 *Interprétation des résultats de la modélisation*

Le **Tableau** suivant présente le modèle final pour le coût moyen des sinistres avec l'estimation des coefficients β associée à chaque occurrence.

Name	*Value*	*Standard Error*	*Standard Error (%)*	*Exp(Value)*
Mean	5,85748701	0,00526	0,1	349,843885
zoneAAc (1)				1
zoneAAc (2)	0,13730896	0,006	2,8	1,14718253
zoneAAc (3)	0,33425937	0,01804	4,2	1,39690541
FRCHBDG (N)				1
FRCHBDG (O)	-0,06338237	0,00522	5,1	0,93858452
New fameng (1)				1
New fameng (2)	0,10861202	0,01277	5	1,11472978
New fameng (3)	0,31166377	0,01396	3,2	1,36569543
New agev (1)	0,18571112	0,01006	3,7	1,20407437
New agev (2)	0,1493104	0,00552	3,3	1,16103332
New agev (3)				1
New puiss (1)				1
New puiss (2)	0,31088195	0,00681	2,1	1,36462811
New puiss (3)	0,39934612	0,01375	2,4	1,49084955
New puiss (4)	0,70031294	0,02619	3,5	2,01438299

Tableau 16 : Modèle Gamma pour la modélisation du coût moyen des sinistres dans le portefeuille AA

Ainsi le modèle Gamma final est composé de cinq variables explicatives qui définissent les critères tarifaires pour notre tarif sur le coût moyen

Exemple :

On considère le même contrat présenté pour la tarification de la fréquence des sinistres attendue:

- ZoneAAc (1)
- Puissance de l'automoteur : 60 CV
- Age du véhicule : 10 ans
- Famille d'engins : T1
- Franchise Bris de Glace : Oui

Le coût moyen attendu est de

$$\boldsymbol{C = 350 \times 100\% \times 100\% \times 100\% \times 94\% = 330\ €}$$

Suite à la modélisation de la fréquence des sinistres et à celle du coût moyen, nous pouvons à présent déterminer le modèle final de prime pure.

2.3.3 Modélisation de la prime pure et étude de la prédictibilité de la sinistralité antérieure

2.3.3.1 Détermination de la prime pure

Les variables explicatives/critères tarifaires de la prime pure sont les variables qui sont significatives pour la fréquence des sinistres ou le coût moyen. Les coefficients de prime pure par modalité sont obtenus en multipliant deux à deux les coefficients pour la modélisation de la fréquence et ceux pour la modélisation du coût moyen. Lorsqu'une variable n'est pas significative pour la fréquence ou le coût moyen, elle contribue dans le calcul de la prime pure pour un coefficient égal à 1.

Nous reprenons l'exemple du contrat pour lequel une fréquence attendue et un coût moyen attendu ont été calculés.

Exemple :

On rappelle les caractéristiques du contrat :

- Formule 1
- Zone 1
- Franchise bris de glace : O
- RETR
- Puissance de l'automoteur : 60 CV
- Age du véhicule : 10 ans
- Famille d'engins : T1
- Niveau de franchise : 2

On rappelle que $\boldsymbol{F}$ désigne la fréquence de sinistres modélisée et $\boldsymbol{C}$ le coût moyen des sinistres modélisé.

La prime pure annuelle est de :

$$\boldsymbol{PP = F \times C}$$

Soit,

$$\boldsymbol{PP = 330\,€ \times 7,42\% = 24,50}^{13}\,€\ \boldsymbol{par\ Année}$$

2.3.3.2 Analyse du modèle Fréquence x Coût moyen

Nous avons effectué une modélisation de la sinistralité du produit AA à l'aide d'un outil statistique rigoureux que sont les GLMs. Nous avons ainsi obtenu une évaluation quantitative des effets purs de facteurs de risque expliquant de façon pertinente la sinistralité AA. Selon les différents facteurs de risque retenus, le modèle s'ajuste bien en fréquence et également en coût moyen. La vérification de la pertinence des variables introduites dans l'équation tarifaire AA a été ainsi établie par le modèle **Fréquence x Coût moyen**.

Ayant considéré l'ensemble des principaux critères de risque du portefeuille, nous souhaitons analyser dans quelle mesure on pourrait intégrer la sinistralité individuelle comme mesure supplémentaire du risque AA. Cette sinistralité individuelle correspond au nombre de sinistres passés d'une police dans le portefeuille.

En assurance automobile, la sinistralité individuelle a permis aux compagnies d'atteindre un équilibre économique à travers le système bonus-malus. Ce dernier a contribué à la baisse du nombre de déclarations de petits sinistres dont les frais de gestion sont supérieurs à l'indemnisation.
Un système comparable a été mis en place par PACIFICA dès 2009. Il s'en est suivi une amélioration des résultats techniques :

- Baisse de la fréquence de sinistres
- Rééquilibrage de l'indice S/C[14]

[13] *Nous rappelons qu'un aléa a été appliqué à l'ensemble des variables pour respecter la confidentialité des chiffres. D'autre part, la prime pure est obtenue sans distinction des garanties.*
[14] *Le **ratio S/C** définit le rapport de la charge de sinistres sur la cotisation. C'est un indicateur de l'adéquation de la cotisation avec la charge de sinistres ou prime pure réelle. En effet, un ratio égal à 100 % signifie que la cotisation permet de régler l'ensemble des sinistres survenus.*

Nous souhaitons donc analyser à une échelle micro, l'impact de la sinistralité individuelle dans la tarification du risque AA.

2.3.3.3 Etude de la prédictibilité de la sinistralité antérieure

Nous créons une nouvelle variable qualitative qui renseigne le nombre de sinistres à chaque exercice **nb_sin**. Cette variable diffère de la variable **nbsin** dans la mesure où elle est qualitative. Les résultats techniques des différents profils sont ensuite ventilés suivant les modalités prises par la variable **nb_sin**, à chaque exercice.

Nous étudions ainsi la sinistralité du portefeuille lors de **l'exercice N2** en fonction du nombre de sinistres à un **exercice N1** passé ($N1 < N2$) comme l'illustre le **Tableau** ci-dessous.

	résultats N2 (Hors AN N2)		
	fréquence	**S/C sous crête**	
nb sin N1	indice freq	indice prime pure	indice S/C
0	73	74	82
1	232	222	142
2 ou +	425	447	205
total	100	100	100

Tableau 17 : Les résultats techniques d'une cohorte de profils de risque pour l'exercice N2[15]

Le **Tableau** présente les résultats techniques de l'exercice N2 associés à chaque profil de risque vu lors de l'exercice N1. Ces résultats techniques sont représentés sous forme d'indice pour faciliter la compréhension.

- Le champ « indice freq » correspond à la fréquence de chaque profil (défini par son nombre de sinistres en N1) rapportée la fréquence totale de la génération des profils en N2.
- Le champ « indice prime pure » correspond à la charge de sinistres sou-crête de chaque profil rapportée à la charge de sinistres sous-crêtes totale de la génération des profils en N2.
- Le champ « indice S/C » correspond au S/C associé à chaque profil rapporté au S/C global de la génération de profils en N2.

[15] AN signifie Affaire Nouvelle

Par exemple, la fréquence d'un contrat ayant eu au moins 2 sinistres lors de l'exercice N1, est 4 fois supérieure $(Indice\ freq = 425)$ en N2, à la fréquence de sinistres de la génération des contrats N1.

En prenant quatre années comme fenêtre d'observation (de **2008 à 2012**), nous répétons la procédure autant de fois que l'on peut observer une génération de profils de risque dans un prochain exercice. Par exemple, si $N1 = 2009$ alors $N2 = 2010, 2011\ ou\ 2012$, la procédure est donc répétée 3 fois. Cela nous permet de déterminer des coefficients de prédictibilité de la sinistralité antérieure sur la sinistralité future. Nous obtenons les résultats présentés ci-dessous.

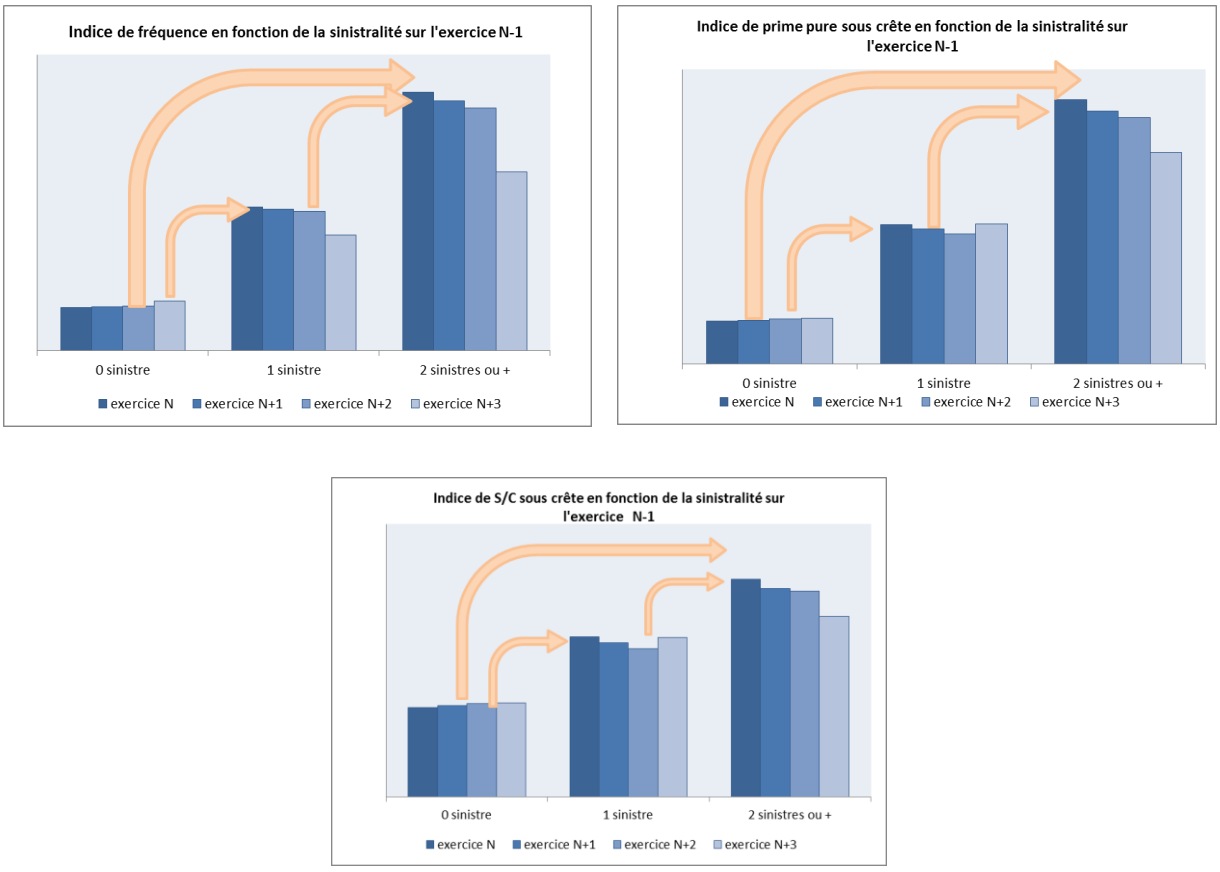


Figure 26 : L'impact de la sinistralité passée sur les résultats techniques

En considérant la fréquence de sinistres (**graphique en haut à gauche**), l'étude nous dit qu'une police qui a eu au moins deux sinistres lors d'un exercice à une probabilité d'être sinistrée chaque année supérieure à une police qui n'en a eu qu'un seul. Cette prédisposition à la sinistralité augmente par rapport à une police qui n'a pas eu de sinistre lors du même

exercice. Le phénomène est retranscrit sur la charge de sinistres sous-crêtes. Il est toutefois moins amplifié sur l'indice S/C[16], dû au fait que la cotisation prend en compte des facteurs explicatifs du risque AA, en particulier ceux liés aux caractéristiques de l'automoteur. Pour autant, les amplitudes observées, sur la représentation de l'indice S/C en fonction de la variable **nb_sin,** sont significatives. Cette étude semble ainsi nous montrer que la variable **nb_sin** constitue un facteur de risque pertinent pour l'étude de la sinistralité AA. Nous allons donc l'analyser sous l'angle des GLMs.

[16] *Le ratio S/C définit le rapport de la charge de sinistres sur la cotisation. C'est un indicateur de l'adéquation de la cotisation avec la charge de sinistres ou prime pure réelle. En effet, un ratio égal à 100 % signifie que la cotisation permet de régler l'ensemble des sinistres survenus.*

2.3.3.4 *Evaluation de l'impact du nombre de sinistres passés sur la fréquence par la méthode GLM*

A. Les coefficients de la sinistralité passée

Dans la base d'étude pour les modèles GLM, on introduit les variables **nbsin_n1** et **nbsin_n2**. Nous les nommerons par la suite **coefficients de sinistralité passée**. Si l'on considère un exercice N, la variable **nbsin_n1** donne le nombre de sinistres déclaré par un contrat lors de l'exercice précédent ($N-1$). Tandis que la variable **nbsin_n2** donne le nombre de sinistres déclaré par ce contrat lors de l'avant dernier exercice ($N-2$). Elles ont ainsi toutes deux les mêmes modalités vues à des exercices différents.

- *NR : Non renseigné*[17]
- *0 : Pas de sinistre*
- *1 : 1 sinistre*
- *2 : 2 sinistres*
- *3 :3 sinistres*
- *3+ : Plus de trois sinistres*

B. Corrélations

Nous représentons à nouveaux les cercles de corrélations entre variables. Les représentations suivantes nous permettent de mesurer le niveau de corrélation entre **nbsin_n1, nbsin_n2 et les autres variables.**

[17] *Cette modalité correspond d'une part, aux exercices 2004 et 2005 pour lesquelles nous ne disposons pas d'informations sur les exercices antérieurs, dans la base d'études. Elle correspond d'autre part, aux affaires nouvelles de 2012.*

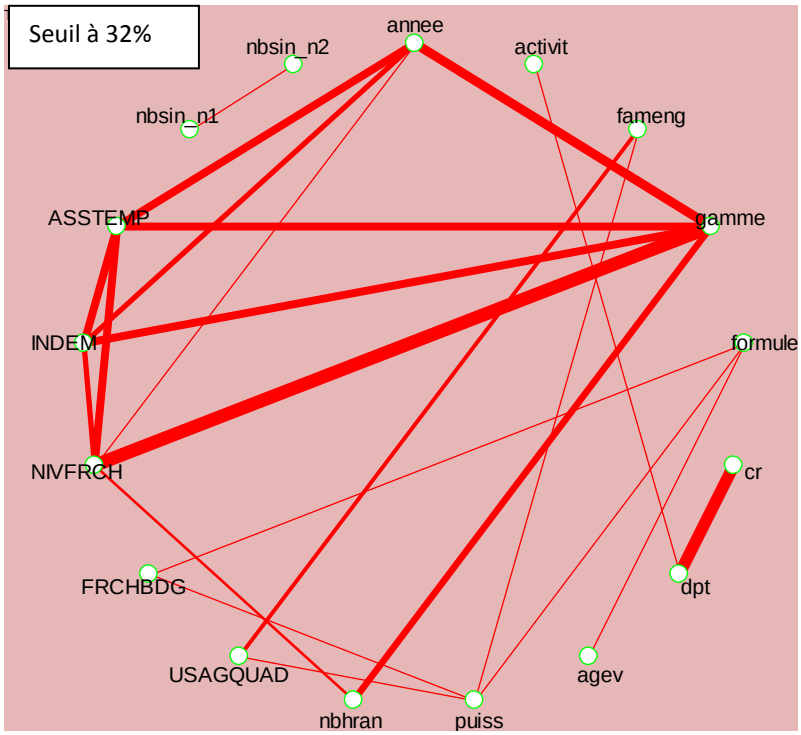


Figure 27 : Mesure de corrélation des variables nbsin_n1 et nbsin_n2

On observe que la corrélation entre les variables nbsin_n1 et nbsin_n2 n'est pas significative – seuil de pertinence fixé à$65\ \% > 32\ \%$. D'après l'analyse des corrélations, on en déduit que les variables **nbsin_n1** et **nbsin_n2** sont très peu corrélées entre elles et avec les autres variables de l'OAT.

C. Etude des coefficients de sinistralité passée dans le modèle fréquence

Nous analysons à présent l'effet pur des coefficients de sinistralité passée sur la fréquence de sinistres. Nous présentons les résultats de l'analyse de la variable **nbsin_n1**[18]. Les tests statistiques montrent que la variable **nbsin_n1** est pertinente pour expliquer la fréquence de sinistres du produit AA. Dans les Figures suivantes, nous analysons l'impact de la prise en compte de la variable **nbsin_n1** dans le modèle fréquence.

[18] *L'analyse portant sur la variable nbsin_n2 a conduit au même résultat que pour celui de la variable nbsin_n1.*

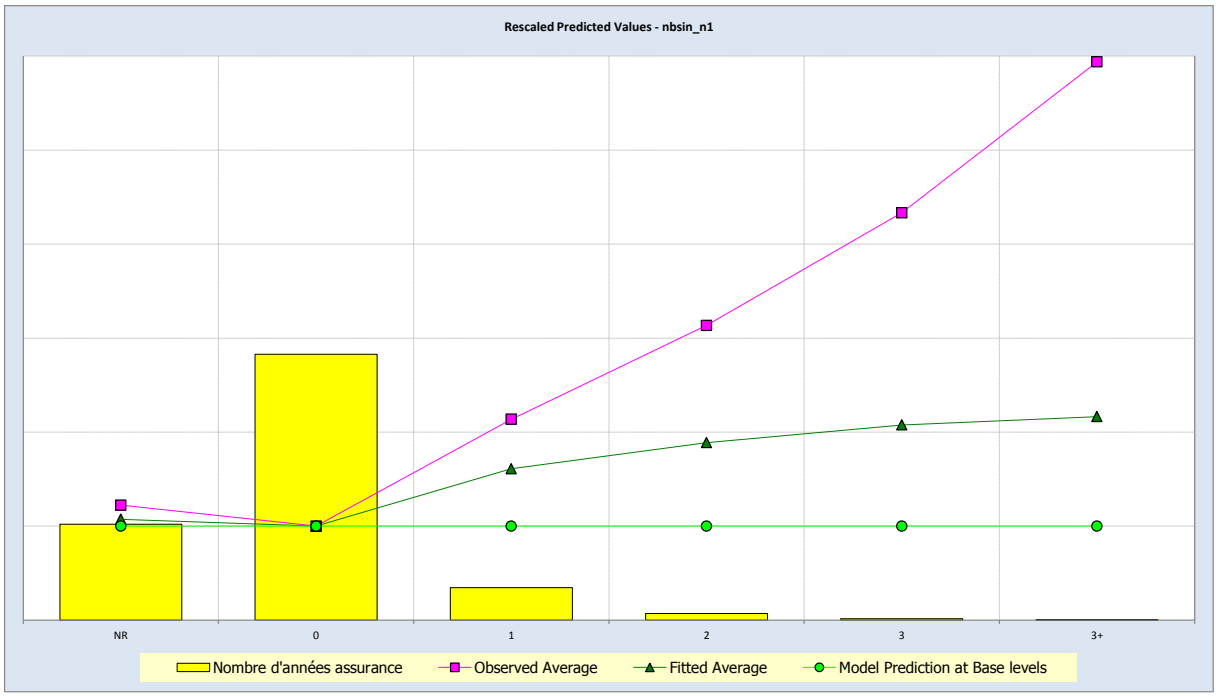


Figure 28 : La modélisation de la fréquence de sans prise en compte de la variable nbsin_n1

Le modèle ne s'ajuste pas du tout à l'observé sur l'axe de la variable **nbsin_n1**. Ainsi, une part de la fréquence de sinistres reste encore à être expliquée.

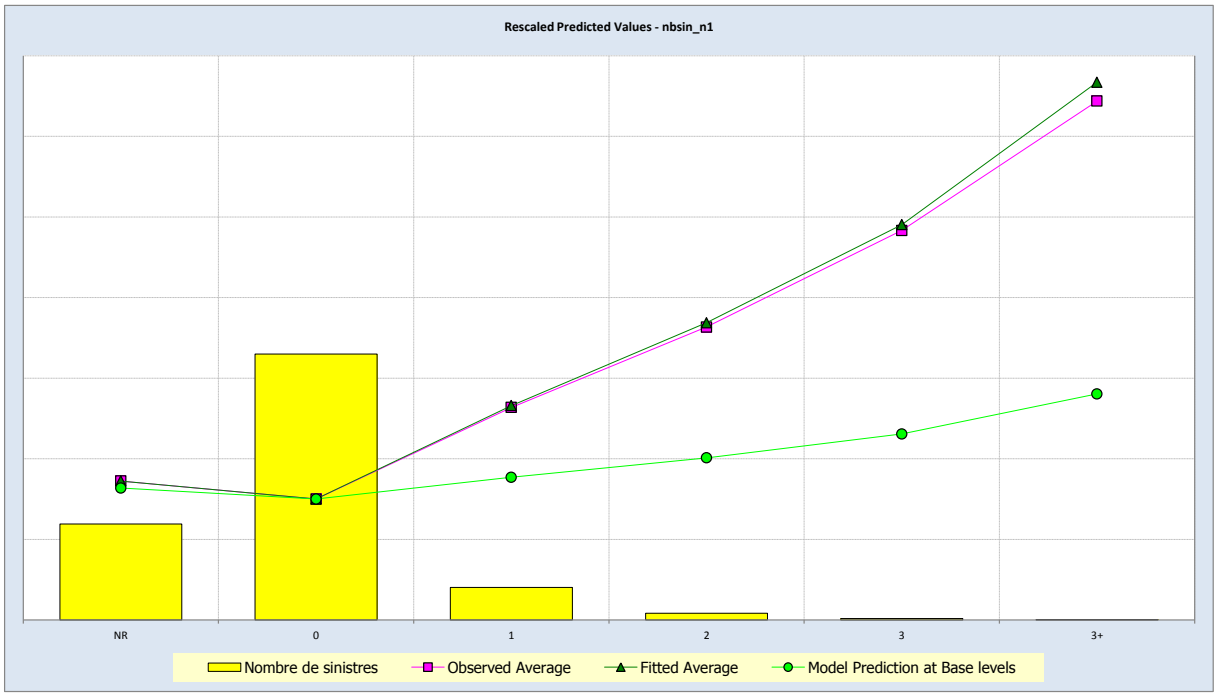


Figure 29 : La modélisation de la fréquence de avec prise en compte de la variable nbsin_n1

On observe que le modèle s'ajuste très bien à l'observé lorsqu'on intègre la variable **nbsin_n1**. Même pour des modalités de faible exposition (*2, 3 et 3+*), l'ajustement est de bonne qualité. On observe un effet pur (courbe en verte claire) qui reproduit ce qui avait été observé à plat. **Ce résultat nous permet de dire que la prise en compte de la sinistralité individuelle dans l'évaluation de la fréquence de sinistres attendue d'un contrat AA est significative d'après le modèle établi.**

Dans le chapitre 3 nous nous proposons d'étudier de quelle façon la théorie de la crédibilité pourrait nous permettre de prendre en compte cette sinistralité individuelle passée. C'est en effet une approche classiquement employée dans le cadre de la tarification a posteriori

Chapitre 3. Prise en compte de la sinistralité passée dans la tarification de la fréquence de sinistres AA

En assurance automobile traditionnelle, les mécanismes de bonus-malus ont été institutionnalisés en 1976, comme techniques de tarification a posteriori. Ces mécanismes sont basés sur les processus de chaînes de Markov. Autrement dit, l'ajustement a posteriori effectué en année N n'est fonction que de l'information sur la sinistralité du contrat en année N-1 via le coefficient de réduction/ majoration.

La théorie de la crédibilité propose quant à elle, un réajustement à posteriori de la variable d'intérêt (prime pure, fréquence de sinistres, etc.) se basant sur l'ensemble de la vie du contrat (et non juste la dernière année). Par conséquent, la théorie de la crédibilité peut-être vue comme une généralisation du système bonus/malus. Pour de plus amples détails sur le bonus-malus, nous renvoyons les lecteurs à l'article d'A. CHARPENTIER [2].

Le service actuariat de PACIFICA souhaitant optimiser le processus existant de majoration de la prime pure du produit AA (dans sa tarification actuelle) en fonction de l'historique de sinistres des contrats, ce chapitre se propose d'appliquer la théorie générale de la crédibilité. Dans la suite, nous introduisons d'abord les principes théoriques de la crédibilité. Ces principes seront après mis en application dans le but d'estimer des coefficients d'ajustement à la sinistralité individuelle.

3.1 Principes théoriques de la crédibilité

3.1.1 Histoire et principe fondamental

La théorie de la crédibilité a été introduite en Amérique du nord pour répondre notamment à des problématiques de tarification par expérience des contrats d'assurance de travail. Les premiers travaux sur le sujet datent donc du début du 20ème siècle par Mowbray [11] et Whitney [12]. En Europe, la théorie de la crédibilité a été développée par Hans Bühlmann sous l'appellation de la crédibilité de précision, voir [13] et [14].

Quel que soit l'approche utilisée, la théorie de la crédibilité est régie par le principe suivant :

Dans l'estimation d'une variable d'intérêt à partir d'un historique de données, la théorie de la crédibilité cherche à déterminer quel poids on doit accorder à l'expérience individuelle et quel poids on doit accorder à l'expérience collective.

3.1.2 Exemple introductif

Pour mettre en évidence une limite de la tarification a priori, nous présentons un exemple issu d'un rapport de R. Norberg publié dans le journal actuariel scandinave [15].

Soit un portefeuille composé de 10 contrats. Les contrats sont considérés a priori équivalents. Les conditions suivantes prévalent:

- *Tout contrat ne peut avoir au plus qu'un sinistre chaque année*
- *Le montant de sinistre est 1*
- *La prime collective est 0,2*

Le Tableau suivant liste les sinistres passés pour les 10 contrats.

Année	$i=1$	2	3	4	5	6	7	8	9	10
1									1	
2	1	1	1						1	
3	1								1	
4			1						1	
5									1	
6		1								
7	1	1		1	1					
8	1			1		1			1	
9	1				1					
10	1								1	
$\bar{S}_i$	**0,6**	**0,3**	**0,2**	**0,2**	**0,2**	**0,1**	**0**	**0**	**0,7**	**0**

Figure 30 : Historique de sinistres sur 10 ans

Les contrats 1 et 9 font énormément payer la mutualité de leur comportement risqué ($0{,}7 > 0{,}6 \gg 0{,}2$). *Les contrats 7, 8 et 10 auront tendance à aller chez l'assureur concurrent qui proposerait une cotisation plus à leur avantage. A long-terme, l'assureur pourrait présenter des résultats techniques défavorables, s'il ne prenait pas des mesures d'ajustement du tarif.*

Cet exemple est une illustration d'un portefeuille supposé homogène a priori et qui ne l'est pas au final. La sinistralité antérieure de chaque contrat permet de mettre en évidence le phénomène d'hétérogénéité résiduelle du portefeuille. La théorie de la crédibilité propose, nous le rappelons, de tenir compte à la fois de **l'expérience individuelle** d'un contrat et de **l'expérience collective** auquel il appartient, pour tarifer ce contrat.

3.1.3 Estimateur de crédibilité

On considère à nouveau un portefeuille composé de K contrats et d'un historique de T années d'observations $\left(X_{k,t}\right)_{\substack{k=1,\dots,K \\ t=1,\dots,T}}$. En fonction, la variable aléatoire $X_{k,t}$ peut correspondre à une fréquence, soit à un coût moyen ou soit à un indice S/C. Pour chacun des contrats k du portefeuille, nous cherchons à estimer la valeur $X_{k,T+1}$ qui est inconnue à la fin de la $T^{ème}$ année.

Formellement, la théorie de la crédibilité cherche à résoudre le problème d'optimisation suivant

$$\min_{\boldsymbol{f_k}(\boldsymbol{X})} \mathbb{E}\left(\boldsymbol{X_{k,T+1}} - \boldsymbol{f_k}(\boldsymbol{X})\right)^2$$

Sous la contrainte que f_k est une fonction linéaire en $X = \left(X_{k,t}\right)_{\substack{k=1,\dots,K \\ t=1,\dots,T}}$.

Par la suite, nous présentons quelques-unes des approches classiques, qui ont fait l'objet d'une étude pratique dans le cadre de ce mémoire.

3.1.4 Les modèles de crédibilité

3.1.4.1 Le modèle de Bühlmann - Straub

Ce modèle est une généralisation du modèle simple de Bühlmann. En effet, à chaque observation $X_{k,t}$, on associe un poids défini a priori $w_{k,t}$. On note Θ_k le paramètre qui définit le profil de risque du contrat k. Ce paramètre est inconnu par nature et propre à chaque individu k.

Avant de présenter l'estimateur de $X_{k,T+1}$ d'après le modèle de Bühlmann-Straub, nous introduisons quelques notations qui précèderont les hypothèses de l'approche.

Notations

On a,

- $w_{k\cdot} = \sum_{t=1}^{T} w_{k,t}$
- $\bar{X}_{kw} = \frac{\sum_{t=1}^{T} X_{k,t} w_{k,t}}{w_{k\cdot}}$

Hypothèses du modèle

(BS1) Sachant Θ_k, les variables $X_k = \left(X_{k,t}\right)_{t=1,\dots,T}$ sont indépendantes avec

$$\mathbb{E}\left[X_{k,t}|\Theta_k\right] = \mu(\Theta_k) \text{ et } \mathbb{V}\left[X_{k,t}|\Theta_k\right] = \frac{\sigma^2(\Theta_k)}{w_{k,t}}.$$

(BS2) Les pairs $(X_1, \Theta_1), (X_2, \Theta_2), \dots$ sont indépendantes et les variables aléatoires $\Theta_1, \Theta_2, \dots$ sont indépendantes et identiquement distribuées (iid).

Estimateur de crédibilité dans le modèle de Bühlmann - Straub

Sous les hypothèses (BS1) et (BS2), l'estimateur de crédibilité a la forme suivante

$$f^{BS}{}_k(X) = \alpha_k \bar{X}_{kw} + (1 - \alpha_k)\mu_0\,,$$

Où $\alpha_k \in [0,1]$ définit le **facteur de crédibilité** et est déterminé comme suit

$$\alpha_k = \frac{w_{k\cdot}}{w_{k\cdot} + \frac{\sigma^2}{\tau^2}}.$$

Les paramètres de structure ont l'interprétation suivante :

- $\sigma^2 = \mathbb{E}[\sigma^2(\Theta_k)]$ représente la **variance intra-risque** des K contrats
- $\tau^2 = \mathbb{V}[\mu(\Theta_k)]$ représente la **variance inter-risque** du portefeuille
- Le paramètre μ_0 est l'espérance des variables $\left(X_{k,t}\right)_{t=1,\dots,T}$, c'est-à-dire :

$$\mu_0 = \mathbb{E}[\mu(\Theta_k)] = \mathbb{E}\left(X_{k,t}\right).$$

En pratique, les paramètres μ_0, σ^2, τ^2 sont remplacés par leur version empirique. Nous renvoyons les lecteurs au livre de BÜHLMANN et GISLER [16] pour de plus amples détails.

3.1.4.2 Le modèle de crédibilité hiérarchique

En 1975, William Jewell [17] introduit le concept de hiérarchie de l'information dans le modèle de Bühlmann - Straub. Ainsi, les K contrats du portefeuille sont répartis suivant différentes mailles de risques. Ces mailles sont définies par des variables de segmentation qui caractérisent le risque sous-jacent au contrat k tel que nous l'avons défini dans le chapitre 2.

Dans la suite, nous allons présenter un modèle de crédibilité hiérarchique à deux niveaux de segmentation, dans lequel :

- Le premier niveau de segmentation est représenté par la variable aléatoire continue Ψ_g avec $g = 1, \dots, G$. Il définit le nœud principal de la hiérarchie.
- Le second niveau de segmentation est représenté par la variable aléatoire continue Φ_h avec $h = 1, \dots, H$.

Dans ce modèle, le profil de risque du contrat k appartenant à la maille de risques (g, h) est représenté par le paramètre Θ_{khg}.

A partir d'un modèle de crédibilité hiérarchique à deux niveaux de segmentation, on peut parfaitement déduire la structure d'un modèle de crédibilité hiérarchique avec plus de niveaux de segmentation.

Comme précédemment, nous introduisons quelques notations, suivies des hypothèses de l'approche, avant d'en déduire l'estimateur de $X_{k,T+1}$.

Notations

On a,

- $\bar{X}_{khw} = \frac{\sum_{t=1}^{T} X_{kht} w_{kht}}{\sum_{t=1}^{T} w_{kht}}$.
- $\bar{X}_{hw} = \frac{\sum_{k=1}^{K_h} \bar{X}_{khw} \alpha_{kh}^{(1)}}{\sum_{k=1}^{K_h} \alpha_{kh}^{(1)}}$.
- $\bar{X}_{gw} = \frac{\sum_{h}^{H_g} \bar{X}_{hw} \alpha_{hg}^{(2)}}{\sum_{h}^{H_g} \alpha_{hg}^{(2)}}$.

Les paramètres $\alpha_{kh}^{(1)}$ et $\alpha_{hg}^{(2)}$ sont les facteurs de crédibilité respectivement associés aux profils de risque Θ_{khg} et Φ_{hg}. Ils seront définis plus en détail dans le troisième paragraphe de cette section.

- K_h le nombre de contrats ayant la caractéristique Φ_h.
- H_g est le nombre de variables Φ_h associé au nœud g.

Hypothèses du modèle

(H1) Les variables aléatoires $(\Psi_g)_{g=1,\dots,G}$ sont indépendantes et identiquement distribuées (i.i.d), de densité $f_\Psi(.)$.

(H2) Sachant Ψ_g, les variables aléatoires $(\Phi_{hg})_{h=1,\dots,H}$ sont i.i.d et de densité conditionnelle $f_\Phi(.|\Psi_g)$.

(H3) Sachant Φ_h, les variables aléatoires $(\Theta_{kh})_{k=1,\dots,K}$ sont i.i.d et de densité conditionnelle $f_\Theta(.|\Phi_h)$.

(H4) Sachant Θ_k, les variables $X_k = (X_{k,t})_{t=1,\dots,T}$ sont indépendantes avec

$$\mathbb{E}[X_{k,t}|\Theta_k] = \mu(\Theta_k) \text{ et } \mathbb{V}[X_{k,t}|\Theta_k] = \frac{\sigma^2(\Theta_k)}{w_{k,t}}.$$

Estimateur de crédibilité dans le modèle de crédibilité hiérarchique à 2 niveaux

Sous les hypothèses (H1), (H2), (H3) et (H4), l'estimateur de crédibilité de $X_{k,T+1}$ est donné en plusieurs étapes par

$$f^H{}_{khg}(X) = \alpha_k^{(1)} \bar{X}_{khw} + \left(1 - \alpha_k^{(1)}\right) f^H{}_{hg}(X), \text{ où } \alpha_k^{(1)} = \frac{w_{k\cdot}}{w_{k\cdot} + \frac{\sigma^2}{\tau_1^2}}, \alpha^{(1)}{}_k \in [0,1].$$

$f^H{}_{hg}(X)$ est donné par

$$f^H{}_{hg}(X) = \alpha_h^{(2)} \bar{X}_{hw} + \left(1 - \alpha_h^{(2)}\right) f^H{}_g(X), \text{ où } \alpha_h^{(2)} = \frac{\sum_k^{K_h} \alpha_k^{(1)}}{\sum_k^{K_h} \alpha_k^{(1)} + \frac{\tau_1^2}{\tau_2^2}}, \alpha_h^{(2)} \in [0,1].$$

$f^{H}{}_{g}(X)$ est donné par

$$f^{H}{}_{g}(X) = \alpha_g^{(3)}\bar{X}_{gw} + \left(1 - \alpha_g^{(3)}\right)\mu_0, \text{ où } \alpha_g^{(3)} = \frac{\sum_h^{Hg} \alpha_h^{(2)}}{\sum_h^{Hg} \alpha_h^{(2)} + \frac{\tau_2^2}{\tau_3^2}}, \alpha_g^{(3)} \in [0,1]$$

Les paramètres de structure ont l'interprétation suivante :

- Avec $\tau_1^2 = \mathbb{E}\left[\mathbb{V}[\mu(\Theta_k)|\ \Phi_h]\right]$, c'est la **variance inter-risque** du second niveau de segmentation.
- $\tau_2^2 = \mathbb{E}\left[\mathbb{V}\left[\mu(\ \Phi_h)\right|\ \Psi_g\right]\Big]$, c'est la **variance inter-risque** du premier niveau de segmentation.
- $\tau_3^2 = \mathbb{V}\left[\mu(\ \Psi_g)\right]$, c'est la **variance inter-risque** du portefeuille.

Comme pour le modèle de Bühlmann - Straub, les paramètres $\mu_0, \sigma^2, \tau_1^2, \tau_2^2, \tau_3^2$ sont remplacés par leur équivalent empirique. Nous renvoyons les lecteurs au livre de BÜHLMANN et GISLER [16] pour plus amples détails sur les modèles de crédibilité hiérarchique. Notons qu'il existe des modèles plus avancés permettant de prendre en compte des variables de régression tels que le modèle de Hachemeister [17], mais ils ne sont pas abordés dans ce mémoire.

3.2 Application des modèles de crédibilité au produit AA

Il s'agit à présent d'appliquer les méthodes de crédibilité présentées dans la section précédente, pour proposer des ajustements a posteriori à la tarification via les GLMs.

La crédibilité est basée sur un principe de prise en compte de l'historique de l'observation. Un historique d'observations important permet d'avoir une robustesse dans les calculs. Les méthodes de crédibilité porteront donc sur un échantillon de contrats présents dans le portefeuille en 2012 et leur historique de sinistralité entre 2004 et 2012.

Nous analyserons la fréquence de sinistres puisque la sinistralité conditionnelle s'appuie avant tout sur la notion de fréquence liée, le coût moyen étant davantage lié au risque assuré et donc essentiellement subi.

Chaque modèle de crédibilité, est caractérisé par un ensemble de **paramètres de structure** qui permettent de déterminer le facteur de crédibilité α_k associé à chaque risque Θ_k. Ces paramètres devront être estimés, car ils sont inconnus[19]. Dans la littérature actuarielle, plusieurs auteurs ont proposé des estimations pour ces paramètres. En pratique, nous nous baserons sur celles proposées par BÜHLMANN et GISLER [16]. Ces estimateurs sont implémentés dans le package ***actuar*** développé par V. GOULET et al [18] sur **R**[20].

Ainsi dans la suite du mémoire,

- **Les observations X désignent un historique de fréquences de sinistres.**
- **Le poids a priori w correspond aux années assurance.**
- **La longueur de l'historique T est égale à 9 ans (2004 à 2012).**
- **Comme caractéristique de risque nous retenons :**
  - **L'activité de l'assuré**
  - **La famille d'engins**
  - **La puissance de l'automoteur.**
- **Afin d'illustrer la méthodologie, nous allons appliquer la crédibilité pour tarifer les risques en portefeuille en plusieurs étapes, en partant d'une seule variable de segmentation pour descendre jusqu'à l'application au niveau individuel.**

[19] Ils dépendent du risque Θ_k qui est lui-même inconnu.

[20] ***R*** *est le logiciel de statistiques permettant notamment l'étude de la crédibilité présentée dans le cadre de ce mémoire.*

3.2.1 Application du modèle de Bühlmann - Straub

A l'aide du modèle de crédibilité de Bühlmann - Straub, nous estimons la fréquence de sinistres de 2013 des différentes modalités de l'activité, de la famille d'engins et de la puissance de l'automoteur.

Autrement dit, on considère trois occurrences :

1. Activité $k \in \{AQCO, ARBO, \dots, VITI\}$
2. Famille d'engins $k \in \{A1, \dots, T6\}$
3. Puissance de l'automoteur $k \in \{Moins\ de\ 66\ CV, \dots, Plus\ de\ 501\ CV\}$

Nous présentons sur les Figures ci-dessous, la fréquence de sinistres passée (**trait de couleur**) et prédite (**rond de couleur**) pour les différents facteurs de risque considérés.

Hypothèse simplificatrice :

L'une des hypothèses de base de la crédibilité est la stationnarité, à savoir que le profil de risque du contrat ne dépende pas du temps.

Dans notre cas, nous pouvons remarquer que depuis 2010, la fréquence de sinistres du portefeuille présente une tendance à la baisse suite à l'instauration à cette date, des premiers coefficients d'ajustement à la sinistralité passée. Cette tendance ne concerne que 3 années sur un historique de 9 ans et ne concerne pas forcément toute la population du portefeuille étudié.

En outre, cette diminution est également liée à la déformation de structure du portefeuille vers des segments moins risqués en fréquence. Le fait d'appliquer la crédibilité sur des granularités d'observations de plus en plus fines, minore ainsi l'impact de la tendance dans les résultats obtenus. L'application de la crédibilité ayant pour objectif de « hiérarchiser » les contrats en fonction de leur sinistralité, le traitement de cette tendance ne constitue pas le cœur de notre approche. Ne souhaitant pas développer dans un premier temps un modèle trop complexe, nous prendrons donc le parti de développer un modèle de crédibilité classique sans retraiter cette tendance. Il est à noter que, dans une approche plus « pure », des modèles plus complexes de type Hachmeister [17] pourraient être appliquées.

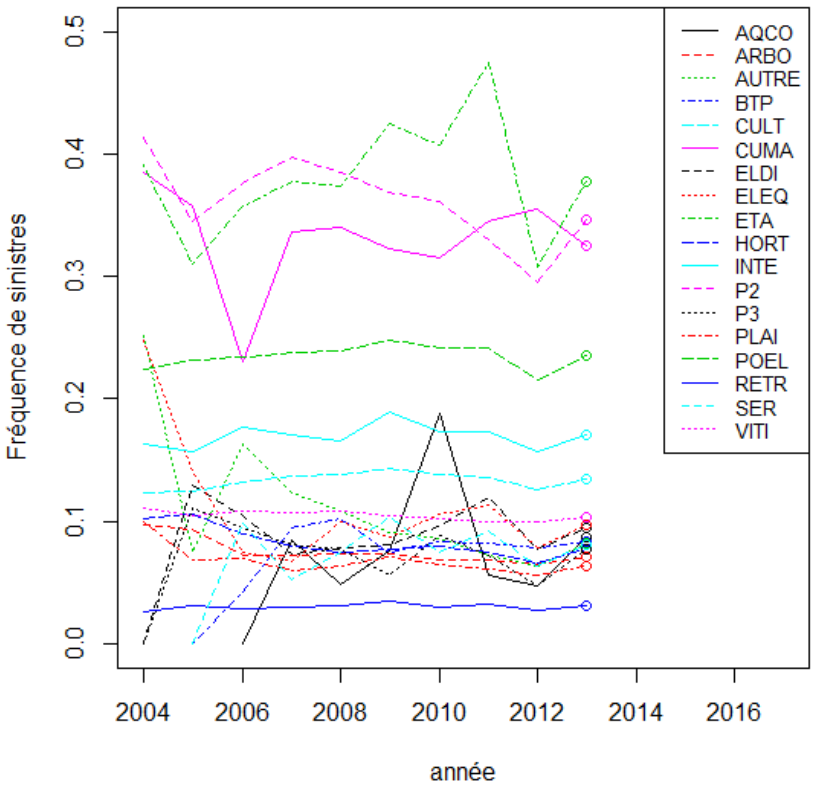


Figure 31 : Cas de l'activité

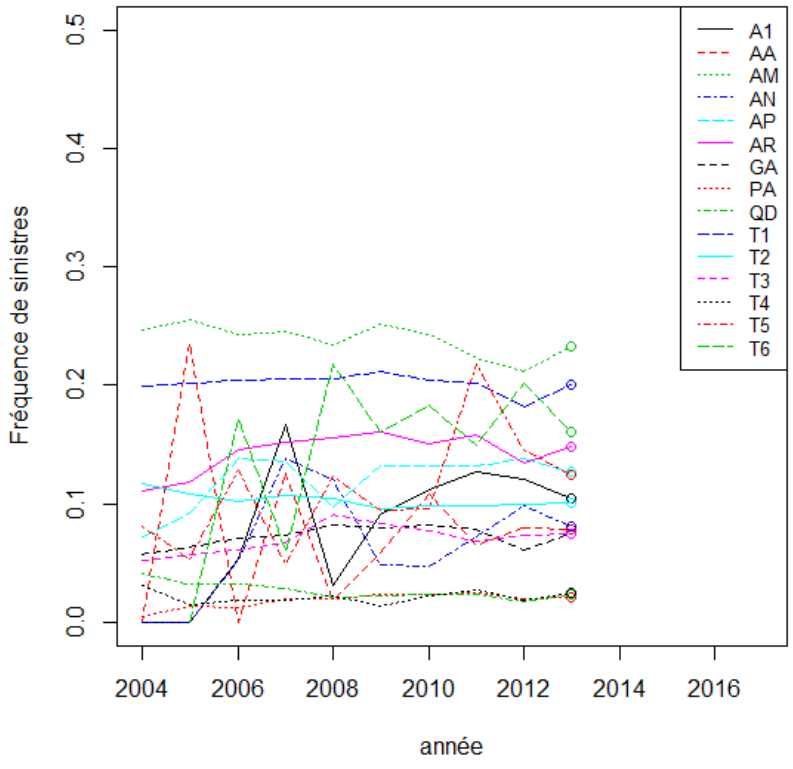


Figure 32 : Cas de la famille d'engins

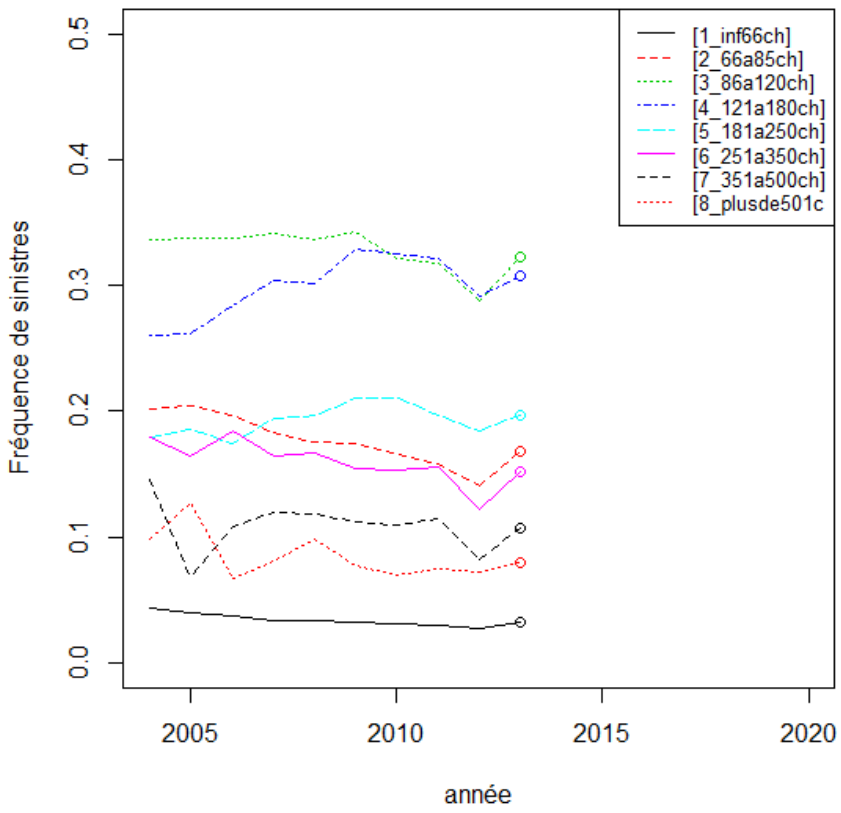


Figure 33 : Cas de la puissance de l'automoteur

La fréquence prédite par le modèle de crédibilité reflète bien la fréquence moyenne 2004-2012. Toutefois, le modèle de Bühlmann - Straub est perfectible dans une telle application de la tarification des produits automoteurs. Ces types de produits sont caractérisés par une anti-sélection très importante. Il est donc nécessaire dans l'évaluation de la fréquence de sinistres attendue, de considérer plus d'un facteur de risque pour affiner la segmentation.

3.2.2 Application du modèle de crédibilité hiérarchique

3.2.2.1 Modèle de crédibilité de Jewell : cas d'un modèle à 1 niveau intermédiaire

C'est un modèle de crédibilité hiérarchique qui fait intervenir un niveau supplémentaire de segmentation par rapport au modèle de Bühlmann - Straub. Nous appliquons donc ce modèle sur une combinaison de 2 variables explicatives.

Il n'existe pas de règle générale pour établir l'ordre des variables de segmentation. Nous décidons de considérer comme 1ère nœud de segmentation, la variable qui paraît la plus significative au sens des GLMs. Nous avons donc les hiérarchies suivantes :

- *Puissance + Puissance : Activité*

Dans ce cas, les indices h et k sont définis de la manière suivante :

- ✓ $h \in \{Moins\ de\ 66\ CV, \dots, Plus\ de\ 501\ CV\}$
- ✓ $k \in \{AQCO, ARBO, \dots, VITI\}$

- *Puissance + Puissance : Famille d'engins*

Dans ce cas, les indices h et k sont définis de la manière suivante :

- ✓ $h \in \{Moins\ de\ 66\ CV, \dots, Plus\ de\ 501\ CV\}$
- ✓ $k \in \{A1, \dots, T6\}$

- *Famille d'engin + Famille d'engin : Activité*

Dans ce cas, les indices h et k sont définis de la manière suivante :

- ✓ $h \in \in \{A1, \dots, T6\}$
- ✓ $k \in \{AQCO, ARBO, \dots, VITI\}$

Remarque

Nous décidons de regrouper dans chaque variable, certaines modalités pour permettre une meilleure visibilité.

- **Processus de regroupement :**

Dans un premier temps, nous appliquons le modèle hiérarchique sur les deux variables en question sans regroupement de modalités.

Dans les applications du modèle de Bühlmann-Straub, nous analysons la valeur des facteurs de crédibilité associés à chaque profil risque Θ_k. Cette analyse consiste à déterminer un seuil de la valeur du facteur de crédibilité tel qu'il nous permettra d'obtenir in fine 2 ou 3 modalités.

En pratique, pour chaque variable nous avons retenu un seuil $\alpha_k = 99{,}9\%$.

Ainsi tous les profils risques $\Theta_k$qui ont un $\alpha_k$inférieur à $99{,}9\%$ sont regroupés. Ce qui nous conduit au résultat suivant :

➔**Pour la variable puissance de l'automoteur :**

$$k \in \{Moins\ de\ 66CV, Entre\ 66CV\ et\ 85\ CV, Entre\ 86\ CV\ et\ \ 120\ CV, autre\ \}$$

➔**Pour la variable famille d'engins :**

$$k \in \{T1, Autre\}$$

➔**Pour la variable activité :**

$$k \in \{POEL, VITI, CULT, Autre\}$$

Nous présentons le résultat de la modélisation des deux derniers cas de figure. Le premier cas de figure, du fait d'un grand nombre de modalités, ne présente pas d'intérêt visuel.

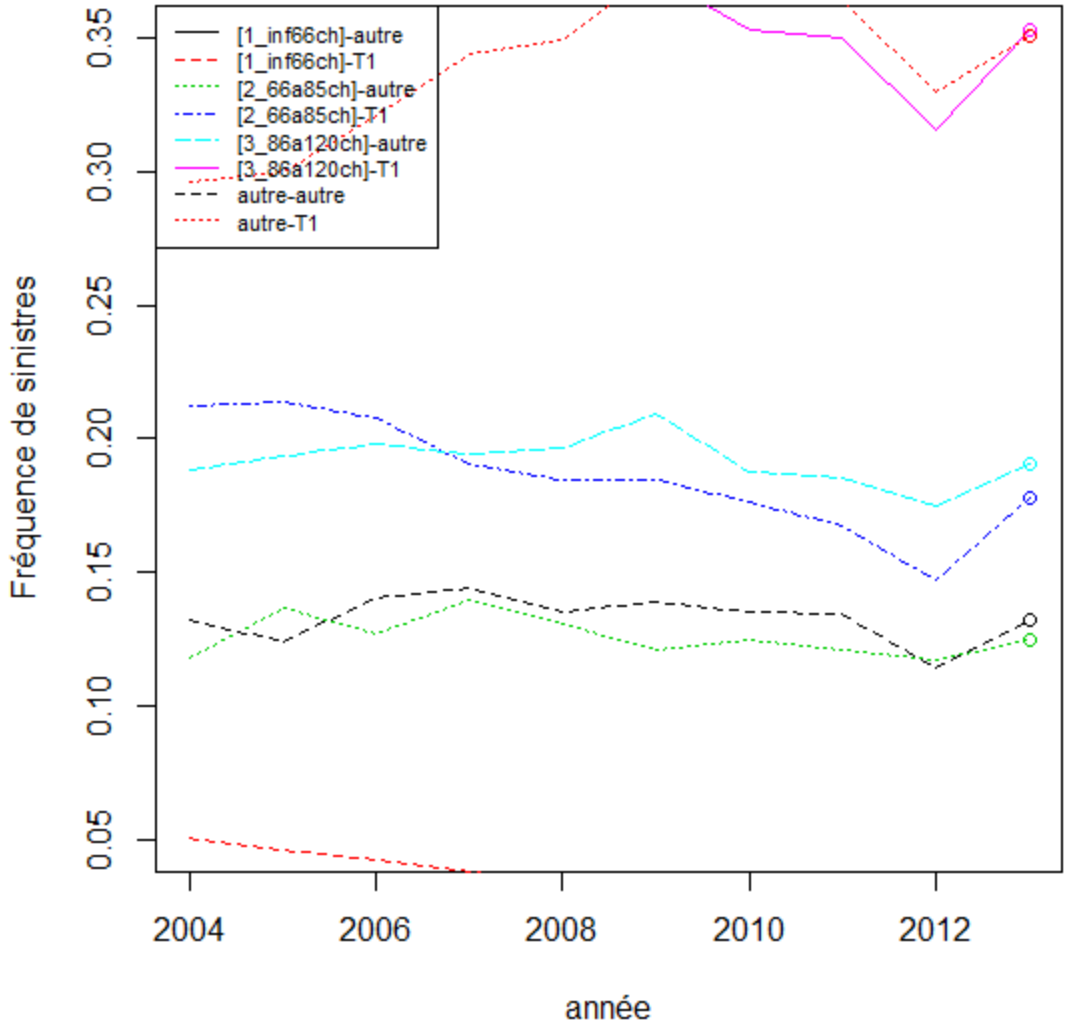


Figure 34 : Modèle de Jewell - Puissance vs Famille d'engins

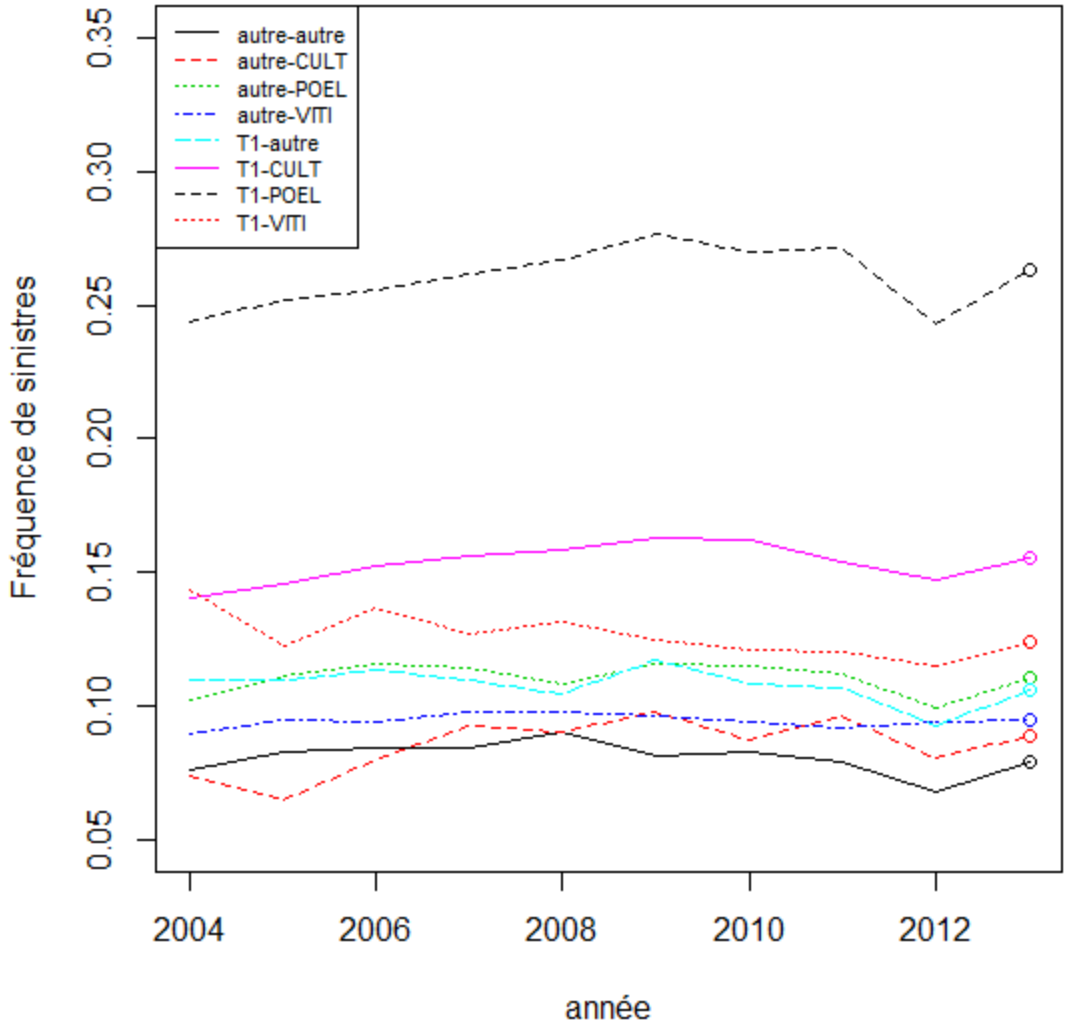


Figure 35 : Modèle de Jewell - Famille d'engins vs Activité

Lorsque nous croisons les caractéristiques de risque, on observe également une relative stabilité de la fréquence de sinistres passée. L'application du modèle de crédibilité hiérarchique sur ces croisements conduit bien à une estimation de la fréquence de sinistres attendue cohérente avec la moyenne 2004-2012. On observe également que la fréquence de sinistres d'un polyculteur éleveur (POEL) qui a un T1 comme automoteur, n'a pas la même fréquence de sinistres qu'un polyculteur éleveur qui a un autre type d'automoteur. L'introduction d'un nouvel axe de segmentation conduit à affiner l'estimation de la fréquence de sinistres en fonction des profils.

3.2.2.2 Modèle de crédibilité de Jewell : Cas d'un modèle à 2 niveaux intermédiaires

Nous combinons à présent les 3 variables explicatives de la fréquence de sinistres que sont la puissance de l'automoteur, la famille d'engins et l'activité de l'assuré. L'ordre d'établissement des nœuds de segmentation dépend de la pertinence des variables par rapport au risque fréquence. Du fait de la significativité de chaque variable, nous retenons la structure suivante :

- *Puissance+ Puissance : Famille d'engins+ Puissance : Famille d'engins : Activité*

Dans lequel, les paramètres d'indices k, h et g sont définis comme suit :

- ✓ $g \in \{Moins\ de\ 66\ CV, \dots, autre\}$
- ✓ $h \in \{T1, autre\}$
- ✓ $k \in \{POEL, CULT, VITI, autre\}$

Nous obtenons les graphiques suivants :

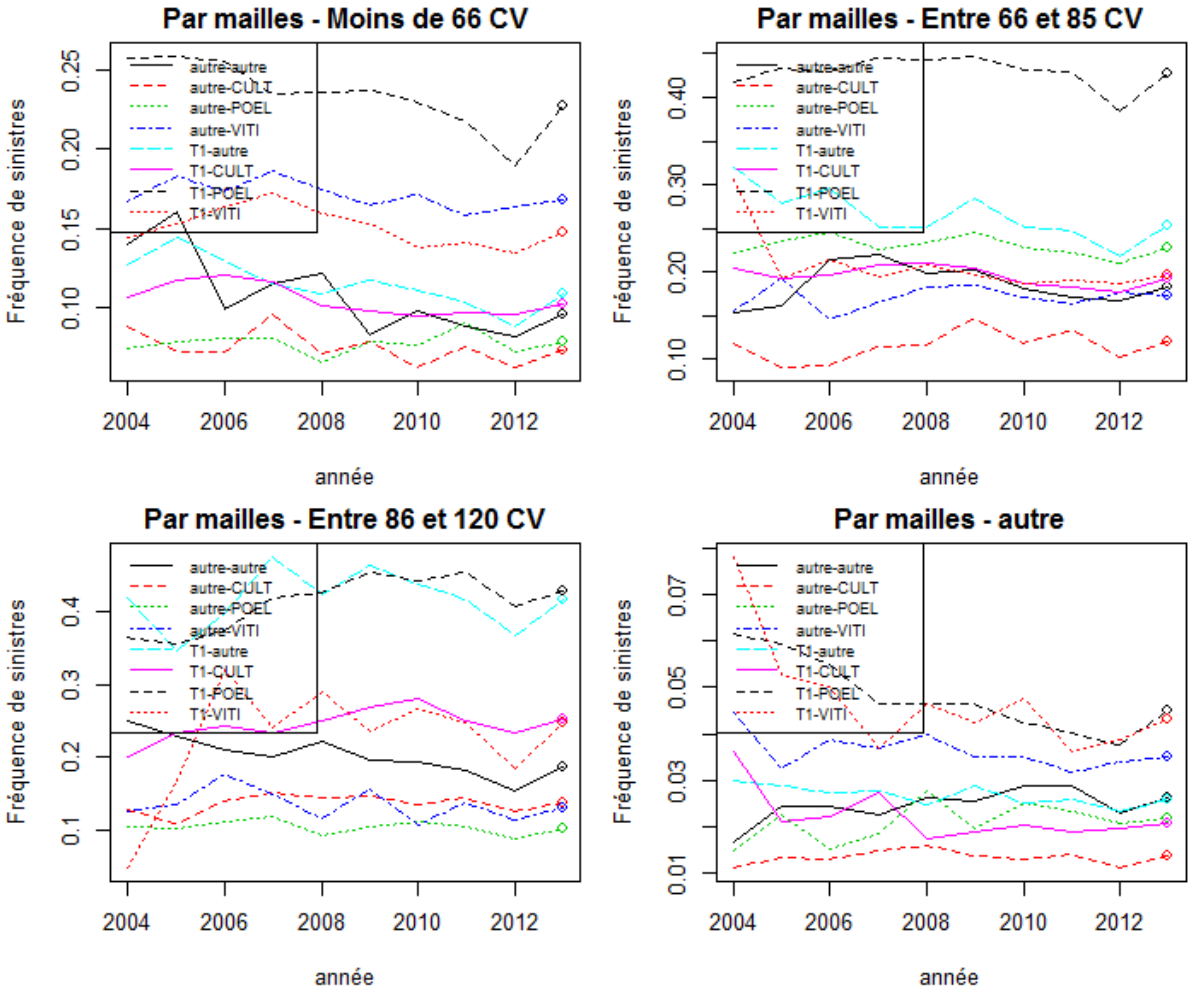


Figure 36 : Fréquence de sinistres estimée suivant différentes mailles de risque - Puissance

On retrouve une stabilité de la fréquence de sinistres sur les différentes mailles de risques et la fréquence de sinistres estimée semble également raisonnable. Le constat sur l'affinage de la fréquence de sinistres en fonction du nombre de variables explicatives considérées est encore valable ici.

3.2.3 Comparaison du modèle de crédibilité hiérarchique au modèle GLM

Nous décidons à présent de comparer les deux principales méthodes statistiques abordées dans ce mémoire, à savoir la théorie de la crédibilité et les modèles GLM basés sur un historique de 9 ans.

Calibrage des modèles

Les 3 variables utilisées pour l'application des méthodes de crédibilité précédemment, sont à nouveau conservées. Le modèle GLM appliqué, à la fréquence de sinistres, est paramétré par une loi Binomiale Négative et une fonction de lien $\log()$.

Lors de l'application des précédents modèles de crédibilité, nous avons observé plus de raffinement et de stabilité des données lorsqu'on considère des granularités plus fines. De ce fait, nous choisissons de ne considérer que le modèle hiérarchique à 2 niveaux intermédiaires.

Echantillon

Les modélisations sont appliquées sur un échantillon de contrats assez représentatif du portefeuille en 2012. Les individus sont définis par des **contrats profils risque**[21] qui représentent trois quart du portefeuille en 2012 en termes d'exposition ou années assurance.

[21] *L'expression « contrat profil risque» est employée pour désigner un ensemble de contrats ayant un profil de risque défini par la puissance de l'automoteur, la famille d'engins et l'activité. Ces groupes sont numérotés, particulièrement dans un ordre qui dépend de la taille de leur exposition total en 2012, d'où le terme* ***contrat profil risque ;***

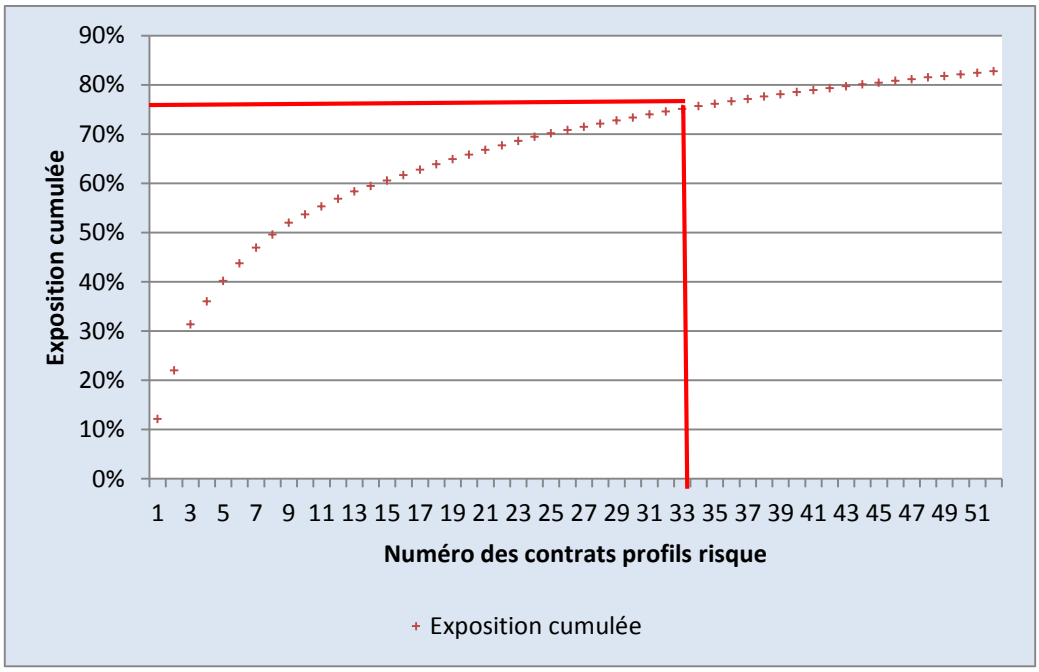


Figure 37 : Principaux profils de risque du portefeuille en 2012

D'après la **Figure** ci-dessus, nous décidons de ne garder que les 33 premiers **contrats profils risque.** Le **Tableau** suivant, présente les caractéristiques de risques des 5 premiers principaux contrats profils risque, soit 40 % du portefeuille en 2012.

N° du contrat profils risque	Caractéristiques de risque		
	Activité	**Famille d'engins**	**Puissance de l'automoteur**
1	POEL	T1	[3_86a120ch]
2	POEL	T1	[2_66a85ch]
3	POEL	T1	[1_inf66ch]
4	RETR	T1	[1_inf66ch]
5	POEL	T1	[4_121a180ch]

Tableau 18 : Caractéristiques des cinq premiers contrats profils risque

D'après le **Tableau** ci-dessus, en 2012, le portefeuille est principalement constitué de polyculteurs (**POEL**), possédant des tracteurs de type **T1**. La puissance de l'automoteur est assez variable suivant les profils contrats risque.

Représentation graphique

On note,

- **GLM_Freq** : La fréquence de sinistres obtenue par le modèle de régression de loi Binomiale Négative défini précédemment sur l'historique de 9 ans.
- **Crédibilité_Freq** : La fréquence estimée par le modèle hiérarchique à 2 niveaux intermédiaires sur l'historique de 9 ans.
- **Observe_Freq** : La fréquence de sinistres observée pour chaque individu dans l'échantillon considéré.

La **Figure** ci-dessous présente les résultats obtenus pour chaque individu de l'échantillon.

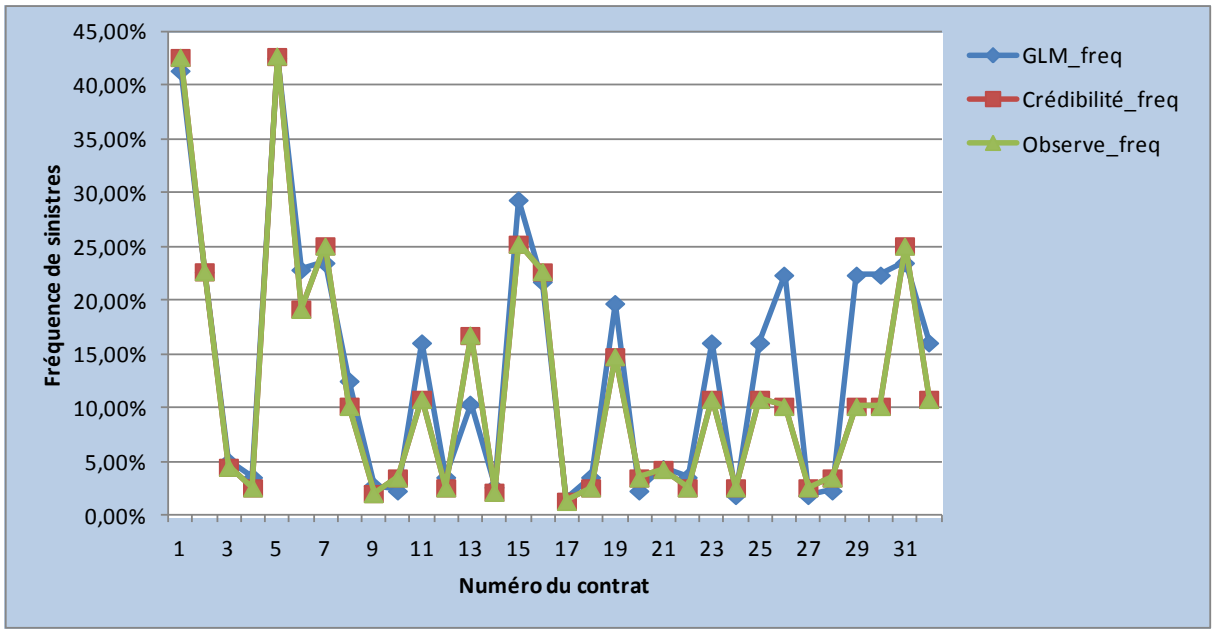


Figure 38 : Des fréquences de sinistres en fonction de différents individus

A l'exception de quelques contrats profils risque – qui représentent une faible part de l'échantillon – nous observons une quasi-juxtaposition des trois courbes. Si l'on considère le contrat profil risque numéro 29, il est caractérisé par la famille d'engins des quads (**QD**). Cette famille d'engins étant regroupée avec d'autres familles d'engins dans la modalité **autres**, c'est l'effet pur de cette dernière qui est estimé par le modèle GLM et non l'effet pur des quads. Les écarts sont donc dus à la simplification par regroupement de modalités, faite sur les modèles.

Cette première partie a consisté à appliquer le modèle de crédibilité sur notre base sinistres et à vérifier que nous pouvions réaliser une tarification, à l'instar des GLM, via cette méthode.

La plus-value majeure de la méthode de crédibilité consiste néanmoins à exploiter l'information (et la sinistralité) individuelle afin de moduler la prime agrégée de la case. C'est ce que nous allons étudier ci-dessous.

3.2.4 Analyse de la sinistralité individuelle

Ce qui fait la particularité d'un modèle de crédibilité, c'est que l'on peut également calibrer le contrat comme nœud de segmentation. Ceci permet d'estimer directement la fréquence de sinistres attendu d'un contrat précis, sur la base de son historique de sinistralité en particulier.

Dans le modèle hiérarchique précédemment présenté, nous rajoutons donc le contrat comme paramètre.

- *Puissance+ Puissance : Famille d'engins+ Puissance : Famille d'engins : Activité + Puissance : Famille d'engins : Activité : Contrat*

Dans lequel, les paramètres d'indices k, h et g sont définis comme suit :

- ✓ $q \in \{Moins\ de\ 66\ CV, \dots, autre\}$
- ✓ $g \in \{T1, autre\}$
- ✓ $h \in \{POEL, CULT, VITI, autre\}$
- ✓ *k correspond à un numéro de contrat*

Premier constat, nous remarquons que si le paramètre$w_k$définit l'année assurance du contrat k alors le paramètre $w_k$définit ainsi l'ancienneté du contrat k. La **Figure** ci-dessous, représente le facteur de crédibilité associé à chaque contrat en fonction de son ancienneté.

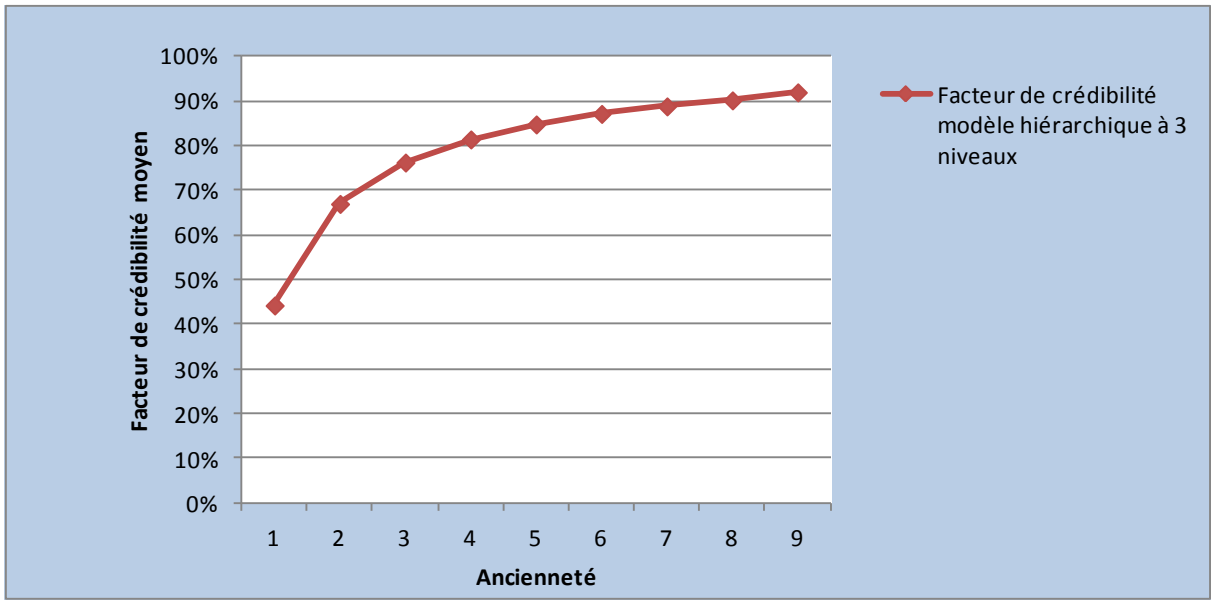


Figure 39 : Evolution du facteur de crédibilité en fonction de l'ancienneté

On observe ainsi que le facteur de crédibilité est bien une fonction croissante de l'ancienneté du contrat qui fait office de poids dans le modèle.

Sensibilité de la fréquence de sinistres estimée à la survenance d'un sinistre

Nous étudions à présent, le réajustement a posteriori obtenu par la crédibilité sur la survenance d'un sinistre. Pour cela on note,

$a = w_k$, l'ancienneté du contrat k.

$\tilde{f}(a)_k$, la fréquence de sinistres estimée par la crédibilité du contrat k d'ancienneté a.

$n(a)$, est le nombre de contrats d'ancienneté a.

Ainsi,

$$\tilde{\tilde{f}}(a) = \frac{1}{n(a)} \sum_{k=1}^{n(a)} \tilde{f}(a)_k.$$

$\tilde{\tilde{f}}(a)$, est la fréquence de sinistres estimée par la crédibilité pour une ancienneté de contrat égale à a.

Trois scénarios de sinistres sont étudiés :

(1) On considère que le contrat ne subit pas de sinistres depuis sa date de souscription jusqu'à la date d'inventaire (2012).

Figure 40 : Evolution de la fréquence de sinistre en cas de non survenance de sinistres dans le passé

(2) On considère que le contrat subit un sinistre en quatrième année de souscription seulement

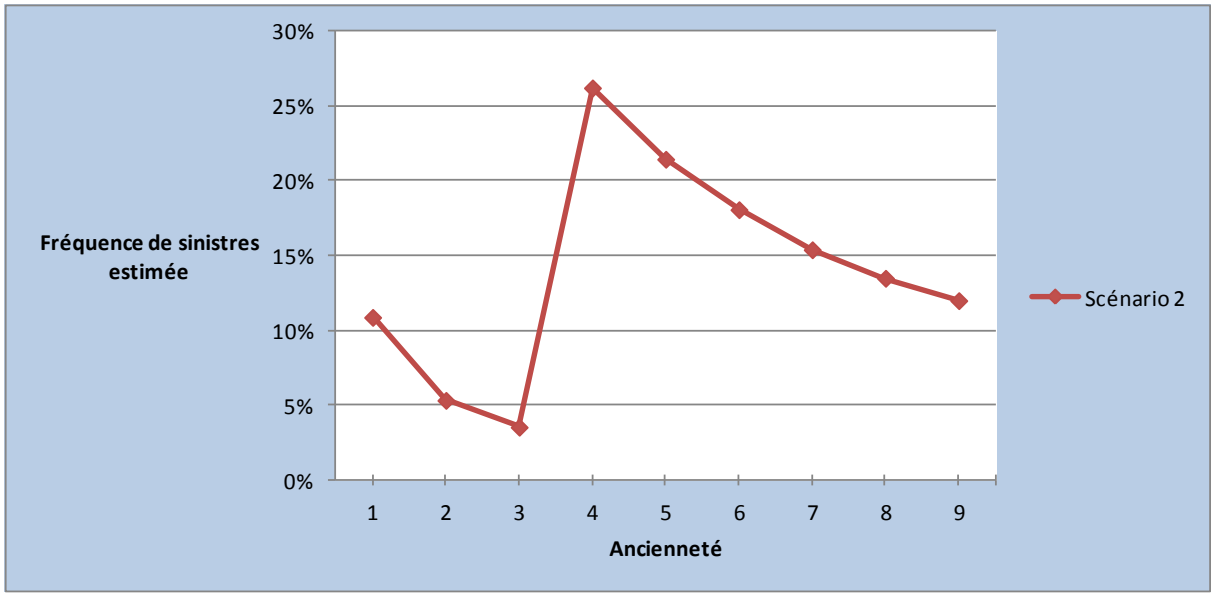


Figure 41 : Evolution de la fréquence de sinistre en cas de sinistre au 4ème exercice

(3) On considère que le contrat subit un sinistre en sixième année de souscription seulement

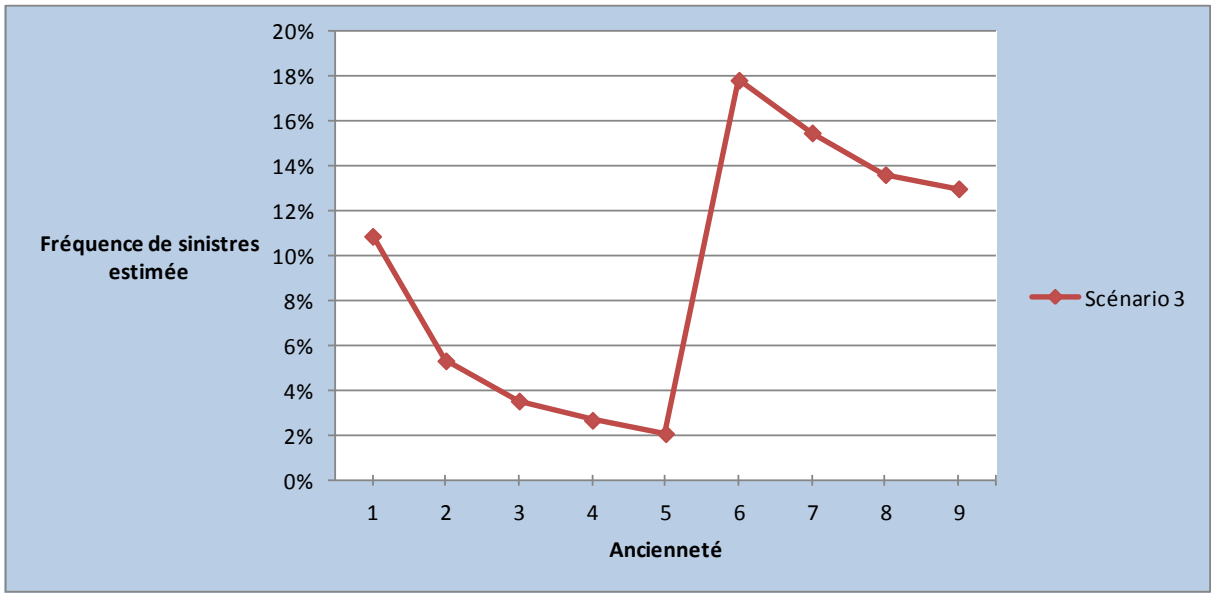


Figure 42 : Evolution de la fréquence de sinistre en cas de sinistre au 6ème exercice

A travers la simulation de 3 scénarios, nous avons déterminé les ajustements de la fréquence de sinistres estimée via un modèle de crédibilité hiérarchique suite à la survenance ou non d'un sinistre :

$$\frac{\tilde{\tilde{f}}(a)}{\tilde{\tilde{f}}(a-1)}.$$

Dans chacune des Figures représentées, nous observons une décroissance de la fréquence des sinistres en fonction de l'ancienneté tant que le contrat ne subit pas de sinistres. Cette décroissance est très forte puisqu'après 2 années sans sinistre, la fréquence crédibilisée est divisée par 3 par rapport à la fréquence de 1ère année, qui est une fréquence a priori. Puis la pente d'amélioration décroît au fil des années.

Parallèlement, la survenance d'un sinistre en 4ème année (4 ans d'ancienneté) entraine une majoration de 741 % de la fréquence de sinistres sur l'ancienneté suivante que l'on peut identifier à l'exercice suivant. Si le sinistre survient en 6ème année (6 ans d'ancienneté), on a une majoration plus sévère de 858%.

Cette application met en lumière un défaut majeur ou tout du moins une limite majeure de la méthode de crédibilité originale, à savoir la très forte volatilité des fréquences prédites et la sensibilité très forte à la survenance d'un sinistre. De tels « sauts » sur la fréquence prédite, et donc sur le potentiel tarif, ne sont pas acceptables dans une approche

opérationnelle et commerciale. Elle décrédibilise l'assureur et ne permettent pas de mettre en place une relation de confiance avec l'assuré.

Ainsi, en réalité, une grande majorité des contrats AA ne sont pas sinistrés sur l'ensemble de notre période d'étude. La forte baisse de la fréquence crédibilisée sur les trois premières années peut donc s'interpréter comme une sorte de « période de probation » durant laquelle l'absence de sinistre est très fortement prédictive d'une absence de sinistres dans le futur. En outre, la décroissance de la fréquence avec l'âge du tracteur (cf. partie 2 sur les GLM) plaide également en ce sens puisque la non déclaration de sinistres durant les premières années - sensibles au sinistre - traduit une forte probabilité de sinistres réduits dans le futur.

Si un système de crédibilité classique semble impossible à mettre en œuvre et renvoie la nécessité de recourir à un système de type bonus/malus plus régulé, cette décroissance très rapide serait quoiqu'il en soit à prendre en compte. A défaut, elle mettra en évidence le biais potentiel et la sur-tarification des contrats non sinistrés et le risque qu'un autre acteur du marché ne vienne capter ce risque. On rappelle ici que l'absence d'AGIRA sur le risque automoteur agricole vient limiter ce phénomène puisqu'un acteur ne peut pas connaître les antécédents sinistre d'un nouveau risque en dehors du déclaratif client.

De façon symétrique, cette forte concentration des sinistres génère une majoration très forte de la fréquence crédibilisée en cas de sinistre après plusieurs années. En effet, la fréquence liée étant très élevée, la survenance d'un sinistre après plusieurs années sans sinistre est très peu probable. En cas de réalisation du sinistre, le coefficient correctif est donc violent. Avec un effet de levier si le sinistre survient encore plus tardivement (après 6 ans, 7 ans…).

Au-delà du caractère très volatile exposé ci-dessus des réajustements proposés, rendant leur utilisation commerciale quasi impossible, l'utilisation pratique au quotidien des modèles de crédibilité n'est pas chose évidente. En effet, pour déterminer périodiquement des taux d'ajustement, il faudrait à chaque fois mener une étude complète : ce qui rend l'utilisation des modèles de crédibilité en pratique fastidieuse, dans le cadre de la tarification a posteriori.

Ainsi, bien que les méthodes de crédibilité permettent de prendre en compte dans la tarification la sinistralité individuelle, une approche plus opérationnelle de type bonus/malus est donc conseillée pour des études au quotidien.

Conclusion

Il existe aujourd'hui encore peu d'études actuarielles traitant spécialement la tarification des contrats d'assurance Automoteur Agricole. Ce produit a comme particularité d'avoir une sinistralité qui se concentre sur une partie du portefeuille. C'est l'une des raisons qui incitent le service actuariat PACIFICA, à faire évoluer la tarification du produit vers une individualisation de la cotisation.

Pour attester du rôle de la sinistralité individuelle, nous avons tout d'abord modélisé la sinistralité Automoteur Agricole en utilisant une approche « fréquence x coût moyen ». Nous avons pu ainsi mettre en évidence l'impact des critères tarifaires historiques présents dans la tarification et leur significativité. Une fois ce risque a priori calibré, nous avons démontré l'impact de la sinistralité antérieure, via la notion de fréquence conditionnelle puis via l'effet pur de cette information en GLM qui complète de façon très pertinente les critères de tarification a priori.

Nous avons ensuite étudié la mise en œuvre d'un modèle de crédibilité général afin d'identifier comment prendre en compte de façon optimale la sinistralité individuelle, et ce sans contrainte a priori. Approche alternative aux modèles GLM dans l'évaluation de la sinistralité, les modèles de crédibilité permettent en plus de prendre en compte la sinistralité individuelle d'un contrat pour évaluer sa prime pure attendue. De la sorte, il est possible de déterminer des coefficients de majoration/minoration en fonction de l'ancienneté du contrat.

Cependant les résultats obtenus n'ont pas été satisfaisants, dans la mesure où les coefficients d'ajustement déterminés étaient assez sévères. Cette volatilité est renforcée sur le contrat automoteur agricole par une concentration des sinistres sur un faible nombre d'individus. De fait, l'absence de sinistres sur courte période étant fortement prédictive d'une faible probabilité future de sinistres, les coefficients de crédibilité donnent un poids important à la sinistralité individuelle et génère des sauts de fréquence crédibilisée importants. Cette application montre là une des limites opérationnelles des méthodes de crédibilité linéaire appliquées sur une population d'individus dont la fréquence collective de sinistres est très faible en particulier dans le cas où la fréquence de sinistres est concentrée. La survenance, d'un sinistre entraine ainsi une correction brutale a posteriori.

Les systèmes bonus/malus semblent être une solution alternative. Ils ont en effet fait leurs preuves dans le cadre de la tarification en assurance automobile traditionnelle. D'autre part, le procédé d'application des systèmes bonus/malus se rapproche davantage d'un cadre opérationnel que les méthodes issues de la théorie de la crédibilité linéaire. Bien sûr, il sera avant tout nécessaire de tenir compte des spécificités du produit AA.

Tables des figures

Liste des Tableaux

Bibliographie

[1] FFSA, «L'assurance des biens agricoles en 2011,» Fédération Française des Sociétés d'Assurance, 2012.

[2] A. CHARPENTIER, «Le Bonus-Malus Français a-t-il encore un avenir ?,» *Risques - Les Cahiers de l'assurance, Numéro 83,* Septembre 2010.

[3] J. BEIRLANT, Y. GOEGEBEUR, J. SEGERS, J. TEUGELS, D. D. WAAL et C. FERRO, Statistics of Extremes: Theory and Applications, WILEY, 2004.

[4] A. C. M. DENUIT, Mathématiques de l'assurance non-vie. Tome 2 : Tarification et Provisionnement, Paris: Economica, 2005.

[5] E. OHLSSON et B. JOHANSSON, Non-Life Insurance Pricing with Generalized Linear Models, Springer, 2010.

[6] G. DARMOIS, «Sur les lois de probabilités à estimation exhaustive,» *C.R. Acad. Sci. Paris,* p. 1265–1266, 1935.

[7] J.-P. DEDIEU, Points Fixes, Zéros et la Méthodes de Newton, Springer Verlag, 2006.

[8] H. AKAIKE, «A new look at the statistical model identification,» *IEEE transactions on automatic,AC-19,* p. 716–723, 1974.

[9] G. SCHWARZ, «Estimating the dimension of a model,» *Annals of statistics, 6,* p. 461–464, 1978.

[10] D. ANDERSON, S. FELDBLUM, C. MOLDIN, D. SCHIRMACHER, E. SCHIRMACHER et N. THANDI, «A practitioner's Guide to generalized linear models,» Watson Wyatt, 2007.

[11] Towers Watson, «Technologie et logiciels - Emblem,» [En ligne]. Available: http://www.towerswatson.com/fr-FR/Services/Tools/emblem.

[12] A. MOWBRAY, «How extensive a payroll exposure is necessary ?,» *Proceedings of the Casualty Actuarial,* p. 25–30, 1914.

[13] A. W. WHITNEY, «The theory of experience rating,» *Proceedings of the Casualty Actuarial Society,* p. 275–293, 1918.

[14] H. BÜHLMANN, «Experience rating and credibility,» *ASTIN Bulletin, volume 4,* p. 199–207, 1967.

[15] H. BÜHLMANN, «Experience rating and credibility,» *ASTIN Bulletin, vomule 5,* p. 157–165, 1969.

[16] R. NORBERG, « The credibility approach to experience rating,» *Scandinavian Actuarial Journal,*

vol. 4, pp. 181-221, 1979.

[17] H. BÜHLMANN et A. GISLER, A Course in Credibility Theory and its Applications, Springer, 2005.

[18] W. JEWELL, «The use of collateral data in credibility theory: a hierarchical model,» *Journal de l'institut italien des actuaires,* n° %138, pp. 1 - 16, 1975.

[19] C. DUTANG, V. GOULET et M. PIGEON, «actuar: An R Package for Actuarial Science,» Journal of Statistical Software, 2008. [En ligne]. Available: http://www.jstatsoft.org/v25/i07.

[20] Ministère de l'alimentaion de l'agriculture et de la pêche, «Réglementation des tracteurs agricoles ou forestiers,» 2009.

[21] Fédération Française des Sociétés d'Assurance (FFSA), «Les assurances de biens et de responsabilités - Données clés 2012,» ASSOCIATION FRANÇAISE DE L'ASSURANCE, 2013.

[22] G. THIRY, «Assurance Automobile, Etudes Actuarielles et Politique Tarifiare en 2001,» *BFA, Vol. 4, N°7,* pp. 61-82, 2001.

[23] A. GUILLOU et A. YOU, «Introduction à la théorie des valeurs extrêmes : Applications en Actuariat,» chez *Université de Strasbourg & Société Générale Insurance*, Strasbourg, 2011.

[24] A. SOULEAU, «Tarification de la branche d'assurance des accidents de travail, Bonus-malus et Crédibilité,» Euria, Mémoire d'actuariat, 2010.

[25] B.CROCHET, 150 ans de machinisme agricole, Éditions de Lodi, 2006.

[26] PACIFICA, «Documents techniques sur Automoteur Agricole».

[27] F. PLANCHET, «Assurances des biens et de responsabilité - Enjeux actuariels,» Formation Groupama, 2012.

[28] G. TAYLOR, «Loss Resserving : An Actuarial Perspective,» *Kluwer Academic Publishers,* 2000.

[29] W. KUTZBACH, «Developments in European combine harvesters,» chez *paper 96 A - 069 - Ageng 96*, Madrid, 1996.

[30] A. M. M. P. L. LEBART, Statistiques Exploratoire Multidimensionnelle, Paris: Dunod, 2000.

[31] M. TIEN, «Etude de la sinistralité sur la garantie incendie du produit Multirisque Habitation,» Master 2 ISF,Université Panthéon-Assas Paris 2, 2011.

[32] M. EFROYMSON, Multiple regression analysis, Mathematical Methods for Digital Computers. Wiley, 1960.

[33] J. N. P. Mc CULLAGH, Generalized Linear Models, New York: Chapaman et Hall, 1989.

[34] T. NGUYEN, «Flottes automobiles : Un nouveau modèle de tarification. Impact de la conservation sur la distribution du ratio sinistres à primes,» CNAM, Institut des Actuaires, 2008.

[35] F. PICAR, A. CLEMENT-GRANCOURT et A. COHEN, «MEMOIRES D'ACTUARIAT : Recommandations,» Institut des Actuaires, Paris, 2010.

[36] F. F. d. S. d'Assurance, «L'assurance des biens agricoles en 2011,» 2012.

[37] PACIFICA, «Assurance tracteurs, matériels et autres automoteurs,» [En ligne]. Available: http://www.credit-agricole.fr/agriculteur/assurances/protegez-vos-biens-professionnels-et-prives/assurance-tracteurs-materiels-et-autres-automoteurs.html.

[38] J. HILBE, Negative Binomial Regresison, Cambridge University Press, 2007.

[39] R. W. KEENER, «Statistical Theory: Notes for a Course in Theoretical Statistics,» *Springer,* p. 27–28;32–33, 2006.

[40] M. U. IQBAL et S. LIM, «A Privacy Preserving GPS-based Pay-as-You-Drive Insurance Scheme,» International Global Navigation Satellite Systems Society , 2006.

[41] S. PITREBOIS, P. D. LONGUEVILLE, M. DENUIT et J.-M. WALHIN, «Etude de techniques IBNR modernes,» *actu-L 2,* pp. 29-62, 2002.

[42] J. DROESBEKE, M. LEJEUNE et G. SAPORTA, «Modèles statistiques pour données qualitatives,» *Société Française de STATISTIQUE,* 2005.

[43] O. VASECHKO, M. GRUN-REHOMME et N. BENLAGHA, «Modélisation de la Fréquence de sinistres en Assurance Automobile,» *Bulletin Français d'Actuariat, Vol. 9, n°18,* pp. 41 - 63, Juillet- décembre 2009.

[44] J. NELDER et R. WEDDERBURN, «Generalized linear models,» *Journal of the Royal Statistical Society, Series A 135,* pp. 370-384, 1972.

[45] A. CHARPENTIER, «LE BONUS-MALUS FRANÇAIS A-T-IL ENCORE UN AVENIR ?,» *Risques,* n° %183, 2010.

Annexes

A. CODE DES ASSURANCES

Article L211-1

« *Toute personne physique ou morale dont la responsabilité civile peut être engagée en raison des dommages subis par des tiers (atteintes aux personnes ou aux biens) dans la réalisation desquels un véhicule terrestre à moteur, ainsi que les remorques ou semi-remorques, est impliqué, doit pour faire circuler les dits véhicules, être couverte par une assurance garantissant cette responsabilité.* »

B. V DE CRAMER

Le V de Cramer est une mesure de corrélation entre deux variables explicatives catégorielles. Il est définie par

$$\sqrt{\frac{\sum_{i,j}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}{\min\big((a-1),(b-1)\big)\times n}}$$

Où :

$a =$ Nombre d'modalités de la 1ère variable

$b =$ Nombre d'modalités de la 2nd variable

n_{ij} =L'exposition de la $i^{ème}$ modalité du 1er facteur et de la $j^{ème}$ modalité du 2nd facteur

$$n = \sum_{i,j} n_{ij}$$

$$e_{ij} = \frac{\sum_i n_{ij} \times \sum_j n_{ij}}{n}$$

Le V de Cramer varie entre 0 et 1.

- Une valeur de 0 traduit une indépendance entre les 2 variables.
- Une valeur de 1 signifie que la connaissance de l'influence d'une des variables explicatives sur la variable à expliquer permet d'enduire celle de l'autre variable explicative sur la variable à expliquer

C. REPRESENTATION DU COUT MOYEN EN FONCTION DE QUELQUES VARIABLES SIGNIFICATIVES

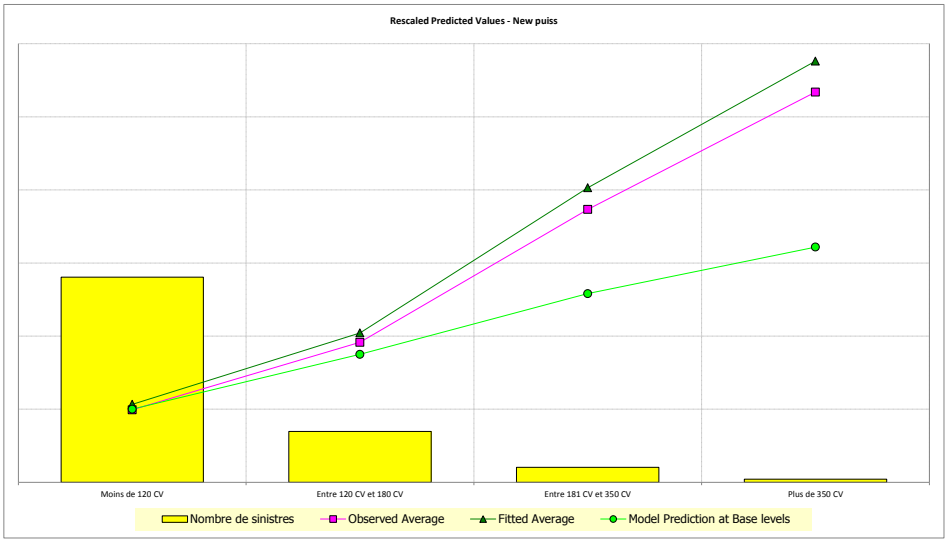


Figure 43 : Le coût moyen en fonction de la puissance

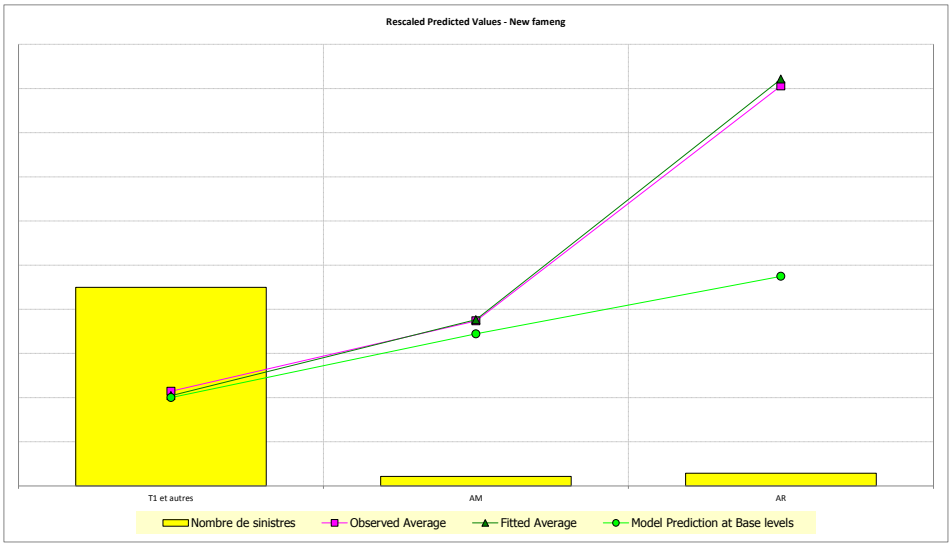


Figure 44 : Le coût moyen en fonction de la famille d'engins

Représentation de la fréquence de sinistres en fonction de quelques variables significatives

D. MODÈLE DE CRÉDIBILITÉ DE JEWELL

Combinaison Puissance vs Activité

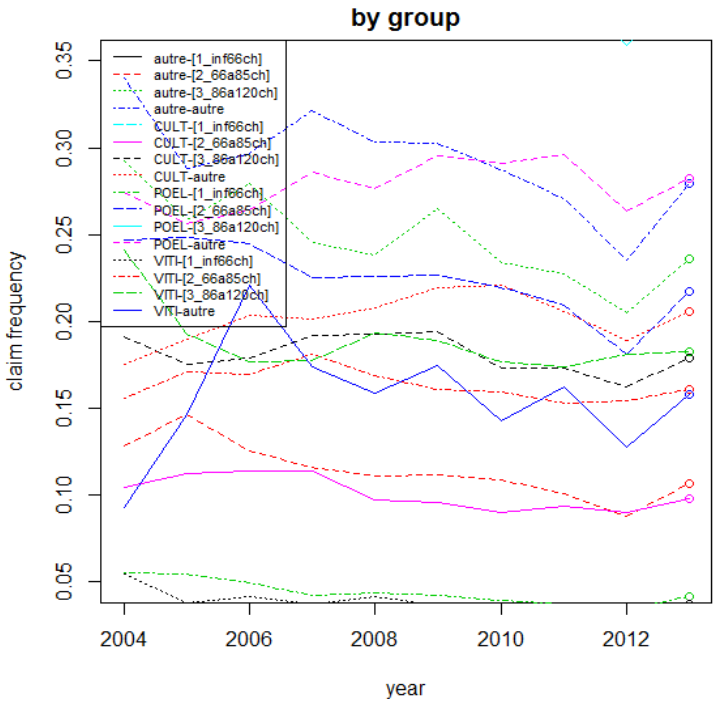


Figure 45 : Modèle de jewell - Puissance de l'automoteur vs Activité

E. ADEQUATION DE LA LOI EMPIRIQUE A UNE LOI THEORIQUE

Dans cette section, nous étudions l'adéquation de la distribution de la fréquence de sinistres et de celle du coût moyen des sinistres sur la base d'études AA par rapport à des lois théoriques usuelles. Si les tests à plat (QQ-plot, test du Chi2 ou test de Kolmogorov-Smirnov) permettent d'avoir un a priori sur la loi théorique que suit la distribution empirique, ils sont toutefois fortement perfectibles. Dans ces tests, on ne tient pas compte de la dépendance de la variable d'intérêt (le nombre de sinistres ou la charge) avec d'autres paramètres (les variables explicatives).

Nous proposons ainsi d'analyser les résidus issus de la modélisation GLM, c'est en effet un bon indicateur de l'adéquation aux lois théoriques.

- **Analyse des résidus**

Ces tests sont effectués en supposant une fonction Log comme fonction de lien et comme variables explicatives, la liste des variables retenues dans le modèle fréquence.

<u>Cas de la fréquence moyenne des sinistres</u>

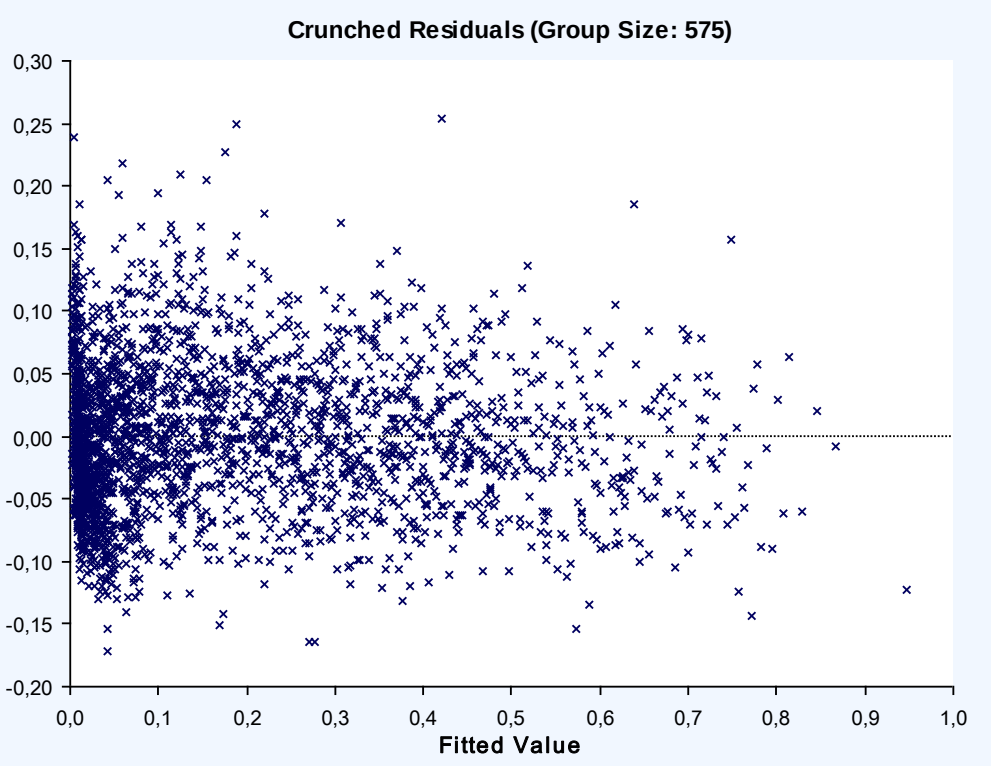


Figure 46 : Nuage des résidus en fonction des valeurs prédites – Cas de la Binomiale Négative

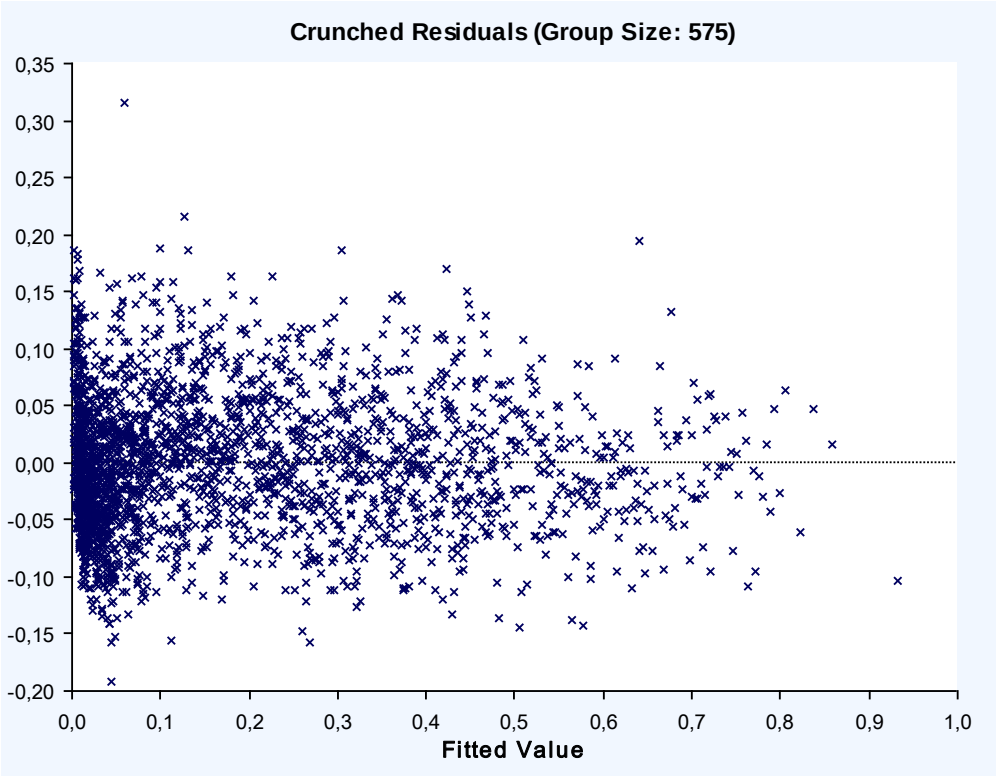


Figure 47 : Nuage des résidus en fonction des valeurs prédites – Cas de la Poisson

Dans les deux cas, la distribution du nuage des résidus autour de l'axe des abscisses est raisonnablement symétrique. Le modèle peut être aussi bien ajusté avec une loi Poisson qu'avec une loi Binomiale Négative.

Cas du coût moyen des sinistres

Dans le cas du coût moyen, l'analyse des résidus ne nous permet pas totalement de conclure d'une adéquation des observations à la loi théorique. Nous avons étudié sous plusieurs angle de représentation des résidus l'adéquation des observations en fonction de :

- La loi gamma
- La Tweedie à dont le paramètre $\xi = 1{,}5$
- La loi Inverse Gaussienne qui équivaut à une Tweedie ($\xi = 3$)

Ces lois sont étudiées sur la base des mêmes variables explicatives et une fonction log comme fonction de lien.

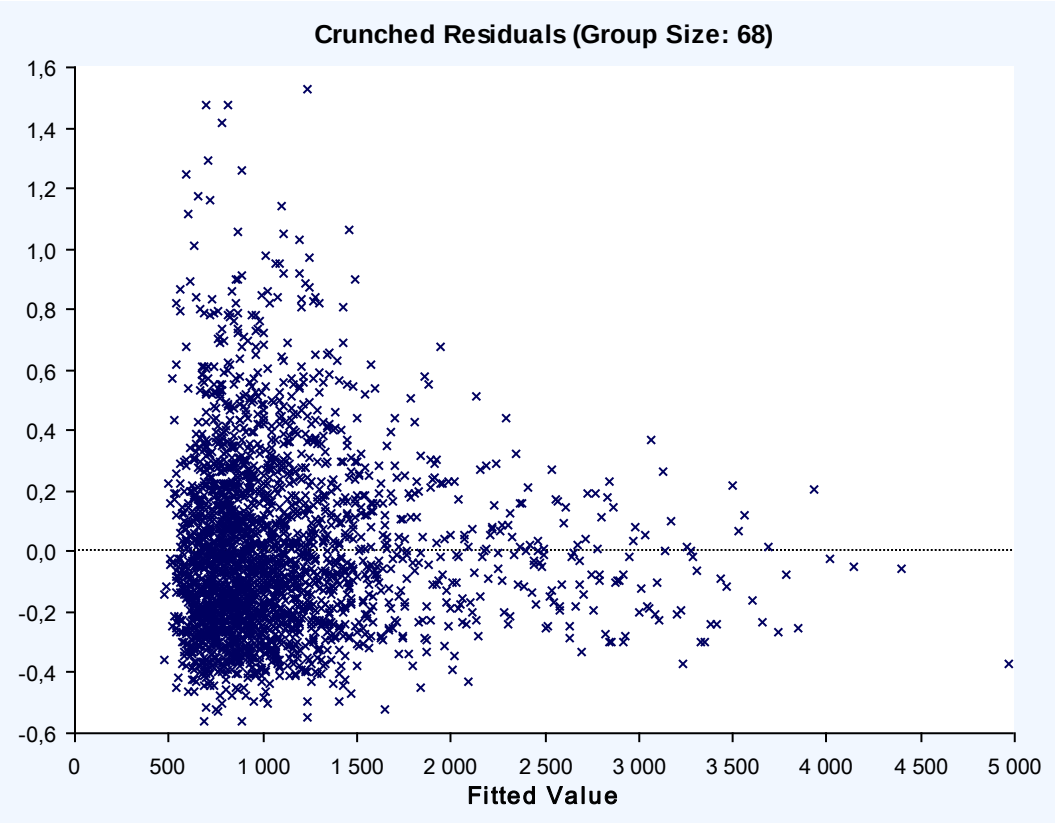


Figure 48 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Gamma

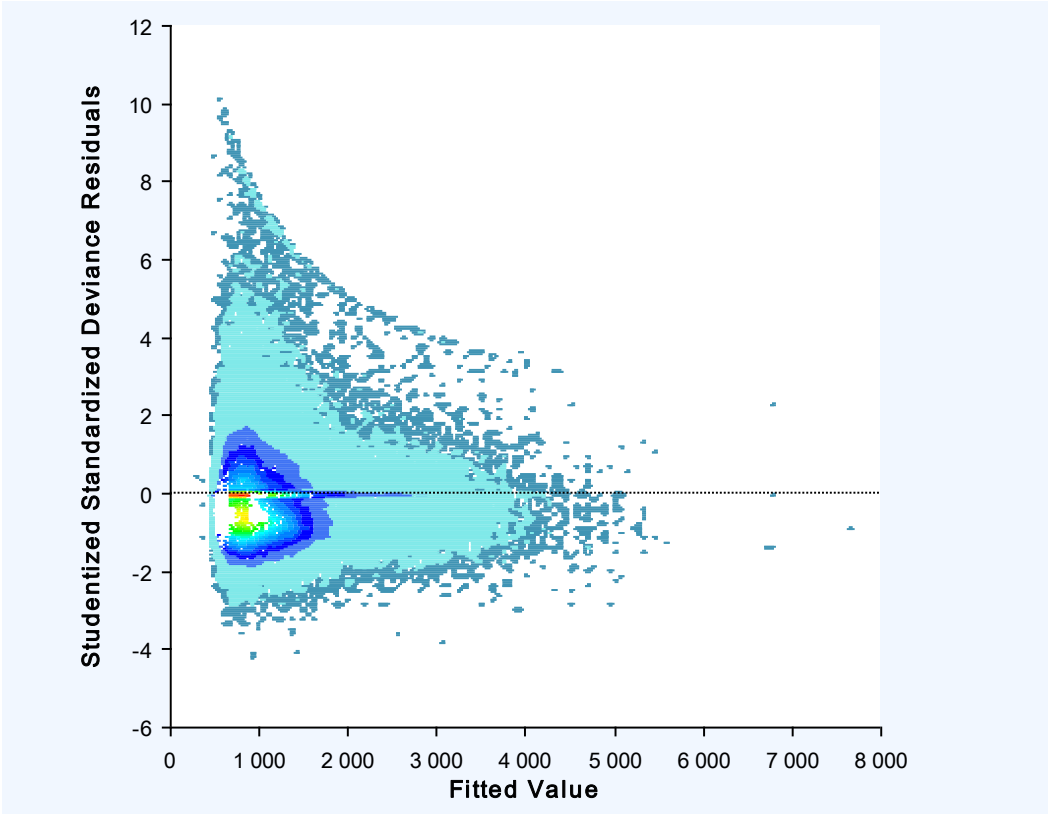


Figure 49 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Gamma

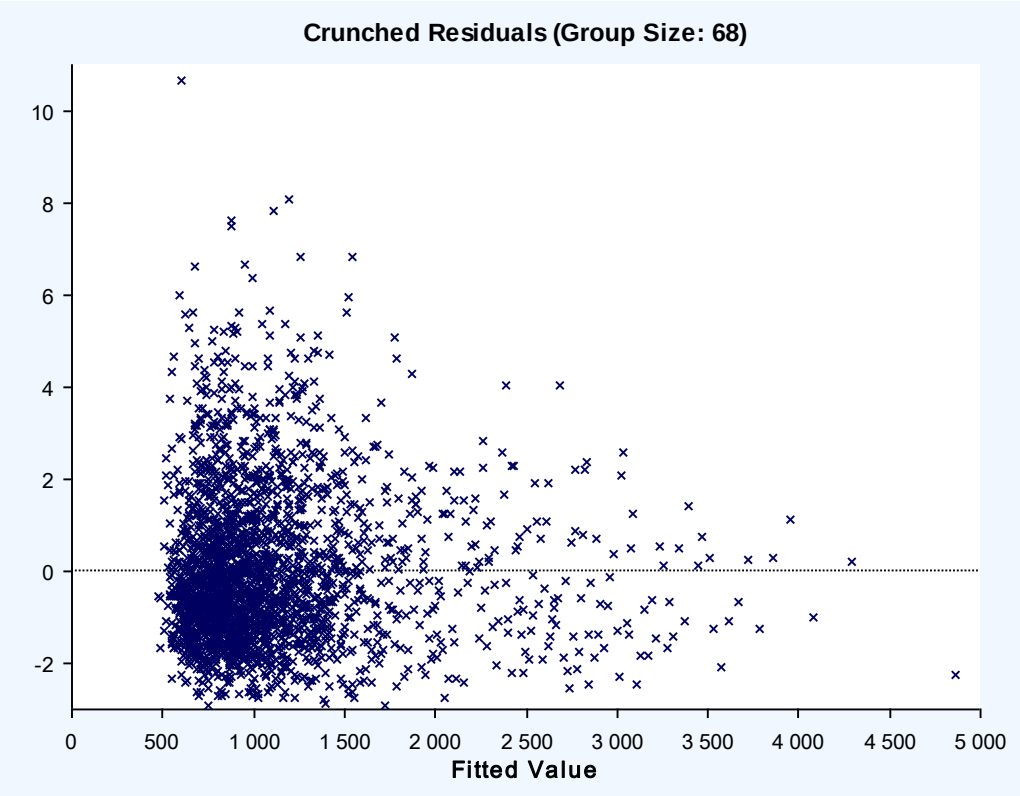


Figure 50 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Tweedie

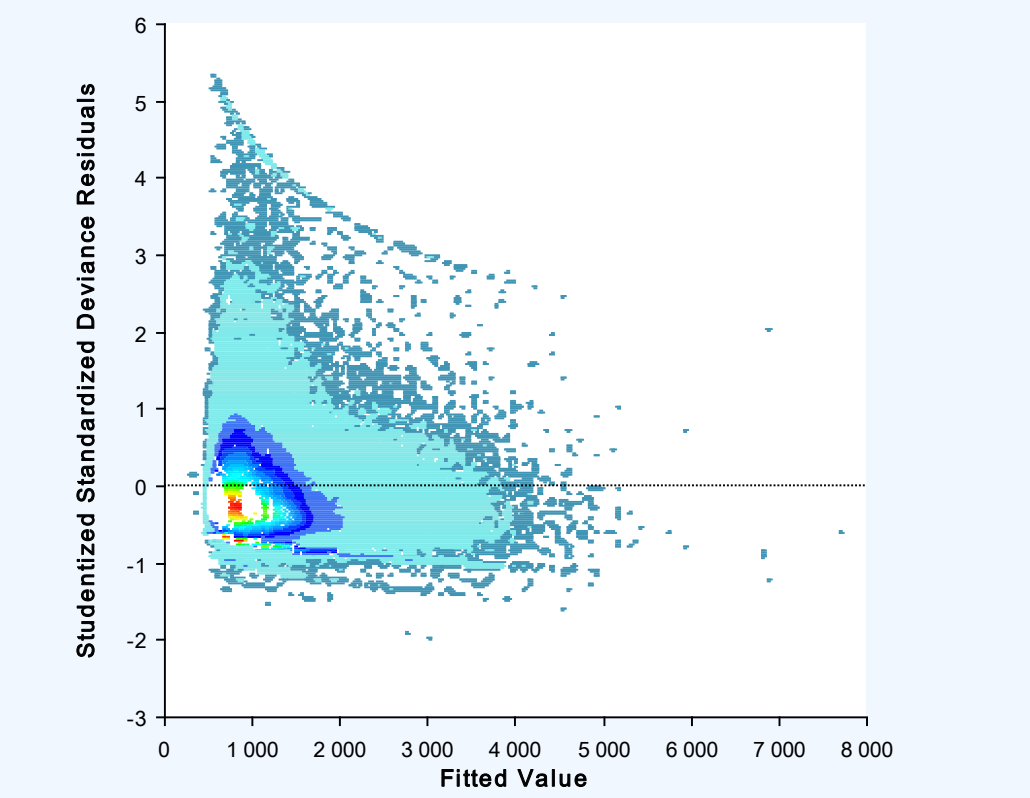


Figure 51 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Tweedie

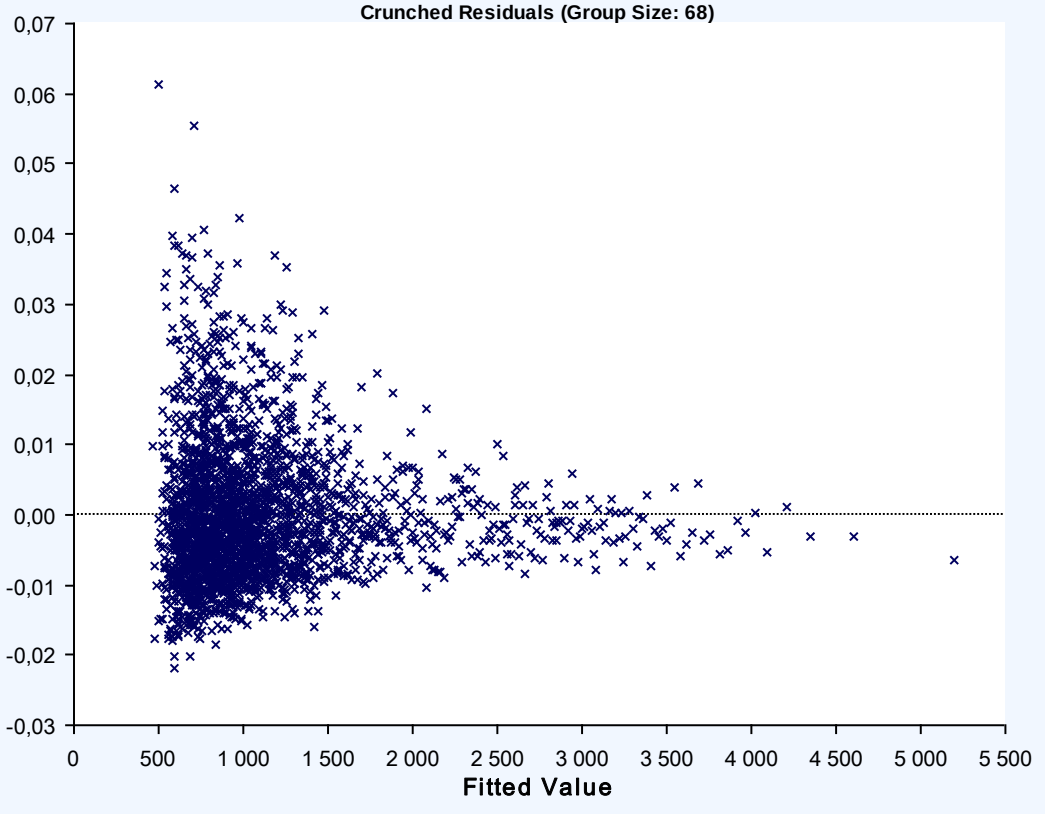


Figure 52 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Inverse Gaussienne

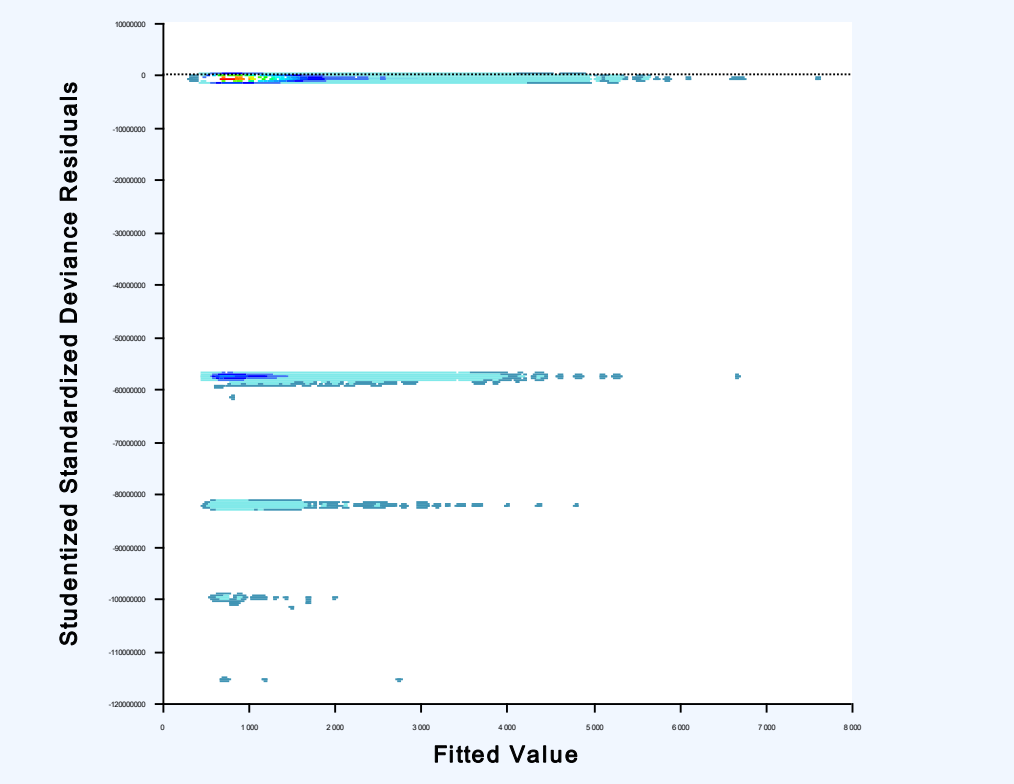


Figure 53 : Nuage des résidus en fonction des valeurs prédites – Cas de la loi Inverse Gaussienne