



Mémoire présenté le :

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

Par : KOUO Kasséa Kévin Axel

Titre TARIFICATION AUTOMOBILE : GLM vs RÉSEAUX DE NEURONES

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

*Membre présents du jury de l'Institut
des Actuaire*

signature

Entreprise : GENERALI France

Nom : GANGNEUX Karine

Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :

Nom : SOURISSEAU Jérôme

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Secrétariat

Signature du candidat

Bibliothèque :

RESUME

Mots clés : Tarification, assurance Automobile, Modèle linéaire généralisé, Modèle coût-fréquence, Machine Learning, Réseau de neurones, Perceptron Multicouche, Approximation universelle, ROC, bagging

Le marché de l'assurance non-vie, en particulier celui de l'Auto est fortement concurrentiel. Celui-ci nécessite de la part de l'assureur d'être à la fois commercialement viable et de couvrir son risque au « juste prix ». Les assureurs tendent à s'orienter vers une tarification ajustée aux besoins et aux caractéristiques propres du client. L'expansion du digital et du recueil de l'information peuvent fournir aux assureurs une quantité de données non structurées.

Les méthodes classiques de tarifications, telles que le GLM, comportant des hypothèses statistiques lourdes ne peuvent modéliser aisément ce type de données. Nous nous sommes donc orientés vers les méthodes de machine learning, en particulier les réseaux de neurones car ces méthodes sont non-paramétriques. Certains réseaux de neurones, comme le perceptron multicouche, possèdent la propriété d'approximation universelle, permettant d'approcher n'importe quelle fonction à peu près régulière.

Nous comparons dans ce mémoire les prédictions issues du GLM et celles du perceptron multicouche sur notre portefeuille d'assurance Automobile. Nous déterminons, par ailleurs un estimateur agrégé issu des deux méthodes à l'aide de la méthode du bagging.

ABSTRACT

Keys words: Pricing, Non-life Insurance, Generalized Linear Modeling, frequency-severity model, Machine Learning, Neural networks, Multilayer Perceptron, Universal approximation, ROC, bagging.

Market of non-life insurance, especially auto insurance, is very competitive. Insurance companies have to be more attractive, and to do so, they must give the fair price for each customer. Pricing products, per their needs and taking count of the customer's characteristics is very important. The growth of digital and the possibility to have more information about customers give a huge quantity of non-structural data.

Traditional pricing methods, such as GLM, cannot model easily those types of data because of statistics hypothesis. It's the reason why we have focused on machine learning methods; more specifically neural networks because they are non-parametric methods. Some types of neural networks as Multilayer Perceptron, owns universal approximation property which permit to approach any regular function.

We compare in this study the prediction based on GLM and those coming from multilayer perceptron on our portfolio. Finally, we determine an aggregated estimator provided on both methods, by using bagging.

REMERCIEMENTS

Tout d'abord, je souhaiterais remercier Karine GANGNEUX, manager du service solution d'assurance dommage pôle traditionnel de m'avoir accueilli au sein de son équipe et de m'avoir permis de réaliser ce mémoire sur le thème de la tarification en Automobile au sein de l'univers client particulier.

Je tiens à remercier tout particulièrement mon maître d'apprentissage Jérôme SOURISSEAU, pour son encadrement, sa disponibilité et la mise à disposition de moyens me permettant de mener à bien ce projet.

Je remercie l'ensemble des collaborateurs de la direction des partenariats pour leur conseil et leur gentillesse. Je remercie particulièrement Samuel CHEMAMA, Kévin YANG pour leur précieux soutien.

Je remercie mon tuteur à l'ISFA, Yahia SAHLI, pour son suivi jusqu'à l'achèvement de ce mémoire.

Aussi, un grand merci à ma famille et mes amis pour leur présence et leurs encouragements.

Enfin, un remerciement spécial à mon épouse pour son soutien moral durant toute cette année académique.

Table des matières

RESUME	2
ABSTRACT	3
REMERCIEMENTS	4
INTRODUCTION GENERALE	7
PREMIERE PARTIE	8
CONTEXTE ECONOMIQUE ET ANALYSE EXPLORATOIRE.....	8
I. Le marché de l'assurance en France	9
A. Un marché concurrentiel.....	9
B. L'assurance automobile en France.....	10
C. Environnement législatif	11
II. Présentation de l'entreprise.....	11
A. GENERALI.....	11
B. La direction des partenariats.....	12
C. L'ÉQUITÉ	12
III. Analyse des variables explicatives.....	13
A. Constitution de la base.....	13
B. Variables d'étude	14
C. Analyse graphique des variables continues	24
D. Discrétisation des variables continues	26
E. Discrétisation non supervisée	27
F. Discrétisation automatique supervisée.....	29
IV. Sinistres graves	36
A. Valeurs extrêmes et dépassement de seuil.....	38
B. Distribution conditionnelle des excès	40
C. Les méthodes d'estimation du seuil.....	41
DEUXIEME PARTIE	45
GLM EN TARIFICATION AUTOMOBILE	45
I. Tarification selon le modèle linéaire généralisé.....	46
A. Notions théoriques du modèle linéaire généralisé	46
B. Modélisation de la fréquence et du coût moyen	52
TROISIEME PARTIE.....	67
TARIFICATION AUTOMOBILE ET RESEAU DE NEURONES	67
I. Tarification selon les réseaux de neurones	68

A.	Introduction aux réseaux de neurones	68
B.	Notions théoriques des réseaux de neurones.....	69
C.	Type d'architecture des réseaux de neurones	73
D.	Algorithme d'apprentissage	77
II.	Application à tarification automobile selon les réseaux de neurones	90
A.	Identification des données en entrées et en sortie	91
B.	Retraitement des données	91
C.	Modélisation de la prime pure à l'aide du perceptron à une couche cachée.....	92
D.	Comparaison entre le GLM et le réseau de neurones.....	94
E.	Modèle agrégé.....	96
III.	Limites et avantages de chaque modèle	98
	CONCLUSION GÉNÉRALE	100
	Bibliographies.....	101
	Tables des figures	102
	Liste des tableaux.....	104
	ANNEXES.....	105

INTRODUCTION GENERALE

Le climat de l'assurance, en particulier celui de l'assurance automobile, est marqué par des réformes réglementaires toutes influant sur le mode de gestion habituel des contrats. On pourrait citer entre autres, solvabilité II qui prône plus de prudence et une, bien meilleure, gestion du risque, la loi Hamon qui donne l'opportunité à tout assuré de résilier son contrat à tout moment. Celles-ci mettent les assureurs automobiles face à deux contraintes fortes, l'une qui est d'adosser le risque convenablement à chaque contrat et l'autre de faire face à la concurrence vis-à-vis des tarifs et de la gestion. Cela fait peser une incertitude importante sur les bénéfices des compagnies. De plus, l'assurance automobile en particulier est un marché fortement concurrentiel, on y trouve une diversité de contrat, de garanties annexes et de nombreuses offres autour de l'évolution du bonus-malus.

Dans ce contexte, la problématique du « juste prix » est plus que jamais au cœur des métiers actuariels. Il est nécessaire de posséder une tarification précise et adaptée à son portefeuille. D'une manière générale, la tarification en assurance auto, se base sur le développement des modèles linéaires généralisés. Bien que performants, ceux-ci imposent des contraintes fortes sur la structure du risque et sur les interactions entre les variables explicatives. Cela peut conduire à une estimation biaisée de la prime d'assurance.

Toutefois, de nouvelles méthodes offrent un cadre différent dans le domaine de la tarification. C'est le cas de l'apprentissage statistique qui par son caractère non-paramétrique n'est pas soumis aux contraintes citées plus hauts.

L'objet de ce mémoire serait donc d'établir un tarif de contrat d'assurance automobile avec en perspective la comparaison de la méthode du GLM avec une méthode d'apprentissage statistique (les réseaux de neurones). Pour ce faire, nous déterminerons les tarifs de trois manières différentes. La première réside en la méthode classique de GLM, la deuxième en utilisant l'algorithme des réseaux de neurones, la troisième en agrégeant les estimateurs issus de ces deux méthodes.

Dans la première partie de ce mémoire, nous décrirons d'abord, l'environnement de l'assurance automobile en France. Ensuite, nous présenterons l'entreprise, la direction et le service dans lequel ce mémoire a été établi. Enfin, nous procéderons à une étude exploratoire et de traitements des données.

Dans la deuxième partie, nous ferons un rappel succinct sur les théories mathématiques qui sous-tendent la méthode GLM, en insistant sur l'importance du choix des lois dans l'utilisation de cette méthode paramétrique. Nous l'appliquerons à notre base de données afin d'obtenir nos tarifs.

Dans la troisième partie, nous présenterons des algorithmes de réseaux de neurones, les concepts mathématiques et leur logique de construction. Nous les mettrons en œuvre sur notre portefeuille.

Enfin, nous construirons un estimateur agrégé des méthodes utilisées plus haut et comparerons par la suite, les tarifs obtenus par ces trois méthodes.

PREMIERE PARTIE

**CONTEXTE ECONOMIQUE ET ANALYSE
EXPLORATOIRE**

Cette partie se distingue en deux grands axes.

Le premier étant de comprendre les enjeux économiques de la tarification plus particulièrement ceux de l'entreprise dans laquelle ce mémoire a été réalisé.

Le deuxième point est axé autour de la problématique de la qualité de la donnée utilisée et le retraitement nécessaire à nos tarifications.

I. Le marché de l'assurance en France

A. Un marché concurrentiel

Avec un taux de croissance du PIB (0,4%) inférieur aux autres grands acteurs mondiaux, la France reste le 5ème acteur mondial en termes de volume de primes (200Md€ en 2014), juste devant l'Allemagne (188Md€ de primes). [Le marché de l'assurance en 2016 note de conjuncture GRAS SAVOYE].

Caractérisé par une concurrence importante et de nombreux canaux de distribution (bancassureurs, agents généraux, courtiers, salariés), le marché français de l'assurance continue d'assister à des fusions et des acquisitions à la fois en assurance de personnes, en assurance de biens et en réassurance. Le durcissement des règles de solvabilité, l'augmentation de la concurrence et la nécessité d'accroître la rentabilité contribuent à accentuer ces mouvements de consolidation.

Dans ce contexte, le marché de l'assurance française a progressé de 6% en 2014, porté principalement par les assurances de personnes (+7,6%).

	2013 en Md €	2014 en Md €	Variation
Cotisations (ensemble)	188,4	200	+6,1%
Assurances de personnes	138,3	148,9	+7,6%
Assurances de biens et responsabilités	50,1	51,1	+1,9%

Tableau 1 -Cotisations par type d'assurance

Source : FFSA

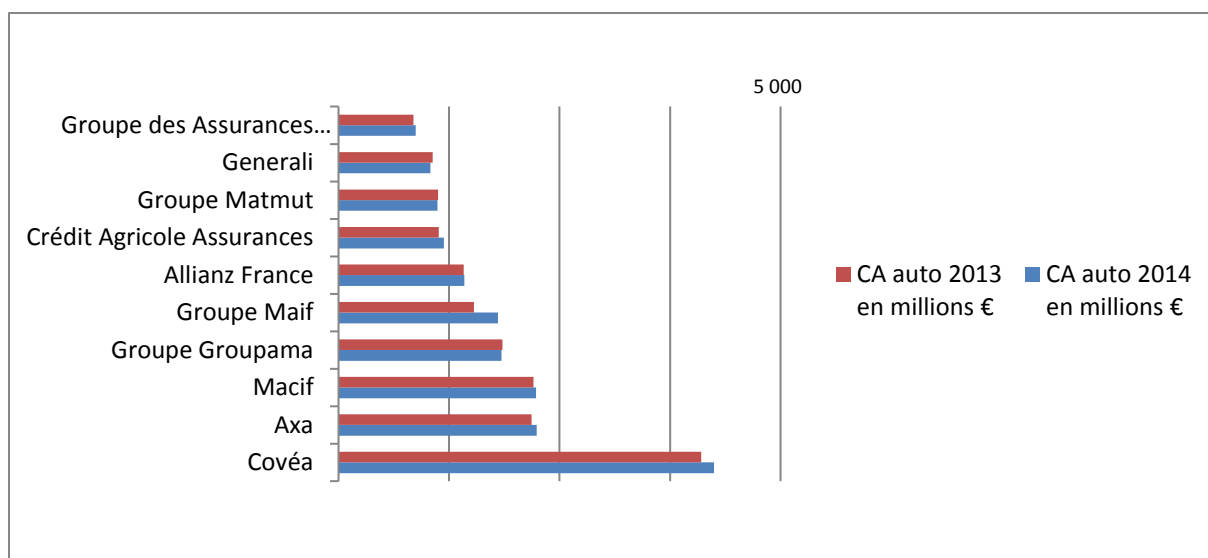


Figure 1- Chiffre d'affaires 2013 et 2014 des principaux acteurs de l'assurance Auto

Source : L'Argus de l'assurance

B. L'assurance automobile en France

Malgré cette progression générale de l'assurance en France la croissance des cotisations de l'assurance automobile, reste tout de même limitée en 2015 (+ 2,2 %). L'environnement du secteur de l'assurance auto est marqué par une sinistralité prononcée. L'année 2015 marque une inversion de tendance de tous les indicateurs de sécurité routière (hausse respectives de +1,7%, 3,7%, +2,5% et +2,9% pour les nombres d'accidents corporels, de tués, de blessés et des blessés hospitalisés. Bien que les conditions climatiques aient pesées dans cette évolution, le relâchement du comportement du conducteur est à prendre en compte.

Au cours de cette même année, le chiffre d'affaires de l'assurance automobile des particuliers représentait 31% des cotisations des assurances de dommages aux biens et de responsabilité civile et 80% de celui de l'assurance automobile. On estime à 40,2 millions le nombre de véhicules assurés dont 39,1 en métropole (véhicules de 1^{ère} catégorie hors flottes). Le marché de l'assurance automobile est un marché concurrentiel avec environ 100 acteurs opérant sur le marché.

Ce retournement se traduit par une hausse de la fréquence RC corporels (+ 2 %) dans un contexte de hausse forte et régulière du coût moyen de ces sinistres. A l'inverse la sinistralité des accidents matériels, bien moins coûteuse, est en léger recul. Étant donné cette mauvaise orientation de la sinistralité, les résultats techniques de la branche sont en dégradation.

C. Environnement législatif

Loi Hamon

Afin de protéger les assurés contre la reconduction tacite des contrats d'assurance, la loi prévoit des mesures facilitant la résiliation d'assurance auto, moto et habitation. Depuis le premier janvier 2015 les assurés peuvent procéder à la résiliation de leur contrat d'assurance à tout moment à partir de leur date anniversaire (un an).

L'impact de la loi Hamon sur les portefeuilles Auto selon l'AFA, est observable par l'évolution de 1,1 point entre 2014 et 2015 du taux de rotation (nombre de résiliations par rapport au nombre de contrat), passant ainsi de 13,8% à 14,9%. Cette loi apporte plus de fluidité sur le marché de l'assurance dommage et donc plus de concurrence. En effet, l'accélération du taux de rotation des portefeuilles alimente la guerre tarifaire à laquelle se livrent les poids lourds de l'assurance dommages en France. Une bataille des prix qui n'est certes pas nouvelle mais qui s'est accrue avec l'entrée en vigueur de la loi Hamon.

Solvabilité II

Les reformes solvabilité II ont un impact significatif dans le mode de gestion des contrats non-vie en particulier auto. Comme principe important, on peut citer la prééminence du fond sur la forme. Cela implique que :

- La segmentation doit refléter la nature des risques sous-jacents au contrat (le fond) plutôt que le versant juridique du contrat (la forme).
- Les engagements d'assurance d'une activité exercée sur une base technique non-vie sont considérés comme des engagements non-vie.
- Les rentes découlant des contrats non-vie sont des engagements vie.

II. Présentation de l'entreprise

A. GENERALI

Fondé il y a plus de 180 ans à Trieste, Generali s'installe dès 1832 en France, sa plus ancienne implantation étrangère. La filiale française acquiert au fil du temps diverses sociétés de l'Hexagone qui sont regroupées progressivement pour aboutir en décembre 2006 à la création de l'entreprise unique Generali France.

Le chiffre d'affaires de Generali France atteint 11,3 milliards d'euros en 2015. La Compagnie s'appuie sur 7600 collaborateurs pour offrir des solutions d'assurance à 7 millions de clients,

particuliers ou bénéficiaires de garanties dans le cadre de leur activité, ainsi que 800 000 entreprises et professionnels

Aujourd'hui, le Groupe Generali est l'un des principaux assureurs au monde. Son chiffre d'affaires en 2015 s'élève à 74 milliards d'euros. Avec 76 000 collaborateurs à travers le monde au service de 72 millions de clients dans plus de 60 pays, le Groupe figure parmi les leaders sur les marchés d'Europe occidentale, et occupe une place d'importance croissante en Europe centrale et orientale ainsi qu'en Asie.

B. La direction des partenariats

La direction des partenariats est un pôle unique sur le marché, entièrement consacré au développement des solutions sur mesure (produits, services et gestion) pour ses partenaires⁽¹⁾ en assurances dommages, en assurances de personnes ainsi que dans le domaine de la protection juridique. Elle s'occupe également de segments de clientèles peu accessibles aux réseaux traditionnels et intervient en complémentarité de ceux-ci.

Elle compte 3,5 millions de clients en France et son chiffre d'affaire est de 622 millions d'euros en 2011.

La direction des partenariats regroupe plusieurs pôles dont :

- La direction des Solutions d'Assurances Dommages à laquelle j'appartiens
- La direction Opérations d'Assurances et Protection Juridique
- La direction du Développement

Le terme partenaire désigne de manière générique un apporteur d'affaires ou un distributeur

C. L'ÉQUITÉ

L'Équité, filiale d'assurance dommages de Generali France depuis 1976, crée et propose, en marque blanche, des solutions d'assurances sur mesure qui tiennent compte des spécificités de ses partenaires. En 2011, elle fusionne avec l'EPJ (Européenne de Protection Juridique) et devient le spécialiste sur les segments d'assurance dommage de particuliers, d'assurance prévoyance accident et santé et sur la protection juridique.

Les partenaires de L'Équité sont divers, il s'agit de professionnels de la distribution d'assurance (courtiers d'assurance, courtiers grossistes, courtiers spécialisés), de mutuelles et institutions de prévoyance ou encore de grands comptes et enseignes agissant pour le compte de leurs clients.

Aujourd'hui, l'Équité représente environ 569 millions de chiffre d'affaires et couvre plus de 3 millions d'assurés.

Les responsables d'études actuarielles des Solutions d'Assurances Dommages ont pour missions principales :

Établir des statistiques partenaires (Stats S/P)

Réaliser des statistiques détaillées par segment

Participer à l'élaboration et à la maintenance des tarifs en liaison avec les Chargés de Comptes

De manière plus large, l'équipe des Partenariats Solutions Dommages est en charge de l'étude des nouveaux partenariats, de l'élaboration de solutions techniques et de la gestion du partenariat et travaille en étroite collaboration avec l'équipe Support technique produits, qui s'occupe de la maintenance des documents contractuels et qui aide à la mise en place des suivis de partenariats.

Notre étude porte sur la reconstruction tarifaire de la garantie responsabilité civile (RC) d'un courtier grossiste.

Le courtier grossiste est un apporteur qui présente la particularité de ne pas être en rapport direct avec le client. Ainsi, celui-ci s'adresse à un réseau d'intermédiaires d'assurance indépendants (courtiers, mandataire d'intermédiaire d'assurance ou agents) qui présentent les produits aux clients finaux. La particularité d'un tel partenariat entraîne une sélection de profil de risque n'ayant pas la possibilité de s'assurer via les réseaux de distribution directe.

III. Analyse des variables explicatives

L'analyse descriptive contient multiples enjeux. Elle permet de déterminer les caractéristiques d'un individu moyen afin de connaître la population assurée et de vérifier son adéquation avec le cœur de cible de l'entreprise. Aussi permet-elle de vérifier la pertinence des variables tout en étudiant de façon plus ou moins brève la dépendance entre les variables, notion primordiale lors de la modélisation et sur laquelle nous reviendrons. Toutefois afin d'avoir des données fiables à l'analyse, une étape préliminaire de construction d'une base adaptée au besoin de notre étude est indispensable.

A. Constitution de la base

Nous avons fait le choix de ne modéliser la prime pure liée à la garantie RC, car celle-ci est la mieux représentée dans notre base. Aussi tenons-nous à préciser que dans un souci de confidentialité

des données, celles-ci ont été translatées afin de ne livrer aucune information sur les coûts réels des sinistres. Nous disposons de deux bases de données : une base des contrats et une des sinistres. Ces bases concernent les contrats ayant été souscrits entre le 01/01/2011 et 31/12/2014 vu au 31/12/2015.

La base contrat est une base image qui associe à chaque contrat une situation du risque rattachée à une exposition. Toute modification ou évolution d'une variable de cette base conduit à la création d'une nouvelle image. On associe un numéro de contrat à chaque contrat qui sert d'identifiant unique pour chacun des assurés.

La base sinistre quant à elle, est constituée ligne par ligne de sinistres identifiés chacun par un numéro de sinistre et le numéro du contrat rattaché à ce sinistre. On dispose aussi de la date de survenance du sinistre afin de rapporter le sinistre à l'exercice qui lui ait attribué.

A l'aide du numéro de contrat et de la date de survenance du sinistre on crée une unique base de données contenant les contrats non sinistrés et ceux sinistrés, rattachés à leurs variables de sinistralité.

Dans le point suivant, nous étudierons les variables, ferons l'analyse descriptive des variables utilisées pour le tarif.

B. Variables d'étude

En pratique les assureurs s'intéressent à l'étude approfondie de l'ensemble des variables explicatives à leur disposition. Cela passe souvent par l'établissement d'un zonier et d'un véhiculier. Dans ce mémoire, Nous travaillerons sur un zonier et un véhiculier déjà existants.

Pour notre étude, nous disposons de la liste de variables explicatives suivantes :

- | | |
|-------------------------------|--|
| 1. Interruption d'assurance : | période de non couverture entre le dernier contrat d'assurance et la date de souscription, |
| 3,4. classe & groupe: | la classe et le groupe représentant les caractéristiques du véhicule assuré
Ceux-ci sont déterminés selon les normes SRA (Sécurité et réparation automobile), |
| 5. mode d'acquisition: | mode d'acquisition du véhicule, comptant, à crédit ou en leasing, |
| 6. zone: | détermine la dangerosité du lieu de stationnement du véhicule, |
| 7. usage: | type d'utilisation du véhicule, |
| 8. parking: | Indique le type de stationnement du véhicule, |
| 9. fractionnement: | période de paiement, |

10. mode de paiement:	mode de paiement de la prime, prélèvement automatique ou chèque,
11. periode sinistre:	Période de recule pour analyser les antécédents de sinistralité,
12. annee:	année d'exercice du contrat,
13. CRM:	coefficient de réduction majoration,
14. conduite accompagnée :	obtention de permis en conduite accompagnée,
15. age du conducteur	âge du conducteur principal,
16. ancienneté du permis:	ancienneté du permis en classe d'année,
17. nombre de conducteur:	nombre de conducteurs du véhicule,
18. antécédent d'assurance :	Nombre de mois pendant lequel le client a été assuré,
19. ancienneté du véhicule:	âge du véhicule,
20. antécédents&responsabilité :	Variable croisée entre le type de sinistre antérieur et la responsabilité

Il est nécessaire de préciser que certaines variables non correctement renseignées ou n'ayant aucune pertinence dans la tarification n'ont pas été prises en compte.

Nous disposons également de variables supplémentaires en lien avec la sinistralité du portefeuille. Il s'agit de :

1. cout du sinistre :	Variable qui désigne le montant de la charge sinistre,
2. nombre de sinistre :	Variable désignant le nombre de sinistres obtenus.
3. nature du sinistre :	Variable qui détermine la nature du sinistre
4. responsabilité sinistre :	Variable qui détermine le niveau de responsabilité de l'assuré dans la survenance d'un sinistre, nulle, partielle ou totale

Ces variables explicatives sont déterminantes dans la tarification étant donnés que leurs modalités définissent les profils de risques. Il faut donc s'intéresser à leur comportement vis-à-vis de la sinistralité, de la fréquence et du coût moyen.

1. Analyse graphique des variables explicatives

Avant toute modélisation, l'analyse des variables étudiées est injonctive. Celle-ci permet de comprendre et d'apprécier l'objectif de la modélisation.

Cette section requiert un distinguo entre variables continues et variables à modalités. D'une part, on observe les variables continues afin d'analyser leur impact sur la sinistralité en vue d'une discrétisation. D'autre part, on étudie les variables à modalités, qui nécessitent des regroupements de modalités selon l'importance ou non de leur coût moyen et de leur fréquence.

2. Analyse graphique des variables à modalités

Analysons à présent à l'aide de quelques graphiques, le coût moyen et la fréquence par variables explicatives afin de comprendre la structure de notre portefeuille.

On définit le coût moyen comme le rapport de la somme des coûts lié à la garantie RC et le nombre de sinistre qui s'y attache :

$$\text{coût moyen} = \text{somme}(\text{coût RC}) / \text{nombre de sinistre}$$

La fréquence quant à elle se définit comme le nombre de sinistres rapporté à l'année d'exposition :

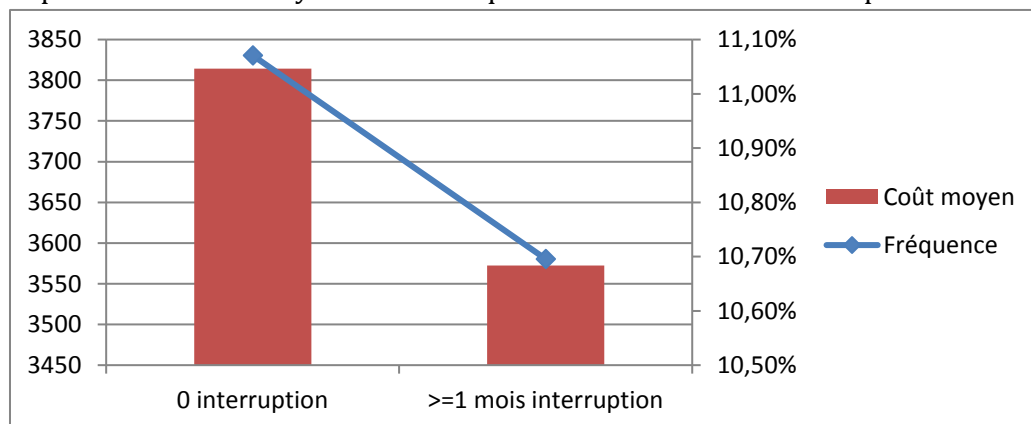
$$\text{fréquence} = \text{nombre de sinistre} / \text{Année d'exposition}$$

On note dans cette partie que les chiffres utilisés ont été translaté dans un souci de confidentialité des données.

Interruption d'assurance

Une variable désignant le nombre de mois durant lesquels l'assuré n'était plus couvert par une assurance. Cette variable dispose de deux modalités.

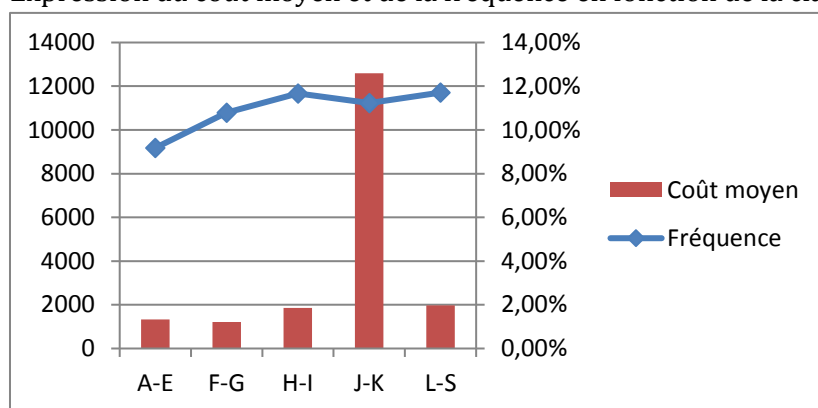
Expression du coût moyen et de la fréquence en fonction de l'interruption d'assurance



Classe

La variable classe correspond à une classification selon la valeur à neuf du véhicule assuré. Plus la lettre de la classe se rapproche de Z, plus la valeur à neuf du véhicule est importante. On constate ici un pic du coût moyen des sinistres pour les classes de J à K mais avec une fréquence plus faible que celle des groupes de classes "H-I" et "L-S". Toutefois, la fréquence des sinistres évolue plutôt linéairement.

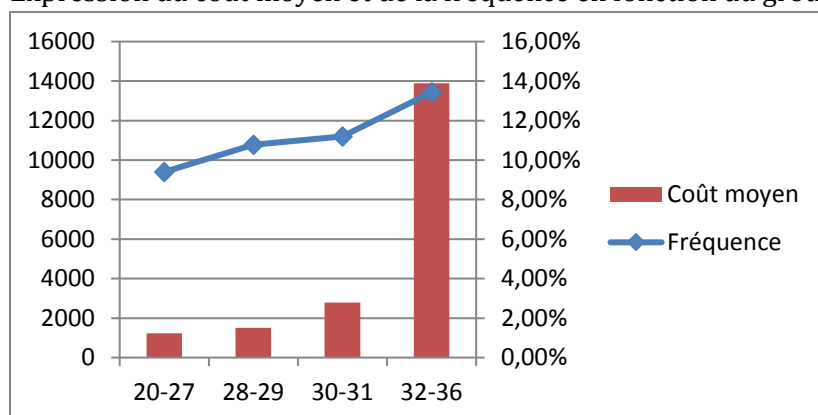
Expression du coût moyen et de la fréquence en fonction de la classe



Groupe

La variable groupe correspond à une classification en fonction des caractéristiques du véhicule telles que la puissance, la vitesse. Nous possédons des groupes allant 20 à 36. Plus le véhicule a de caractéristiques importantes, plus son groupe est élevé.

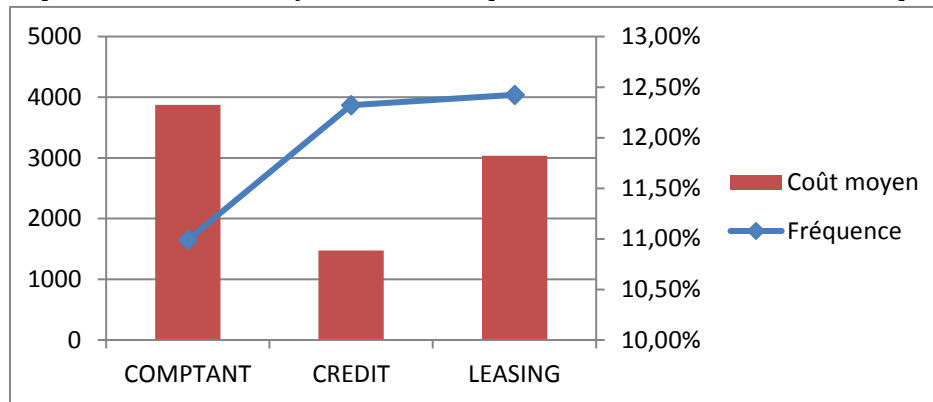
Expression du coût moyen et de la fréquence en fonction du groupe



Mode d'acquisition

Le mode d'acquisition est une variable qui détermine le mode d'achat du véhicule. Elle prend les modalités comptant, credit, leasing. La modalité comptant possède la fréquence la plus basse mais le coût moyen le plus élevé.

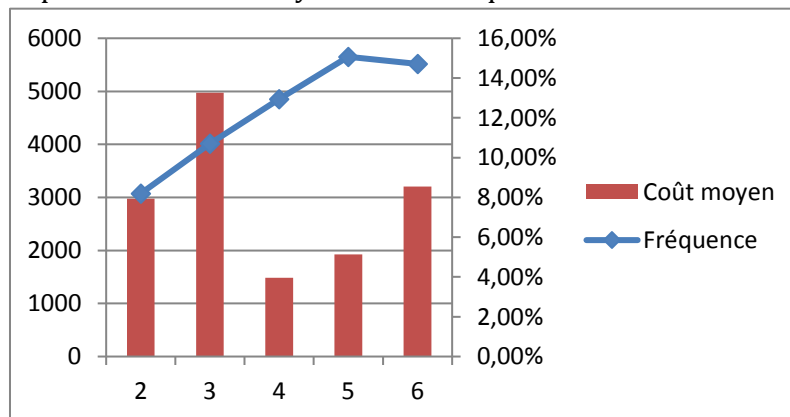
Expression du coût moyen et de la fréquence en fonction du mode d'acquisition



Zone

La zone est une variable qui détermine la dangerosité du lieu de stationnement. Les modalités iront de 2 à 6. Plus la modalité est grande, plus la zone est supposée dangereuse. Ceci est en adéquation avec la représentation de la fréquence. A l'inverse, les coûts n'ont pas de tendance apparente et difficilement interprétable.

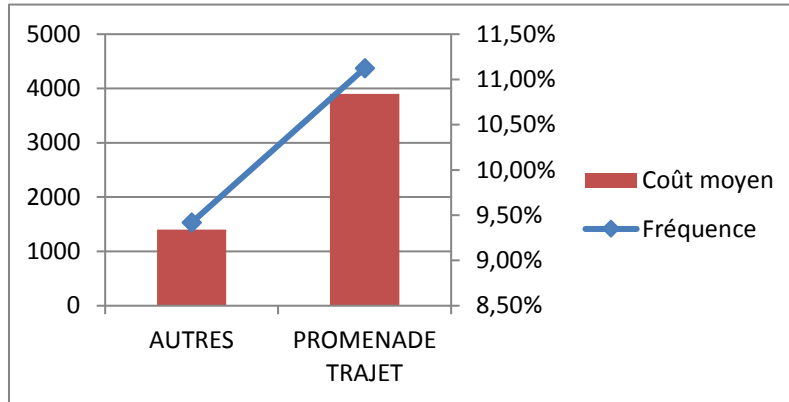
Expression du coût moyen et de la fréquence en fonction de la zone



Usage

La variable usage correspond aux types d'utilisation du véhicule. Cette variable a été retraitée sous deux modalités la modalité "promenade trajet" et "autre", qui est le regroupement de toutes les autres modalités représentant moins de 10% en terme d'effectifs.

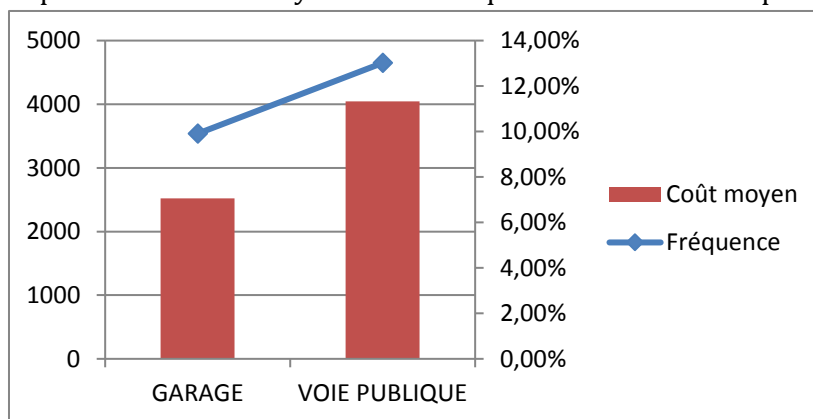
Expression du coût moyen et de la fréquence en fonction de l'usage



Parking

Cette variable indique le mode de stationnement du véhicule, dans un garage ou sur une voie publique.

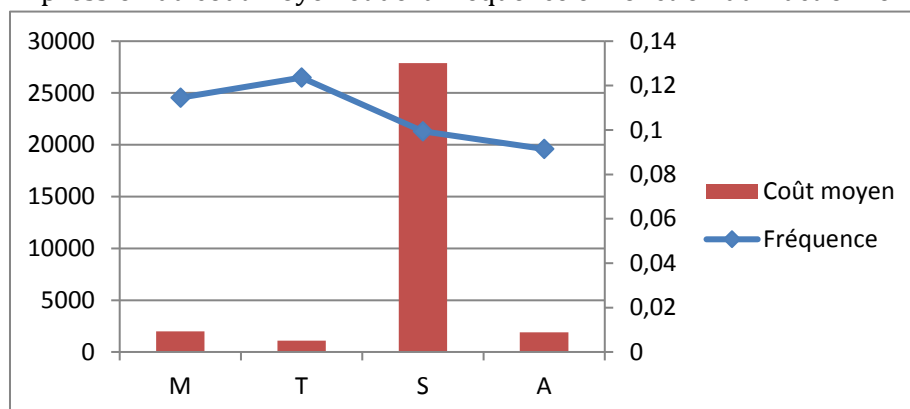
Expression du coût moyen et de la fréquence en fonction du parking



Fractionnement

Le Fractionnement désigne la cadence de paiement des primes d'assurance par l'assuré. Celui-ci peut s'en acquitter mensuellement (M), trimestriellement (T), semestriellement (S) et annuellement (A). L'analyse est biaisée par un sinistre extrême représentant la moitié du coût total et possédant la modalité S.

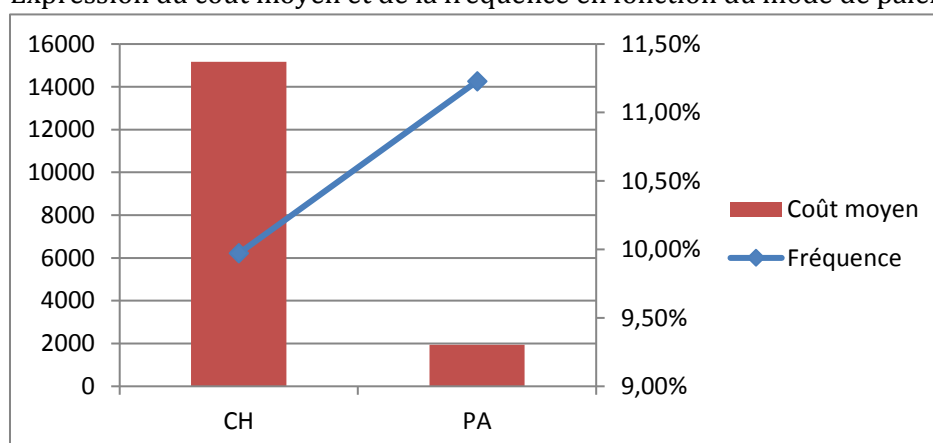
Expression du coût moyen et de la fréquence en fonction du fractionnement



Mode de paiement

Le mode de paiement correspond au canal de paiement utilisé par l'assuré afin de régler ses primes. Il peut être sous forme de chèque (CH) ou de prélèvement automatique (PA).

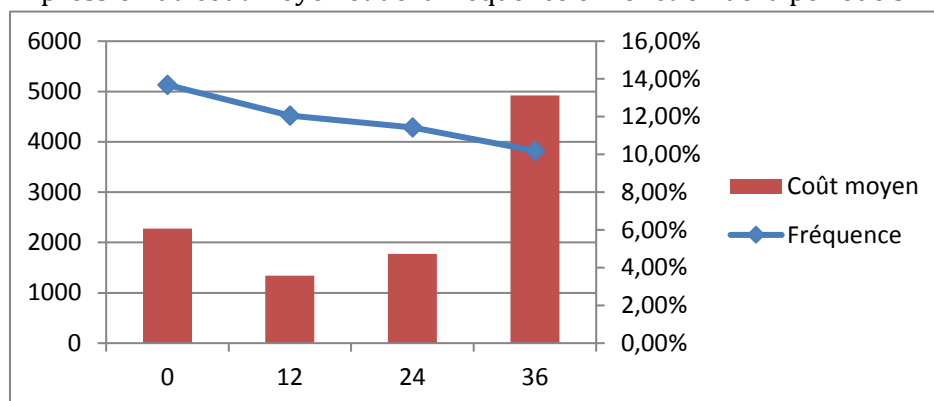
Expression du coût moyen et de la fréquence en fonction du mode de paiement



Période sinistre

Cette variable correspond au recule, en nombre de mois, que l'on se donne pour analyser les antécédents de sinistralité de l'assuré. Elle peut correspondre à 0, 12, 24 et 36.

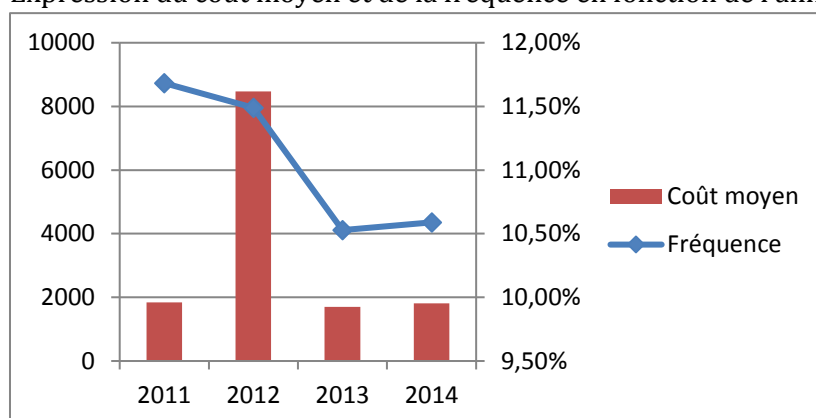
Expression du coût moyen et de la fréquence en fonction de la période sinistre



Année

La variable Année correspond à l'exercice auquel les sinistres et les contrats sont rattachés. On peut voir que 2012 a été une année plutôt difficile tant la fréquence et les coûts qui y sont rapportés sont importants.

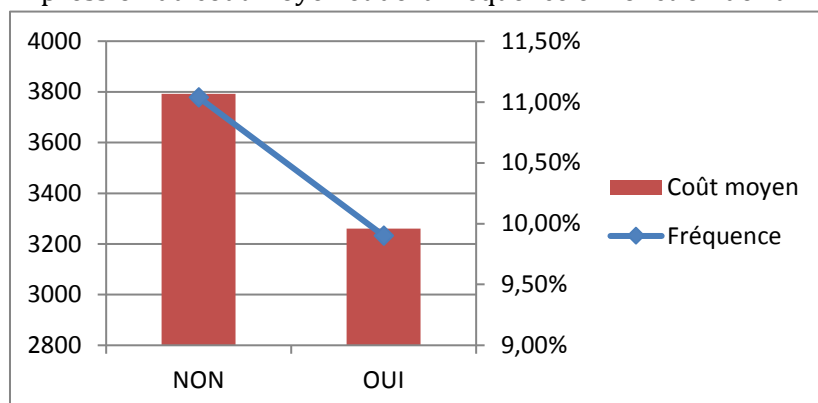
Expression du coût moyen et de la fréquence en fonction de l'année



Conduite accompagnée

Une variable binaire indique si oui ou non, le permis du conducteur principal a été obtenu en conduite accompagnée. Notre portefeuille est constitué en grande partie d'assurés n'ayant pas obtenu leur permis en conduite accompagnée.

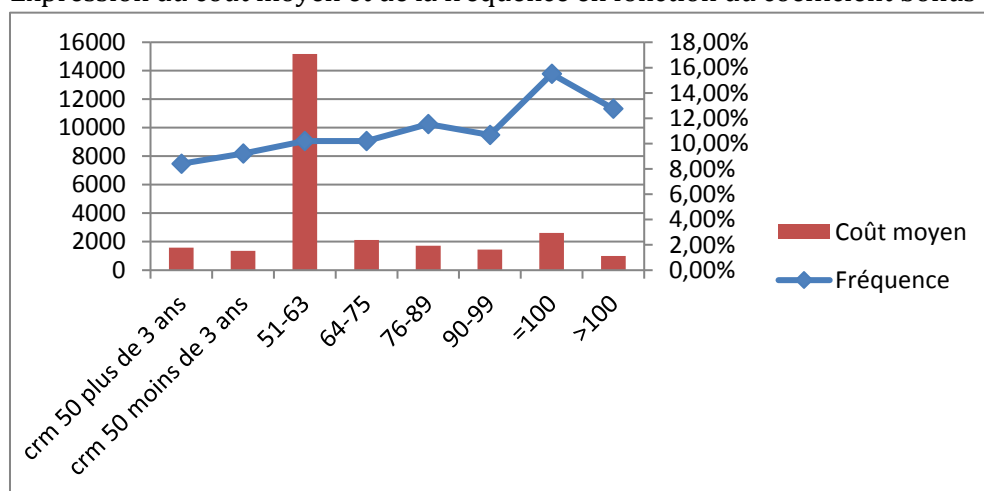
Expression du coût moyen et de la fréquence en fonction de l'année



CRM

Le coefficient de réduction majoration aussi appelé coefficient bonus-malus est un mécanisme de pondération de l'appréciation du risque par la sinistralité. Le coût moyen possède un pic pour les groupes "51-63" lié à une accumulation de sinistres graves durant l'année 2012.

Expression du coût moyen et de la fréquence en fonction du coefficient bonus-malus



Antécédents du sinistre & type de responsabilité

Une variable croisée entre le type de sinistres antérieurs (36 derniers mois) et le niveau de responsabilité de l'assuré. Cette variable est éclatée en quatre modalités :

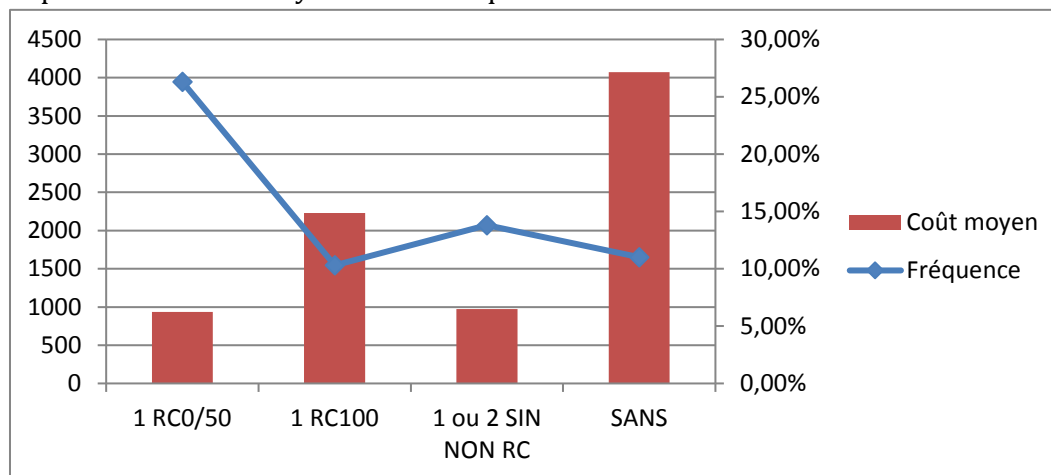
1RC0/50 : l'assuré ayant eu un sinistre dont la responsabilité ne lui ait pas incombée (0) ou la lui ait partiellement (50) mettant en jeu la garantie responsabilité civile (RC).

1RC100 : l'assuré ayant eu un sinistre dont il a la totale responsabilité (100) mettant en jeu la garantie RC.

1 ou 2 SIN NON RC : l'assuré ayant un ou deux sinistres mettant en jeu d'autres garanties.

SANS : l'assurant n'ayant pas eu de sinistres.

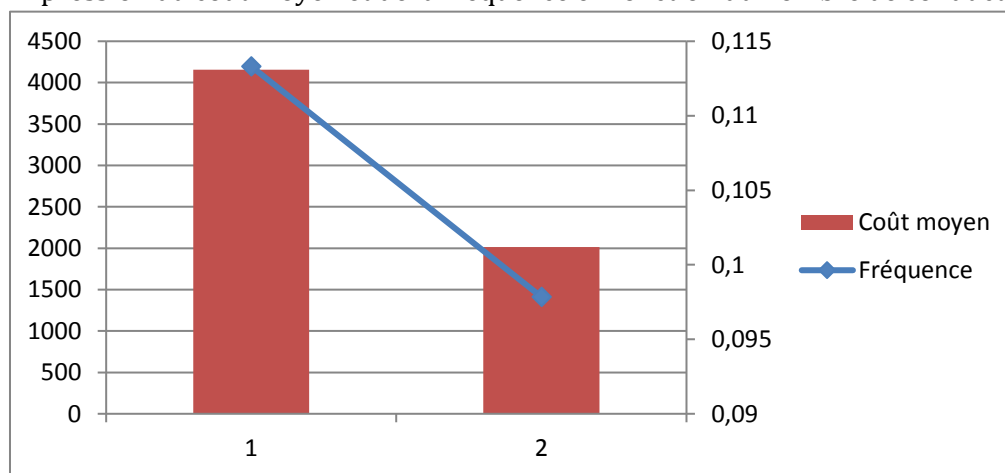
Expression du coût moyen et de la fréquence en fonction des antécédents de sinistralité



Nombre de conducteur

La variable nombre de conducteur désigne le nombre de personne habilité à conduire le véhicule. La fréquence et le coût moyen des assurés ayant déclarés un unique conducteur sont plus élevés que ceux ayant déclarés deux conducteurs.

Expression du coût moyen et de la fréquence en fonction du nombre de conducteur



C. Analyse graphique des variables continues

Nous analysons nos trois variables continues à savoir : l'âge du conducteur, l'ancienneté du permis et l'ancienneté du véhicule. Nous le faisons en juxtaposant la distribution de ces variables sur les contrats ayant déjà été sinistrés (OUI) et ceux qui ne l'ont jamais été (NON).

Ces graphiques permettent d'expliquer aussitôt des liaisons entre les variables continues et la sinistralité, ainsi que d'éventuelles anomalies, par exemple des valeurs anormalement grandes ou des pics inattendus.

Age du conducteur

Cette variable est déterminée en prenant l'âge du conducteur principal à la date du début de contrat.

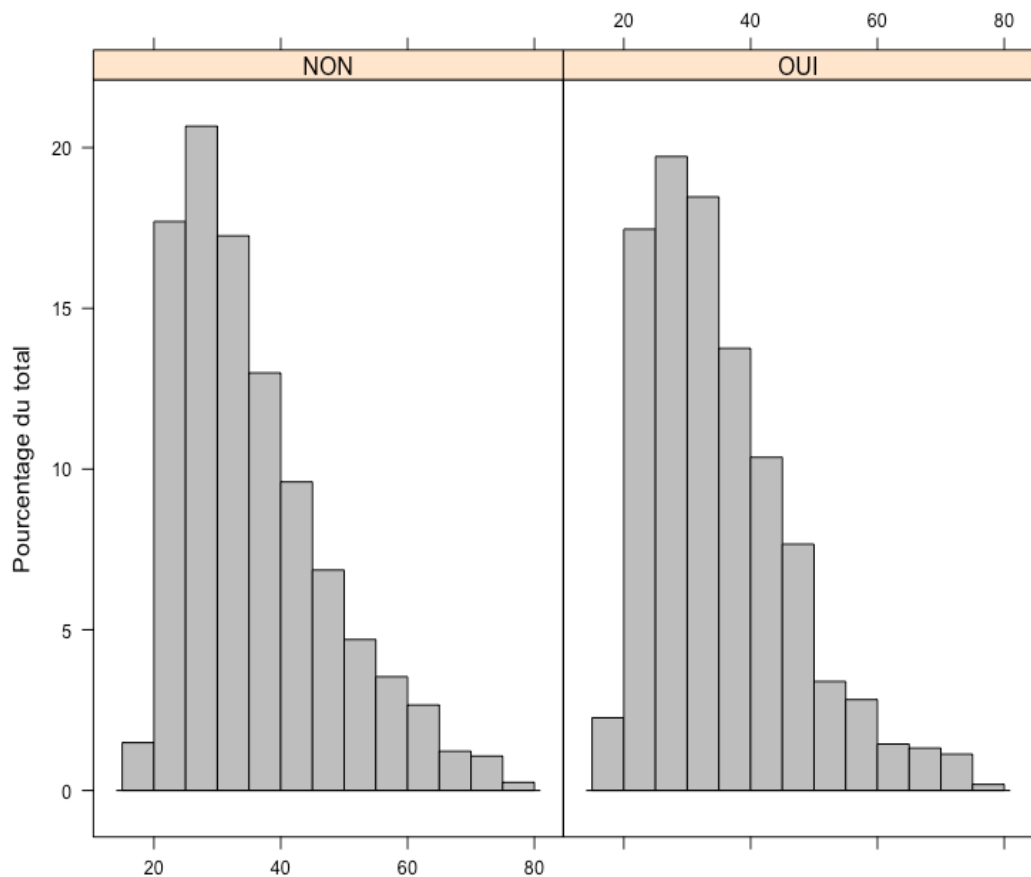


Figure 2 - Histogramme de l'âge du conducteur

On remarque un pic inattendu en [25-30[chez les assurés non sinistrés, qui sans doute lié à la forte propension de jeunes dans le portefeuille. Toutefois, l'accumulation de moins de 35 ans est plus importante pour les sinistrés que pour les non sinistrés. Cependant, après 35 ans la distribution est indifférente à la sinistralité.

Ancienneté du véhicule

L'ancienneté du véhicule est la différence entre la date de mise en circulation du véhicule et la date d'effet du contrat.

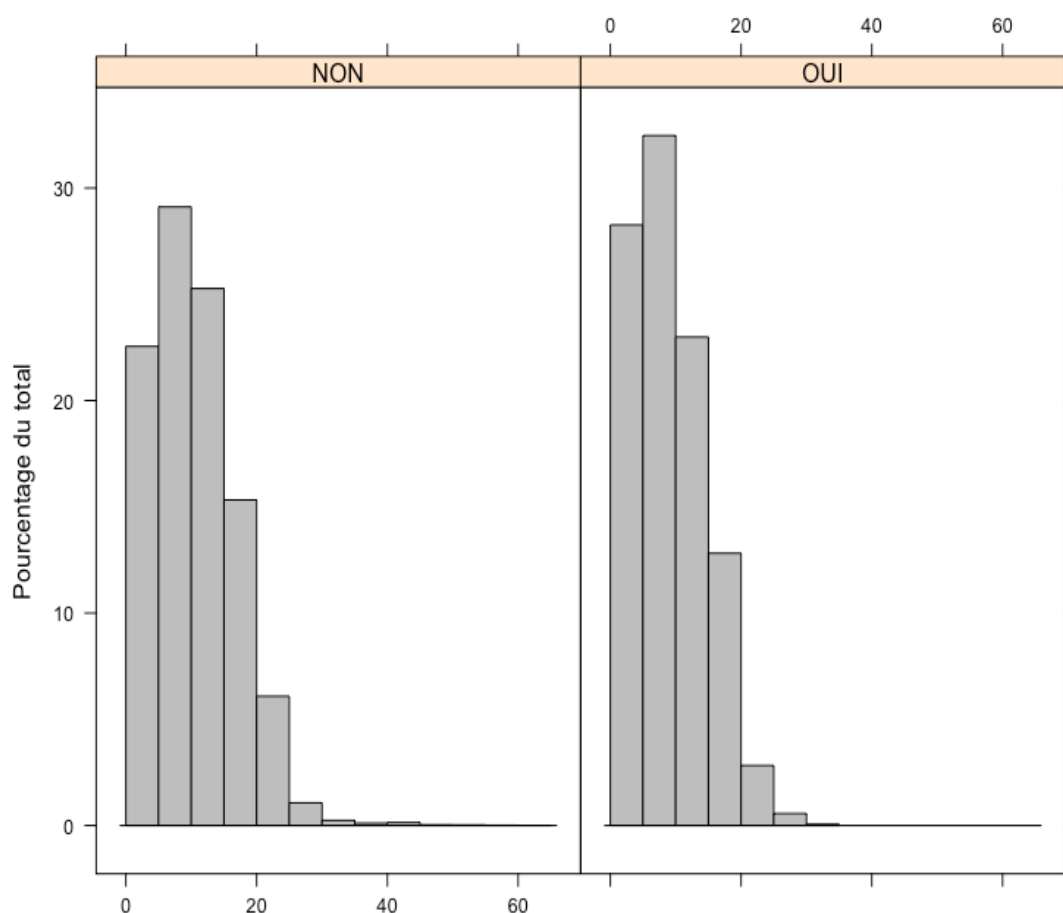


Figure 3- Histogramme de l'ancienneté du véhicule

L'ancienneté du véhicule est comprise entre 0 et 61 ans. On constate une proportion d'ancienneté moins élevée parmi les contrats sinistrés. Les véhicules ayant passés 15 ans d'ancienneté semblent être de moins en moins sujets aux sinistres.

Ancienneté du permis

La variable ancienneté du permis permet représente la différence entre la date du permis de conduire et la date à laquelle le contrat a pris effet. Notre portefeuille contient des assurés ayant une ancienneté de permis comprise entre 0 et 60 années. Les deux distributions conditionnelles de l'ancienneté du permis sont décroissantes. En revanche, on constate un pic entre 0 et 5 ans d'ancienneté de permis pour les assurés sinistrés. De plus, après 10 ans d'ancienneté du permis, la distribution semble ne plus dépendre du conditionnement avec un décroissement régulier.

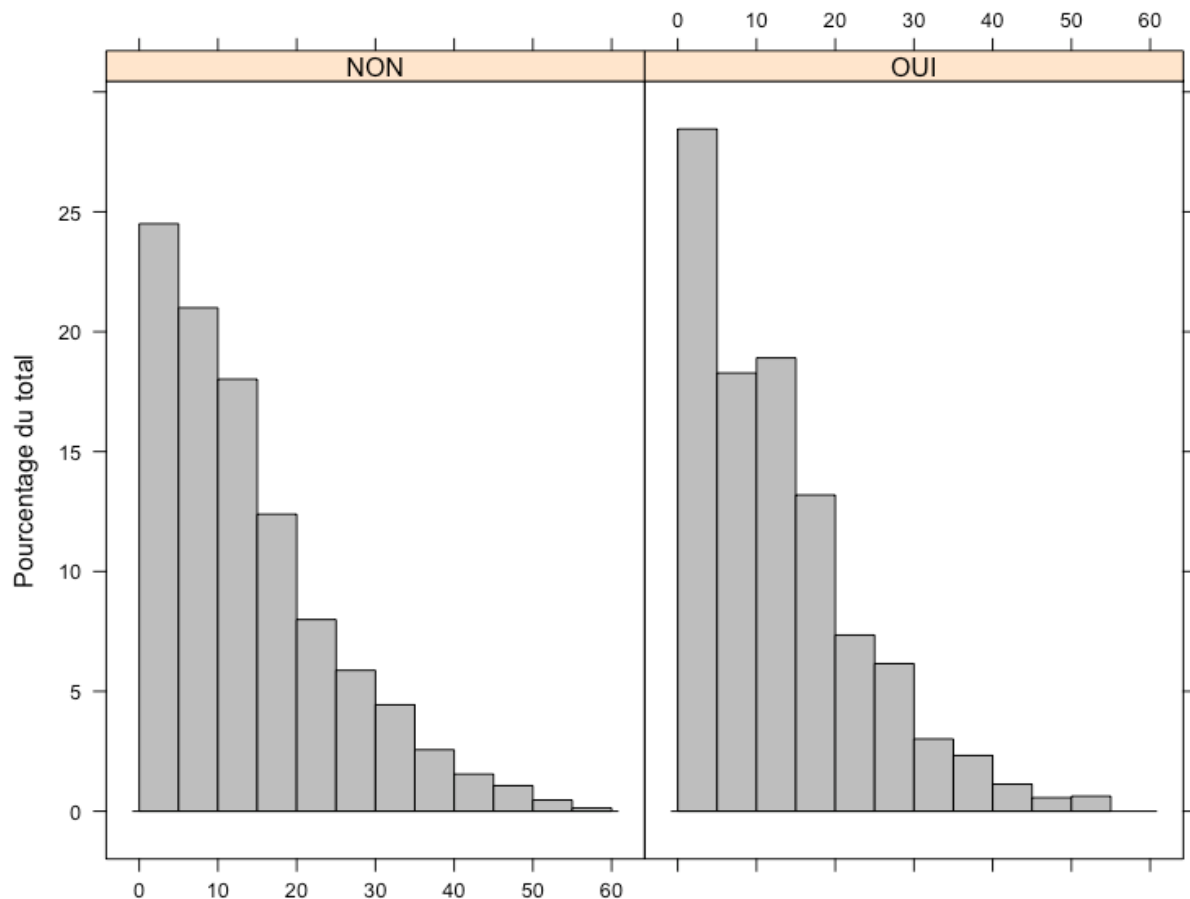


Figure 4 - Histogramme de l'ancienneté du permis

La modélisation linéaire nécessite une uniformisation du type des variables et nous oblige à rendre discret nos variables continues.

D. Discrétisation des variables continues

La discrétisation de variables continues consiste à découper une variable quantitative en intervalles. Il s'agit d'une opération de recodage. De quantitative, la variable est transformée en qualitative ordinale.

Plusieurs méthodes sont possibles, des plus usuelles au plus sophistiquées. Nous présenterons deux approches, qui nous conduiront à une discrétisation à peu près similaire.

E. Discrétisation non supervisée

Cette approche consiste à s'appuyer sur les caractéristiques intrinsèques de la variable pour en déduire un découpage pertinent.

Les caractéristiques disponibles pour toute statistique descriptives sont : le Min, le Max, la Moyenne, les quantiles, la variance, l'écart-type.

L'objectif est donc de se servir de ceux-ci pour établir le découpage en classe.

Les méthodes les plus connues (discrétisation en intervalles de largeur égales, en intervalles de fréquence égales, algorithme de Fisher) requièrent le nombre d'intervalles en paramètres. Leur véritable avantage est leur popularité, ce sont des approches « passe-partout », que l'on peut mettre en œuvre dans tout contexte.

Leur principal inconvénient est d'ignorer l'information en provenance de la variable cible.

C'est pourquoi, nous croiserons cette méthode au taux de sinistralité, ce qui nous permettra d'avoir une pertinence dans notre démarche.

On commence par déterminer le nombre K d'intervalles initiaux pour notre étude. Il existe des formules adaptées aux techniques non supervisées permettant d'estimer le nombre d'intervalles.

Appellation	K
Brooks-Carruthers	$5 \times \log_{10}(n)$
Huntsberger	$1 + 3,332 \times \log_{10}(n)$
Sturges	$\log_2(n + 1)$
Scott	$\frac{b - a}{3,5 \times \sigma \times n^{(-\frac{1}{3})}}$
Freedman-Diaconis	$\frac{b - a}{3,5 \times q \times n^{(-\frac{1}{3})}}$

Tableau 2- Nombre d'intervalle en fonction du nombre d'observations

Où n représente le nombre d'observations, a le minimum, b le maximum, σ l'écart-type et q l'intervalle interquartile.

On calcule les K-tiles de la variable X, puis on découpe X en K-tiles et enfin on calcule le taux de sinistres par intervalle. On regarde les points d'inflexions entre les K-tiles puis on regroupe les intervalles de sortes à avoir une liaison linéaire au niveau du taux de sinistres.

Le trait horizontal sur les trois graphiques, représente le taux moyen de sinistres par contrat en portefeuille. Ici l'on souhaite répartir les classes de sorte à ce que l'on ait une liaison linéaire du taux de sinistres. Les classes sont réparties pour créer une liaison plus linéaire entre les modalités et le taux de sinistres, ainsi qu'en ayant le moins de classes possibles au-dessus du taux moyen de sinistres.

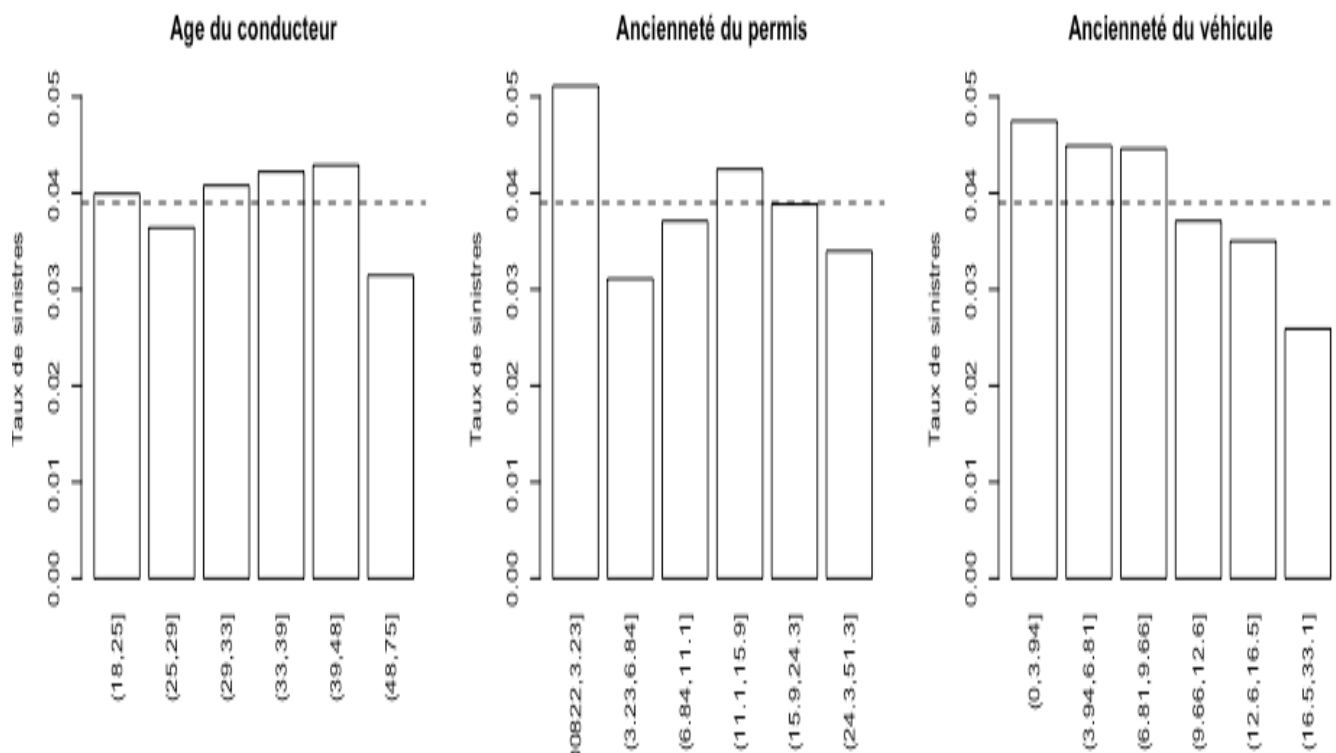


Figure 5 – Taux de sinistres en fonction du découpage des variables

Nous faisons le choix dans un premier temps de conserver les résultats bruts issus du découpage. Nous affinerons les intervalles et ne conserverons que les entiers aux bornes.

Age du conducteur	Au-dessus du taux moyen de sinistres	Ancienneté du permis	Au-dessus du taux moyen de sinistres	Ancienneté du véhicule	Au-dessus du taux moyen de sinistres
[18-25]	OUI	[0-3,23]	OUI	[0-3,94]	OUI
[25-29]	NON	[3,23-6,84]	NON	[3,94-6,81]	OUI
[29-33]	OUI	[6,84-11,1]	NON	[6,81-9,66]	OUI
[33-39]	OUI	[11,1-15,9]	OUI	[9,66-12,6]	NON
[39-48]	OUI	[15,9-24,3]	NON	[12,6-16,5]	NON
>48	NON	>24,3	NON	>16,5	NON

Tableau 3 - Découpage des variables continues

Pour la variable Age du conducteur, nous constatons deux points d'inflexion en [25-29] et [39-48]. De plus, les classes [18-25], [29-33], [33-39], et [39-48] ont des taux de sinistres supérieurs au taux moyen. Nous découpons la variable Age du conducteur en trois modalités en affinant les bornes.

La variable Ancienneté de permis a elle aussi deux points d'inflexion, [3,23-6,84] et [11,1-15,9]. Les classes [0-3,23] et [11,1-15,9] ont des taux de sinistres supérieurs au taux moyen. Nous

reconstruisons les modalités en ne gardant que trois modalités. Pour des raisons opérationnelles, nous utilisons des bornes entières.

En revanche, la variable Ancienneté du véhicule est linéaire par rapport au taux de sinistres. Toutefois, elle dispose d'un nombre important de classes au-dessus du taux moyen de sinistres.

Après regroupement de certaines des classes nous obtenons les graphiques suivants pour chacune de nos variables.

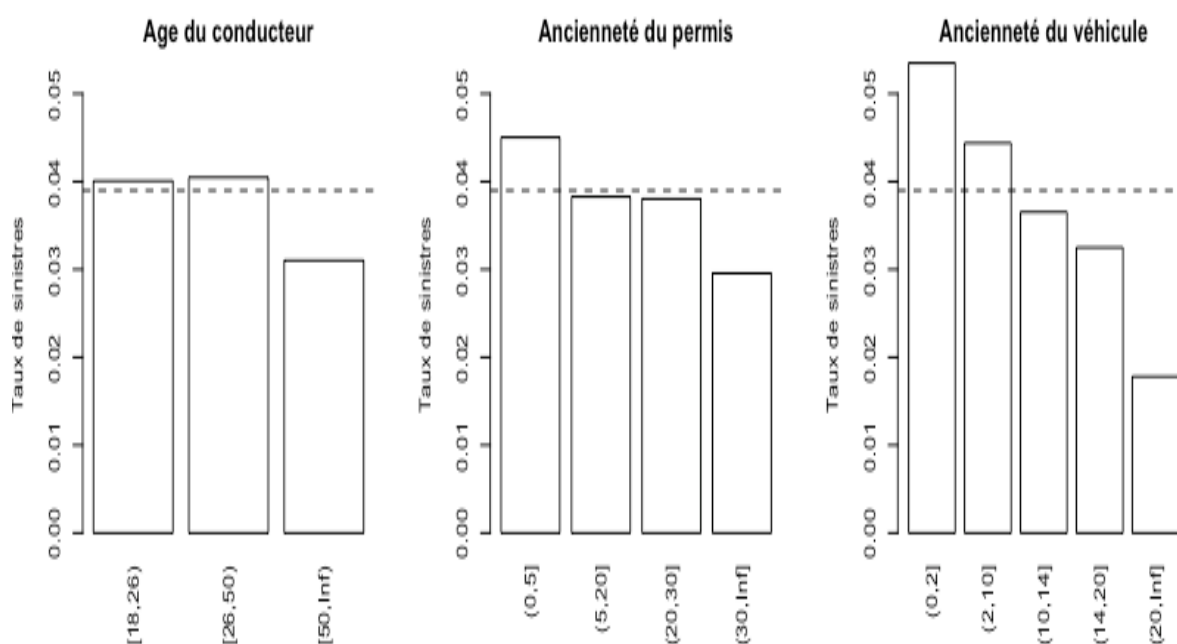


Figure 6 – Taux de sinistres en fonction du découpage des variables

Nous récapitulons les modalités obtenues pour chaque variable dans les tableaux suivants :

Variables	Discrétisation				
Age du conducteur	[18-26[[26-50[>=50		
Ancienneté du permis	[0-5]	[5-20]	[20-30]	>30	
Ancienneté du véhicule	[0-2]	[2-10]	[10-14]	[14-20]	>20

Tableau 4-Discrétisations retenues

Cette technique nous permet d'obtenir une discrétisation assez subjective, dans la mesure où certains regroupements sont préférés à d'autres de manière arbitraire. Pour pallier ce manque d'objectivité, nous introduisons une méthode dite de discrétisation supervisée.

F. Discrétisation automatique supervisée

Nous proposons maintenant une méthode permettant d'obtenir automatiquement une discrétisation supervisée des variables.

Cette méthode requiert de fixer le nombre K de classes souhaitées. Nous allons considérer tout au plus 6 classes pour chaque variables et faire des tests en faisant varier le nombre de classes souhaité de 2 à 6. L'idée est de rechercher le découpage en K classes, maximisant l'aire sous la courbe ROC d'un modèle logistique ajusté sur la variable ainsi discrétisée.

Cherchant à discrétiser la variable X par rapport à une variable à expliquer Y , l'algorithme se décrit comme suit :

1. Fixer le nombre de classes K .
2. Calculer les K -tiles.
3. Ajuster un modèle logit $Y \sim X_d$, X_d étant la variable catégorielle dont les facteurs sont les classes déterminées à l'étape précédente : $] Min; q_1],] q_1; q_2], \dots,] q_k; +\infty[$.
4. Mesurer l'aire sous la courbe ROC du modèle logit.
5. Modifier les seuils q_i des classes et revenir à l'étape 3, de façon à obtenir une variable X_d modifiée et un modèle logit correspondant dont l'aire sous la courbe ROC soit plus élevée que l'aire précédente.
6. S'interrompre lorsque l'aire sous la courbe ROC n'augmente plus significativement.

1. Principe des courbes ROC

Les courbes ROC (Receiver Operating Characteristic) ont fait leur apparition lors de la deuxième guerre mondiale pour une mise au point de moyens efficaces de détection d'avions Japonais. Cette méthode a été ensuite appliquée à la détection de signal puis en médecine.

La problématique est la suivante : On étudie un phénomène souvent de nature binaire, dans, notre cas, présence ou non de sinistre. On souhaite mettre en place un test, permettant de détecter la survenance d'un événement (la présence d'un sinistre).

Soit S une variable binaire ou multinomiale décrivant le phénomène pour N individus.

On appelle positifs, les individus pour lesquels ces événements se produisent et négatifs ceux pour lesquels il ne se produit pas. Soit T un test dont le but est de déterminer si le phénomène se produit ou non.

Nous reprenons les notations et les définitions du site xlstat [15].

On souhaite définir la valeur seuil au-delà/ en-deca de laquelle l'événement se produit. Pour ce faire, on étudie un certain nombre de valeurs seuils possibles. Les plus simples sont :

- Vrais positifs (VP): nombre d'individus déclarés positifs par le test et qui le sont réellement.
- Faux positifs (FP) : nombre d'individus déclarés positifs par le test mais qui ne le sont pas en réalité.

- Vrais négatifs (VN): nombre d'individus déclarés négatifs par le test et qui le sont effectivement.
- Faux négatifs (FN): nombre d'individus déclarés négatifs par le test mais qui sont en réalité positifs.
- Prévalence de l'événement : fréquence de survenance de l'événement dans l'échantillon total $(VP+FN)/N$

Il existe des indices synthétiques mis en place pour évaluer la performance du test à une valeur seuil donnée :

- Sensibilité (aussi appelé fraction de vrais positifs) : Proportion d'individus positifs effectivement bien détectés par le test. La sensibilité permet de mesurer la force du test lorsqu'il est appliqué sur ces individus-là. Une sensibilité à 1 indique que le test est parfait pour les vrais positifs. Une sensibilité valant 0,5 signifie que le test est équivalent à un tirage au hasard. Une sensibilité inférieure à 0,5, signifie que le test est contre-performant et on aurait intérêt à inverser la règle pour qu'il soit supérieur à 0,5. $\text{Sensibilité} = VP / (VP + FN)$.
- Spécificité (aussi appelée fraction des vrais négatifs) : proportion des vrais négatifs effectivement bien détectés par le test. Autrement dit, la spécificité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus négatifs. Le test est parfait pour des individus négatifs si la spécificité est égale à 1, est un tirage au hasard si elle vaut 0,5 et est contre-performant sur des négatifs si elle est inférieure à 0,5. Pour un test contre-performant sur les négatifs, il est judicieux d'inverser la règle pour que la spécificité soit supérieure à 0,5.
- Fraction des faux positifs (FFP) : proportion des négatifs détectés comme des positifs par le test. $FFP = 1 - \text{spécificité}$.
- Fraction des faux négatifs (FFN) : proportion des positifs détectés comme des négatifs par le test. $FFN = 1 - \text{sensibilité}$.
- Valeur prédictive positive (VPP) : C'est la proportion des cas effectivement positifs parmi tous les positifs détectés par le test. $VPP = VP / (VP + FP)$ ou $VPP = \text{sensibilité} \times \text{prévalence} / [\text{sensibilité} \times \text{prévalence} + (1 - \text{spécificité})(1 - \text{prévalence})]$. Cette valeur dépend de la prévalence et est indépendante de la qualité du test.
- Valeur prédictive négative (VPN) : C'est la proportion des cas effectivement négatifs parmi tous les négatifs détectés par le test. $VPN = VN / (VN + FN)$ ou $VPN = \text{spécificité}(1 - \text{prévalence}) / [\text{spécificité}(1 - \text{prévalence}) + (1 - \text{sensibilité})\text{Prévalence}]$. Tout comme la VPP, la VPN dépend de la prévalence et est indépendante de la qualité du test.
- Rapport de vraisemblance positif (LR+) : Ce rapport indique à quel point un individu a de chance d'être positif en réalité, si le test est positif. On a : $LR+ = \text{sensibilité} / (1 - \text{spécificité})$.
- Rapport de vraisemblance négatif (LR-) : Ce rapport indique pour un individu la chance qu'il a d'être en réalité positif si le test le détecte négatif. Ce rapport est aussi appelé risque relatif. On a : $LR- = (1 - \text{spécificité}) / \text{sensibilité}$.

Quid, de la courbe ROC ?

La courbe ROC correspond à la représentation du couple (1-spécificité ; sensibilité) pour les différentes valeurs seuils. Son allure est soit en escalier, soit en droites par morceaux.

L'aire sous la courbe (ou Area Under the Curve-AUC) est un indice synthétique utilisé pour les courbes ROC. L'AUC correspond à la probabilité qu'un individu positif soit détecté comme tel par le test sur l'étendue des valeurs seuils possibles. Pour un modèle idéal $AUC=1$; pour un modèle aléatoire $AUC=0,5$. On considère habituellement qu'un modèle est bon dès lors où l'on a AUC supérieure à 0,7. Un modèle bien discriminant doit avoir une AUC comprise entre 0,8 et 0,9. Un modèle qui a une AUC supérieure à 0,9 est dit excellent.

Un mot sur l'homogénéité des effectifs des classes

Pour privilégier les effectifs homogènes, nous utiliserons l'indice de Gini¹ $1 - \sum_i f_i^2$, où f_i est la proportion de la population dans la $i^{ème}$ classe. Cette quantité est d'autant plus grande que les fréquences f_i sont proches. Elle vaut d'ailleurs 0 si l'une des classes contient l'ensemble de la population et $1-(1/m)$ si les m classes ont le même effectif.

Nous multiplions l'AUC par le ratio :

$$\frac{1 - \sum_i f_i^2}{1 - (1 - h) \sum_i f_i^2}$$

De sorte à ce que l'effet soit nul si $h=0$ (pas de correction d'homogénéité) et que l'AUC est multipliée par $1 - \sum_i f_i^2$ si $h=1$. Si les classes sont hétérogènes $1 - \sum_i f_i^2$ est plus petit.

Considérons un découpage tel que l'AUC vaut 0,9, ayant deux classes. Une avec un effectif contenant 90% de la population et l'autre 10%. Son indice de Gini vaut 0,18.

Soit le deuxième découpage avec une AUC valant 0,6, deux classes de répartition égale. Son indice de Gini vaut 0,5.

Bien qu'ayant une AUC plus importante le premier découpage ne sera pas choisi car $0,6 \times 0,5 = 0,3 > 0,18 = 0,9 \times 0,9$.

On peut aussi souhaiter privilégier des découpages plus hétérogènes. Il suffit de multiplier l'AUC par l'inverse du ratio précédent.

Nous décrivons les résultats obtenus pour chaque variable.

❖ Age du conducteur

¹ L'indice ou le coefficient de Gini est une mesure statistique de la dispersion d'une distribution dans une population donnée développée par le statisticien Corrado Gini. Ce coefficient est compris entre 0 et 1. Il vaut 0 si l'égalité (l'homogénéité) est parfaite, et 1, si l'inégalité (hétérogénéité) est parfaite.

L'aire sous la courbe ROC avant découpage des variables est de 0,53. Nous cherchons à obtenir le découpage qui maximisera cette aire. Nous indiquons dans le tableau suivant l'évolution de l'AUC à chaque découpage pour l'itération retenue.

Nombre de classes	Bornes	AUC	Evol AUC
2	18-48+	0,535	0,94%
3	18-32-34+	0,553	3,36%
4	18-27-37-43+	0,563	1,81%
5	18-24-28-47-52+	0,574	1,95%

Tableau 5- AUC maximale par nombre de classes

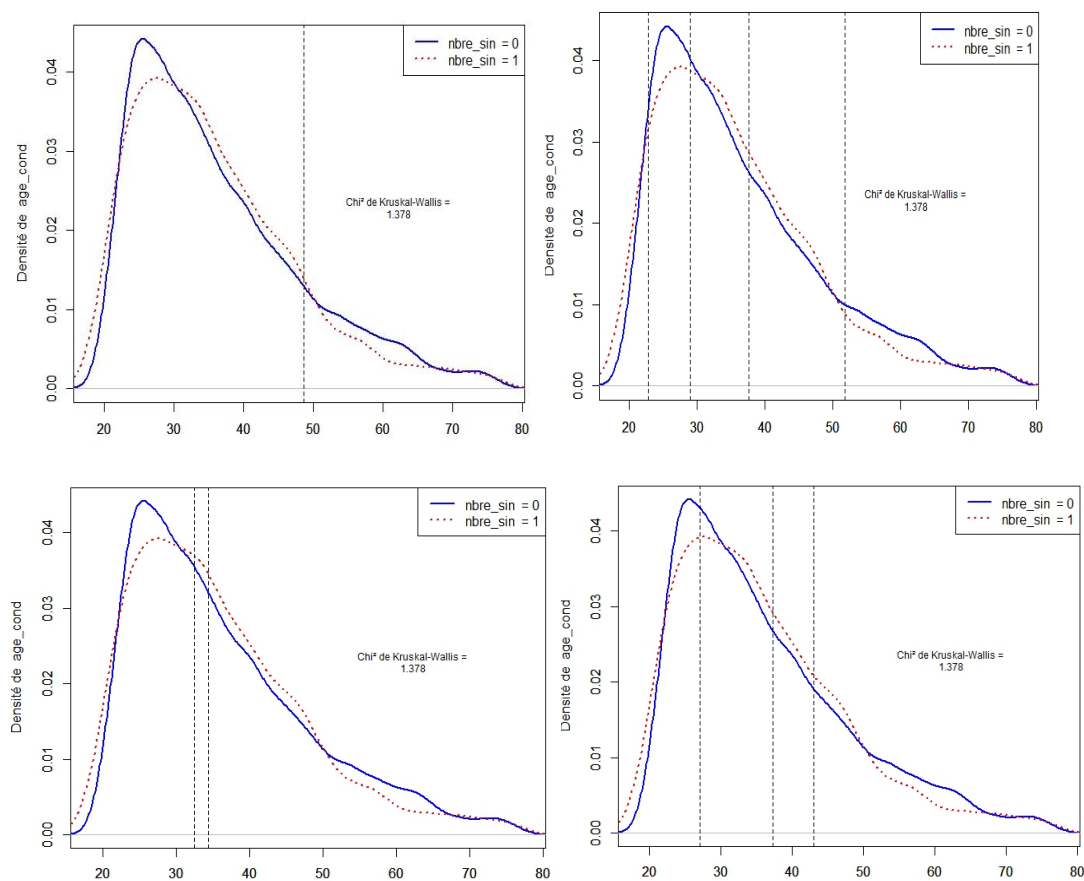


Figure 7- Découpage de la variable âge du conducteur

On ne constate qu'une discrimination bien évidente dans le découpage en fonction de l'âge du conducteur. En effet, de 18 à 24 ans, la densité d'avoir au moins un sinistre conditionnellement à l'âge est supérieure à la densité conditionnelle au sans sinistre. Une inversion se produit entre 24 et 28 ans. Puis de 28 à 47 une légère supériorité de la courbe des sinistrés. Enfin après 50 ans les assurés sont moins risqués.

❖ Ancienneté du permis

L'AUC initial est de 0,576 avant découpage. Le découpage de l'ancienneté du permis avec l'aire sous la courbe ROC et son évolution sont contenus dans le tableau suivant :

Nombre de classes	Bornes	AUC	Evol AUC
2	0-12+	0,561	-2,60%
3	0-8-16+	0,569	1,43%
4	0-2-13-22+	0,578	1,58%
5	0-2-8-17-24+	0,583	0,87%

Tableau 6- AUC maximale par nombre de classes

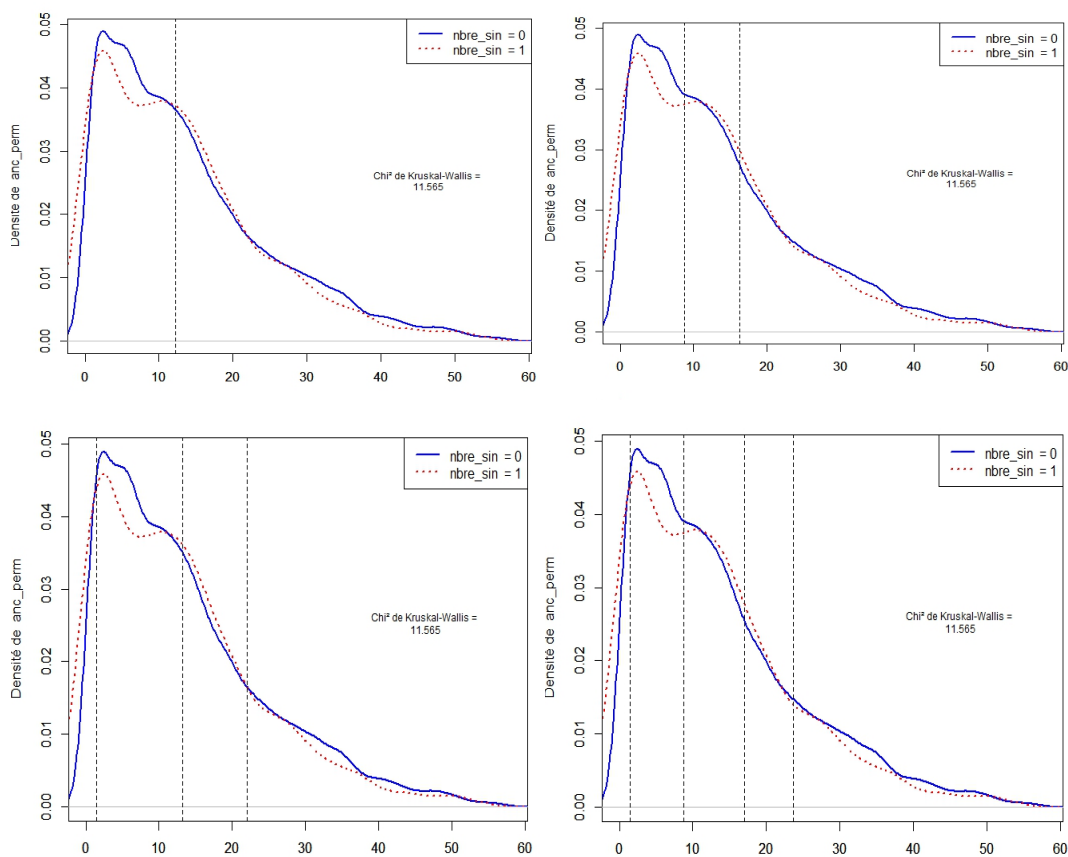


Figure 8 – Découpage de la variable ancienneté du permis

La variable, ancienneté du permis, a une discrimination peu prononcée. Seule la classe [2-8] présente un décrochage significatif entre les densités des sinistrés et des non sinistrés.

❖ Ancienneté du véhicule

Le tableau ci-dessous présente l'évolution de l'aire sous la courbe ROC pour chaque découpage de l'itération optimale. Celle-ci est de 0,521 avant le découpage.

On a le tableau suivant :

Nombre de classes	Bornes	AUC	Evol AUC
2	0-9+	0,526	0,96%
3	0-5-15+	0,528	0,38%
4	0-6-10-26+	0,532	0,76%
5	0-6-9-12-17+	0,537	0,94%

Figure 9- AUC maximale par nombre de classes

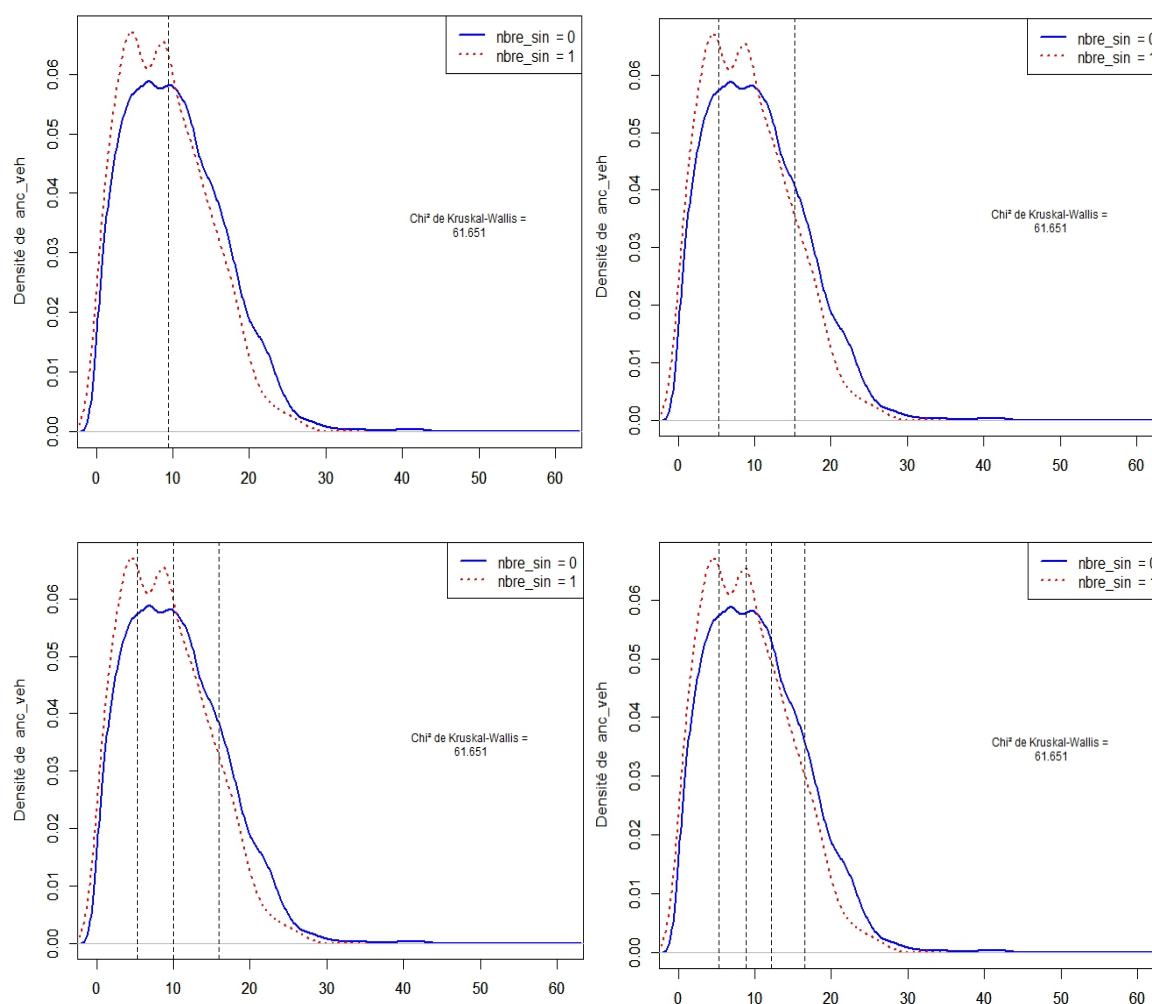


Figure 10 – Découpage ancienneté du véhicule

De 0 à 10 ans la courbe des sinistrés est supérieure à celle des non sinistrés. On constate une inversion des positions après 10 ans.

❖ Comparaison des résultats obtenus par les deux méthodes

Les résultats du découpage présentent des différences considérables. En regardant la sinistralité, les découpages issus de la méthode supervisée, ne créent pas de classes où les différences de sinistralité par classe de regroupement ne sont pas marquées. On préférera la méthode non supervisée, l'objectif étant de dissocier clairement les profils de risque.

Tout au long de notre analyse des variables explicatives qualitatives, nous n'avons cessé de relever la présence de pic dans certaines modalités des variables d'étude. En effet, ces effets sont dus à des sinistres dont les coûts représentent une part très importante du total de la charge sinistre comme l'indique le boxplot suivant.

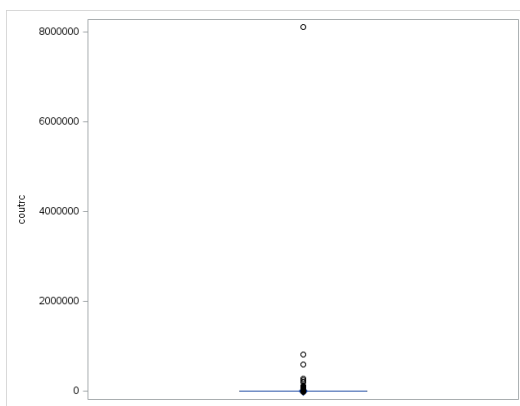


Figure 11 - Boxplot de la charge sinistre RC

Nous avons comparé graphiquement en annexe 0, l'effet du top 15 des sinistres les plus élevés sur l'ensemble des modalités en observant les coûts moyens avec et sans ces sinistres.

Nous constatons que les coûts moyens des sinistres sont plutôt homogènes par modalité lorsqu'on élimine les sinistres très élevés. Toutefois, nous ne pouvons pas déterminer le caractère exceptionnel du montant d'un sinistre en ne considérant qu'un certain nombre de sinistres. Nous devons donc exploiter des techniques nous permettant de déterminer ces sinistres.

IV. Sinistres graves

En assurance automobile, certains sinistres, notamment les dégâts corporels significatifs en responsabilité civile, peuvent causer des sinistres majeurs. En tarification, certains sinistres à faible apparition et forte sévérité peuvent donc perturber l'estimation. Pour conserver l'hypothèse d'homogénéité des risques, les sinistres observés sont souvent écrêtés. L'écrêtement consiste à plafonner les sinistres à un montant maximum M . La charge résiduelle est quant à elle répartie à l'ensemble des risques suivant une clé de répartition.

Le choix de ce plafond est très délicat car, s'il est mal calibré, peut conduire à une sous-estimation ou une surestimation de la prime sur certains segment. Par exemple, les jeunes conducteurs qui attirent plus de sinistres graves auraient une prime trop élevée.

Pour déterminer ces plafonds, il existe plusieurs principes du moins théorique au plus théorique.

1^{ère} Approche : Ce principe consiste à choisir un montant (d'expérience) forfaitairement

2^{ème} Approche : Une technique consiste à choisir un seuil tel que la charge sur-crête représente un pourcentage de la charge totale de sinistres.

3^{ème} Approche : Le seuil peut être choisi sur la base d'un quantile de la distribution historique des coûts unitaires.

4^{ème} Approche : La théorie des valeurs extrêmes peut aider dans le choix du seuil d'écèlement. Pour ce faire on suppose que les sinistres unitaires sont extrêmes. Il suffit d'un test d'adéquation avec la loi de Pareto. Dans ce cas l'observation selon laquelle l'espérance de coût résiduel $\mathbb{E}(X - M|X > M)$, au-delà du seuil est linéaire, permet d'envisager des techniques de détermination de seuil.

L'inconvénient de la 1^{ère} approche réside dans l'absence de fondements techniques.

Les 2^{ème} et 3^{ème} approches bien plus techniques que la première possède un caractère arbitraire. En effet, on a pour la 2^{ème}, le choix du pourcentage de charge sur-crête par rapport à la charge totale, et pour la 3^{ème} le choix du quantile (en pratique 0,5%, 1% ou 2%).

La dernière approche est celle que l'on retient, au vu de l'adéquation de nos coûts à la loi de Pareto.

VarName	Distribution	Test	TestLab	Stat	pType	pValue
coutrc	Pareto	Kolmogorov-Smirnov	D	0,23	Pr > D	0,000
coutrc	Pareto	Cramer-von Mises	W-Sq	48,60	Pr > W-Sq	0,000
coutrc	Pareto	Anderson-Darling	A-Sq	218,08	Pr > A-Sq	0,000

Les p_values sont toutes très faibles et donc l'hypothèse nulle d'ajustement à la loi de Pareto n'est rejetée.

De plus à l'aide des fonctions de répartition théorique et empirique, on peut confirmer cette intuition. Comme le montre le graphique suivant

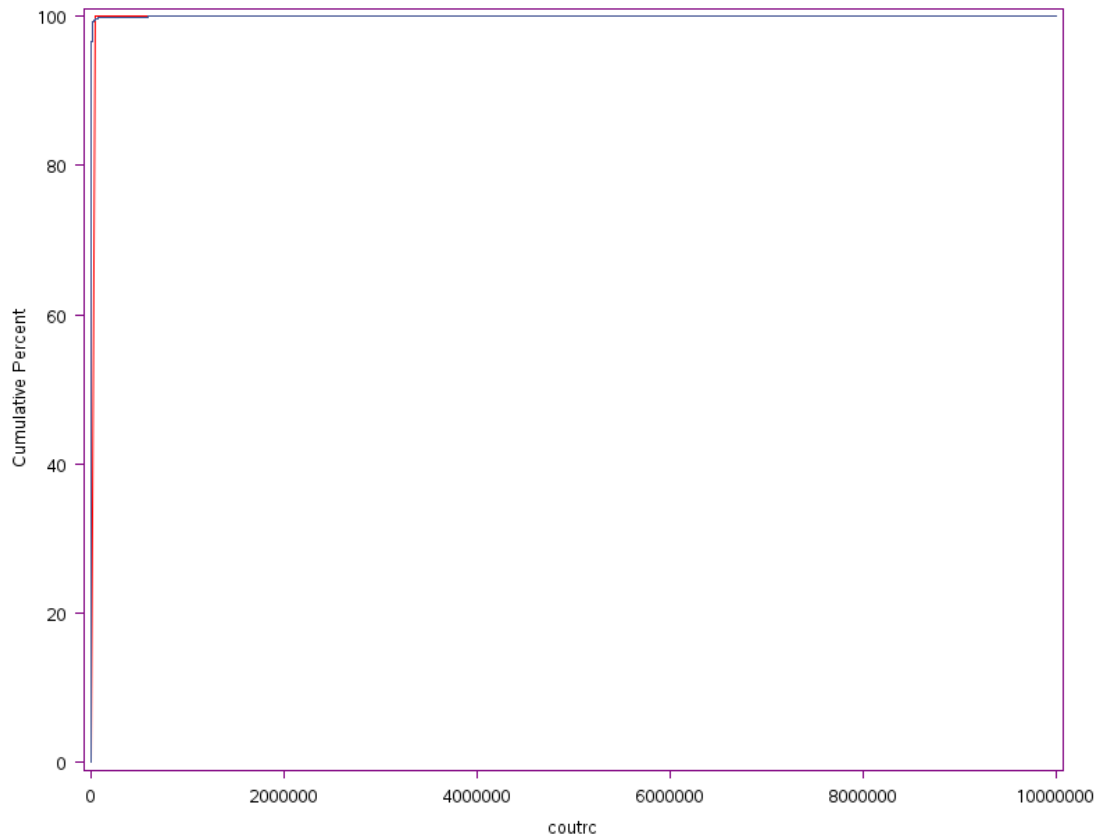


Figure 12 - Répartition empirique de la charge RC

- Fonction de répartition théorique de la loi de Pareto
- Fonction de répartition empirique de la charge RC

Cette approche conduit à plusieurs méthodes de détermination de seuil. Parmi elles, nous évoquerons les plus utilisées en pratique après un bref résumé des théories qui sous-tendent ces méthodes.

A. Valeurs extrêmes et dépassement de seuil

Soient $(X_i)_{i=1}^n$ une suite de variables aléatoires indépendantes et identiquement distribuées et leur maximum $M_n = \max(X_i)_{i=1}^n$.

Le but de cette théorie est de déterminer la loi limite normalisée que suit le maximum (ou le minimum) en fonction de celle de la variable aléatoire X . Notons F_X , la fonction de répartition de la variable X de loi de probabilité \mathbb{P} . La fonction de répartition de M_n est alors définie par :

$$F_{M_n}(x) = \mathbb{P}(M_n < x) = \mathbb{P}(X_1 < x, \dots, X_n < x) = [F_X(x)]^n$$

La fonction de répartition de X n'étant généralement pas connue, il est difficile de déterminer celle du maximum à partir de ce résultat. Les valeurs extrêmes se trouvant à droite et en fin de

distribution, il apparait naturel de décrire la queue de distribution X à l'aide de la loi asymptotique de M_n .

Notons $x_F = \sup\{x \in R: F_X(x) < 1\}$ le point extrême à droite de la fonction F_X . Ce point peut être fini ou infini (Embrechts et al., 1997, exemple 3.3.22, p.139).

La distribution asymptotique de M_n s'en déduit aisément :

$$\lim_{n \rightarrow \infty} F_{M_n}(x) = \lim_{n \rightarrow \infty} [F_X(x)]^n = \begin{cases} 0 & \text{si } x < x_F \\ 1 & \text{si } x \geq x_F \end{cases}$$

On obtient une masse de Dirac en x_F , la distribution asymptotique du maximum est donc une loi dégénérée. Cette loi ne fournit pas assez d'information, d'où l'idée d'opérer une transformation de standardisation du maximum.

Théorème 1 (Fisher et Tippett, 1928, Gnedenko, 1943)

Soient $(a_n)_{n \geq 1}$ et $(b_n)_{n \geq 1}$ avec $b_n > 0$, et une fonction de répartition G non dégénérée telle que, $\frac{M_n - a_n}{b_n} \xrightarrow{d} G$ lorsque n tend vers l'infini, alors G est nécessairement de l'un des trois types suivant :

$$G_0(x) = \exp(-\exp(-x)), -\infty < x < +\infty$$

$$G_{1,\alpha}(x) = \begin{cases} 0 & x < 0 \\ \exp(-x^{-\alpha}) & x \geq 0, \alpha > 0 \end{cases}$$

$$G_{2,\alpha}(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & x < 0, \alpha < 0 \\ 1 & x \geq 0 \end{cases}$$

Avec $G_0, G_{1,\alpha}, G_{2,\alpha}$ respectivement appelé loi de Gumbel, loi de Fréchet et loi de Weibull.

On dit alors que X appartient au domaine d'attraction maximum de G et l'on note $X \in MDA(G)$. Jenkinson (1955) a proposé une écriture unifiée de ces trois types de loi limite du maximum. On obtient la forme plus générale de la distribution des valeurs extrêmes notée GEV (Generalized Extreme Value Distribution).

Elle correspond à :

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

Où ξ est appelé paramètre de forme ou indice de queue qui ne dépend que de la loi de X . Plus cet indice est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On parle alors de distribution à queue épaisse. La fonction de répartition $G_{\mu,\sigma,\xi}$ est dans le domaine d'attraction maximum, respectivement, de Fréchet, Gumbel ou Weibull selon que $\xi > 0$, $\xi = 0$ ou $\xi < 0$ (Galambos, 1987, pp.53-54, Embrechts et al., 1997, p.152).

B. Distribution conditionnelle des excès

Soit u un réel suffisamment grand, inférieur au point terminal x_F , appelé seuil.

Notons $E_u = \{j \in \{1, 2, \dots, n\} | X_j > u\}$ l'ensemble des indices de dépassement de seuil.

On définit comme excès au-delà du seuil la variable aléatoire Y_j telle que : $Y_j = X_j - u$, $j \in E_u$.

On cherche à caractériser la loi F_u de la variable $X | X > u$.

$$F_u(x) = \mathbb{P}(X < x | X > u) = \frac{F_X(x) - F_X(u)}{1 - F_X(u)} \quad \text{Pour } x \geq u. \quad (1)$$

Ce qui équivaut à :

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F_X(u+y) - F_X(u)}{1 - F_X(u)} \quad \text{Pour } 0 \leq y \leq x_F - u. \quad (2)$$

Théorème 2 (Pickands, 1975 ; Balkema et de Haan, 1974)

Soit $F_{\xi, \sigma}^{GPD}$ la fonction de répartition de la loi de Pareto généralisée (GPD), définie par :

$$F_{\xi, \sigma}^{GPD}(x) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - \exp(-y/\sigma) & \text{si } \xi = 0 \end{cases}$$

$$\text{Pour : } \begin{cases} 0 \leq y \leq x_F - u & \text{si } \xi \geq 0 \\ 0 \leq y \leq \text{Min}(-\sigma/\xi, x_F - u) & \text{si } \xi < 0 \end{cases}$$

F_X appartient au domaine d'attraction du maximum $G_{0,1,\xi}$ si et seulement si il existe une fonction positive σ :

$$\lim_{u \rightarrow x_F} \sup_{0 < y < x_F - u} |F_u(y) - F_{\xi, \sigma(u)}^{GPD}(y)| = 0$$

Ce théorème exprime le lien entre le paramètre de la loi du domaine du maximum et le comportement limite des excès au-delà d'un seuil assez grand. Plus particulièrement l'indice de queue ξ , obtenu dans l'étude du maximum, est identique au paramètre de la loi de Pareto généralisée.

C. Les méthodes d'estimation du seuil

La théorie des valeurs extrêmes présente plusieurs méthodes de détermination du seuil au-delà duquel une valeur sera considérée comme extrême. On distingue trois méthodes généralement rencontrées dans la littérature qui sont :

Les valeurs record

La fonction moyenne des excès

L'approximation GPD (par la distribution de Pareto généralisée)

1. Les valeurs record

Soit le processus de comptage défini comme suit :

$$N_1 = 1, \quad N_n = 1 + \sum_{k=2}^n 1_{\{X_k > M_{k-1}\}}, \quad \text{Pour } n \geq 2$$

Avec cette méthode, le seuil correspond à une valeur de la distribution. L'espérance de N_n permet d'estimer le nombre de valeurs extrêmes puis d'en déduire le seuil au-delà duquel les observations sont extrêmes à partir de la statistique d'ordre.

L'espérance et la variance de N_n s'écrivent comme suit :

$$\mathbb{E}(N_n) = \sum_{k=1}^n \frac{1}{k}, \quad \text{Var}(N_n) = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k^2} \right).$$

n	$\mathbb{E}(N_n)$	$\text{Var}(N_n)$
1 000	7,49	6,84
2 000	8,18	7,53
3 000	8,58	7,94
5 000	9,09	8,45
10 000	9,79	9,14
25 000	10,70	10,06
50 000	11,40	10,75
100 000	12,09	11,45

Tableau de l'espérance et de la variance du nombre de dépassement de seuil en fonction de la taille n de l'échantillon.

L'estimation de notre seuil se situe à environ 22 000 €.

Cette méthode ne tient pas compte de la distribution des coûts et impose au seuil d'être une observation de la distribution empirique.

2. La fonction moyenne des excès

La fonction moyenne des excès (FME) est définie par : $e_n(u) = E(X - u | X > u)$ permet de prédire les dépassements de seuil u . Celle-ci est estimée par la somme des excès dépassant un seuil élevé u , divisé par le nombre d'observations dépassant ce seuil.

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

Où $(X - u)^+ = \text{Sup}(X_i - u, 0)$

Cette approche consiste à choisir le seuil u tel que la fonction \hat{e}_n soit linéaire pour tout $x \geq u$. En effet, la fonction de l'espérance résiduelle d'un GPD de paramètre $\xi < 1$, étant affine (Schirmacher 2005), on cherchera un u aussi petit que possible de sorte à ce que l'estimateur empirique de l'espérance résiduelle des excès soit approximativement linéaire. Le seuil est donc déterminé à partir de l'instant où le graphe présente une partie affine stable.

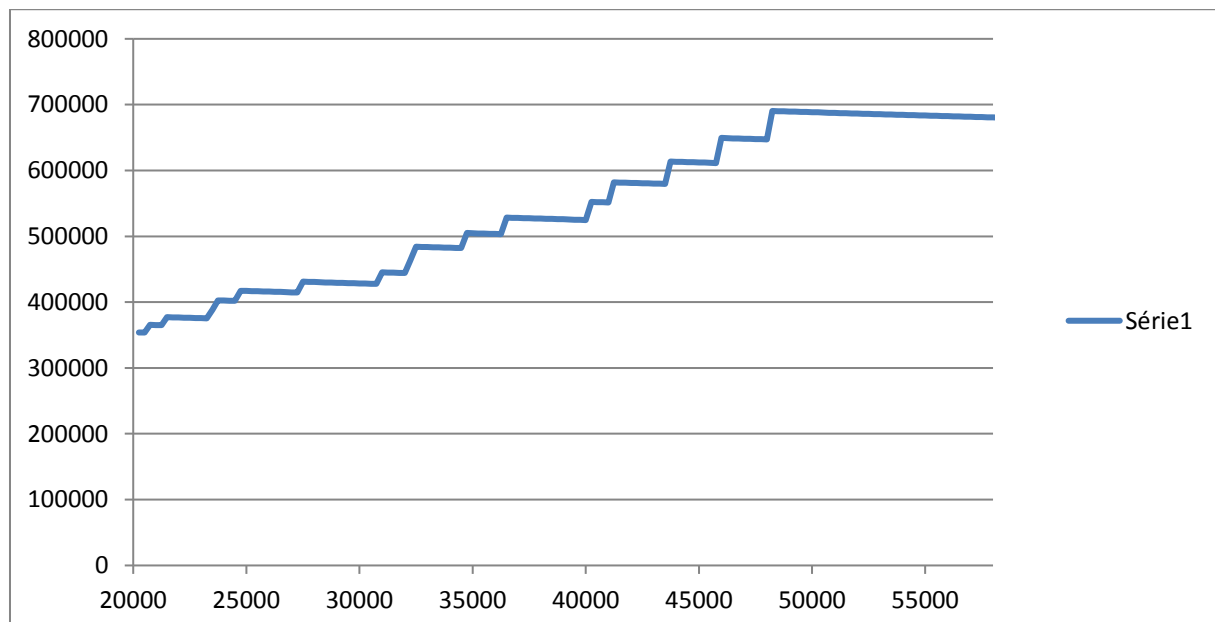


Figure 13 - Fonction moyenne des excès des coûts sinistre

L'estimation graphique du seuil consiste à repérer les plages de linéarité. Nous observons plusieurs plages de linéarité mais elles ne sont pas acceptables car on observe un changement de pente après celle-ci. Le seuil retenu serait donc $\hat{u} = 48\,500 \text{ €}$.

Cette méthode est parfois délicate car le graphique peut présenter des aspects irréguliers donc difficile à interpréter.

1. L'approximation GPD

La méthode est basée sur l'estimation du quantile extrême. Au-delà de ce quantile, les observations sont jugées extrêmes et le nombre de valeurs extrêmes correspond à celui des observations qui dépassent ce quantile estimé par l'approximation GPD. Cette méthode tire son fondement du théorème 2 et utilise donc la distribution de Pareto généralisée $F_{\xi, \sigma}^{GPD}$.

Pour ce faire, on définit un seuil aléatoire u_n comme étant le quantile d'ordre $1 - \frac{m_n}{n}$ où m_n le nombre d'excès est choisi tel que $\lim_{n \rightarrow \infty} m_n = +\infty$ et $\lim_{n \rightarrow \infty} \frac{m_n}{n} = 0$.

Des solutions pour fixer m_n de manière à obtenir un estimateur asymptotiquement sans biais ont été proposées par Goldie et Smith (1978) puis De Haan et Peng (1998). Nous avons utilisé dans notre étude l'approximation du nombre moyen de dépassement de seuil par $\ln(n) + \gamma$. Où γ est la constante d'Euler qui vaut environ 0,577.

Afin de déterminer le quantile extrême, nous devons vérifier si un modèle paramétrique permet d'obtenir une bonne approximation de la loi des données, particulièrement en queue de distribution. Reiss et Thomas (1997) ont développé un test d'adéquation de queue de distribution aux données. Le modèle doit s'adapter autant à la queue qu'à l'ensemble de la distribution.

En effet, pour $x \geq u$ (des points appartenant à la queue de distribution), en utilisant les équations (1) ou (2) on a :

$$F_X(x) = \mathbb{P}(X \leq x) = (1 - \mathbb{P}(X \leq u))F_u(x - u) + \mathbb{P}(X \leq u)$$

Dans la relation précédente, on peut estimer $F_u(x - u)$ par $F_{\xi, \sigma}^{GPD}(x - u)$ pour u élevé. Cette approximation est déduite du théorème 2. Nous pouvons également estimer $\mathbb{P}(X \leq x)$ à partir des données par la fonction de repartition empirique $F_n(u)$ évaluée au point u :

$$F_n(u) = 1 - \frac{1}{n} \sum_{i=1}^n 1_{\{X_i > u\}}.$$

Cela signifie que pour $x \geq u$ nous pouvons utiliser l'estimation de la queue :

$$\hat{F}(x) = [1 - F_n(u)]F_{\xi, \sigma}^{GPD}(x - u) + F_n(u), \text{ pour approximer la fonction de repartition } F_X(x)$$

L'estimateur du quantile extrême $q_{GPD, n}$, issu de l'approximation par la loi de Pareto généralisée et qui sera utilisé par la suite pour la détermination du seuil se présente comme suit :

$$q_{GPD, n} = u_n + \frac{\sigma_n}{\xi_n} \left[\left(\frac{np_n}{m_n} \right)^{-\xi_n} - 1 \right]$$

Les paramètres σ_n et ξ_n sont déterminés par la méthode du maximum de vraisemblance.

Dans cette étude nous essayons d'estimer un quantile extrême avec une probabilité de 99,9% d'être une valeur extrême (un sinistre grave) pour la distribution du coût des sinistres. Nous avons effectué une estimation ponctuelle.

Paramètre	Estimation
σ_n	1 258,654
ξ_n	0,31054
u_n	40 000

Tableau 7- Paramètres obtenus sur notre base

on obtient 36 527 € comme quantile extrême.

La méthode de GPD offre une méthode de détection des valeurs extrêmes minimale (avec peu de valeurs extrêmes).

En considérant la moyenne des seuils issues des trois méthodes, nous retiendrons un seuil de sinistres graves à 35 000 € dans notre tarification.

Notre étude exploratoire, nous a permis d'analyser nos variables explicatives lesquelles nous permettrons de réaliser notre tarification à l'aide de notre GLM.

DEUXIEME PARTIE

GLM EN TARIFICATION AUTOMOBILE

Les GLMs sont les méthodes de tarification les plus utilisées en tarification IARD. Ils permettent de modéliser les liaisons linéaires qui existent entre la variable à expliquer et les variables explicatives. Cette partie a pour but de présenter les fondements théoriques puis d'appliquer ces modèles pour calculer la prime pure de la garantie RC à notre portefeuille de contrats automobiles.

I. Tarification selon le modèle linéaire généralisé

Les modèles linéaires généralisés en anglais Generalized Linear Model (GLM) sont les techniques statistiques les plus répandues en matière de tarification actuarielle. Ces concepts furent introduits en 1972 par les statisticiens John NELDER et Robert WEDDERBURN. Historiquement, les actuaires utilisaient les modèles linéaires pour quantifier les relations entre la valeur des contrats des risques assurés et les variables décrivant ce risque. Suite à la complexité grandissante des problématiques actuarielles, les méthodes de régression simple se sont avérées très vite désuètes et les actuaires se sont ainsi tournés vers l'utilisation de GLM. L'une des premières utilisations de ce modèle a eu lieu vers la fin du 20^{ème} siècle à la City University, par des actuaires londoniens.

Ces modèles permettent de modéliser à la fois des comportements non linéaires et des distributions de résidus non gaussiens.

Pour modéliser la prime pure nous utiliserons le logiciel R.

A. Notions théoriques du modèle linéaire généralisé

1. Présentation du modèle linéaire généralisé

Les GLMs sont formés de trois composantes :

La variable de réponse Y , **composante aléatoire** à laquelle est associée une loi de probabilité.

La composante déterministe définie comme combinaison linéaire des variables explicatives X_1, \dots, X_k

La fonction lien, définissant la relation fonctionnelle entre la combinaison linéaire des variables X_1, \dots, X_k et l'espérance mathématique de la variable réponse Y .

Ce qui s'écrit mathématiquement comme suit :

$$g[\mathbb{E}(Y)] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

❖ La composante aléatoire

La loi de probabilité de la composante aléatoire appartient à la famille exponentielle ce qui signifie que sa densité peut s'écrire sous la forme suivante :

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

Où :

- le support S est un sous-ensemble de \mathbb{N} ou \mathbb{R}
- θ est un paramètre réel, appelé paramètre canonique (ou paramètre de la moyenne)
- ϕ est un paramètre réel, appelé paramètre de dispersion
- a est une fonction définie sur \mathbb{R}^*
- b est une fonction définie sur \mathbb{R} et deux fois dérivable
- c est une fonction définie sur \mathbb{R}^2

❖ La composante déterministe

La composante déterministe, exprimée sous forme d'une combinaison linéaire $\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$ (appelée aussi prédicteur linéaire) précise quels sont les prédicteurs.

Certaines variables peuvent se déduire de variables explicatives étudiées dans le modèle, par exemple :

$X_j = X_k * X_l$ se définit comme l'interaction des variables X_k et X_l

Ou encore $X_j^2 = X_k$ de façon à prendre en compte l'effet non linéaire de X_k

L'estimation des paramètres $\beta_0, \beta_1, \dots, \beta_k$ sont estimées en maximisant la log-vraisemblance du GLM.

❖ La fonction de lien

La fonction de lien permet de traduire la relation fonctionnelle entre la composante aléatoire et la composante déterministe. Elle spécifie comment l'espérance de Y , notée μ , est liée au prédicteur.

La fonction de lien $g(\mu) = \log(\mu)$ permet par exemple de modéliser le logarithme de l'espérance. Les modèles utilisant cette fonction de lien sont les **modèles log-linéaires**.

La fonction de lien $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ modélise le logarithme du rapport des chances. Elle est appelée logit et est adaptée au cas où μ est comprise entre 0 et 1.

2. Comparaison de modèles : les indicateurs statistiques

❖ Déviance

Afin d'évaluer la qualité d'ajustement d'un modèle linéaire généralisé, il est courant de calculer la statistique dite des écarts, appelée la déviance. Elle compare le modèle estimé au modèle saturé, c'est-à-dire le modèle qui comporte autant de paramètre à estimer que d'observations et représentant exactement les données.

On désigne respectivement par $D, \mathcal{L}, \mathcal{L}_{sat}$ la déviance, la log-vraisemblance du modèle estimé et la log-vraisemblance du modèle saturé.

$$D = -2(\mathcal{L} - \mathcal{L}_{sat})$$

Une faible valeur de la déviance traduit un ajustement de bonne qualité. En effet, plus la valeur de la déviance est faible et plus les deux log-vraisemblances sont proches. L'objectif est donc de minimiser la valeur de la déviance.

Lorsque le modèle étudié est exact, D suit approximativement une loi de χ^2 à $n - k$ degrés de liberté.

❖ Critère d'information d'Akaike (AIC)

Le critère AIC est une mesure de la qualité d'un modèle statistique proposé par Hirotugu Akaike en 1973.

L'AIC utilise le maximum de vraisemblance, mais en pénalisant les modèles comportant trop de variables, qui « sur apprennent » et généralisent mal.

On note par k le nombre de paramètres du modèle. Le critère AIC est défini de la manière suivante :

$$AIC = -2\mathcal{L} + 2k$$

On utilise le critère d'information d'Akaike corrigé (AICc) lorsque le nombre de paramètres k est grand par rapport au nombre d'observations n dans l'échantillon (si $n/k < 40$). Le critère AICc est une correction du critère AIC pour les petits échantillons.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Le meilleur modèle est celui qui a le plus petit critère AIC (resp. AICc)

❖ Critère d'information bayésien (BIC)

Le critère d'information Bayésien est un critère d'information dérivé du critère d'information d'Akaike proposé par Gideon Schwartz en 1978. A la différence du critère d'information d'Akaike, la pénalité dépend de la taille et pas seulement du nombre de paramètres.

Le critère se définit comme :

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

Il faut choisir le modèle ayant le plus petit BIC.

Dans notre étude nous enterons de minimiser les trois critères présentés. Toutefois en cas de contradiction entre les critères nous privilégieront le critère AIC.

3. L'analyse des résidus

On note y_i la valeur observée, μ_i la valeur prédite, w_i la contribution de l' $i^{\text{ème}}$ observation à la statistique de Pearson et $V(\cdot)$ la fonction variance.

Pour l'observation i le résidu $r_i = y_i - \hat{u}_i$ n'a qu'un intérêt limité, les modèles étant hétéroscédastiques. D'autres types de résidus sont utilisés pour rendre compte de l'adéquation du modèle aux données observées. Les plus connus sont les résidus de Pearson et les résidus de déviance.

❖ Résidus de Pearson

Les résidus de Pearson mesurent la contribution de chaque observation à la statistique de Pearson.

$$r_i^P = \frac{\sqrt{w_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

❖ Résidus de déviance

Les résidus de déviance mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Les résidus de déviance se définissent de la manière suivante :

$$r_i^D = \text{signe}(y_i - \mu_i) \sqrt{d_i}$$

Où d_i représente la contribution de la $i^{\text{ème}}$ observation à la déviance.

Afin de ne pas avoir des effets doubles, liés à une forte corrélation entre les variables explicatives, dans la modélisation, les variables explicatives, il faut en supprimer certaines ou ne garder que leur interaction.

4. Etudes de corrélations

Cette étude de dépendance entre les variables explicatives est basée sur l'analyse de la matrice de corrélation avec comme mesure de dépendance le V de Cramer.

Le V de Cramer (ou phi de Cramer) mesure l'intensité de la relation entre deux variables qualitatives et est basé sur la statistique du Chi2. Contrairement au Chi2, cette mesure reste stable si l'on augmente la taille de l'échantillon dans les mêmes proportions inter-modalités.

Soient un tableau de contingence avec :

p lignes et q colonnes

n le nombre total des observations,

n_{ij} l'effectif de ligne i ($i = 1, \dots, p$) et de colonne j ($j = 1, \dots, q$)

$n_{i.}$ le total de la ligne i

$n_{.j}$ le total de la ligne j

Nous définissons la statistique du Chi2 de la façon suivante :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Le V de Cramer s'exprime alors comme la racine carrée du rapport entre la statistique du Chi2 et du produit de la taille de l'échantillon et de la dimension maximale :

$$V = \sqrt{\frac{\chi^2}{n \times \min(p-1; q-1)}}$$

Plus le V de Cramer entre deux variables est proche de 0, plus les variables sont indépendantes. Inversement, plus le V de Cramer est proche de 1 plus les variables sont dépendantes. Nous obtenons la matrice partielle de corrélation dans le tableau ci-dessus.

	Groupe	Mode de paiement	de CRM	Age du conducteur	Ancienneté du permis	Ancienneté du contrat
<i>Interruption d'assurance</i>	0,02	0,01	0,07	0,05	0,08	0,01
<i>Classe</i>	0,47	0,05	0,12	0,13	0,11	0,03
<i>Groupe</i>	1	0,03	0,09	0,1	0,09	0,03
<i>Mode d'acquisition</i>		0,04	0,03	0,04	0,03	0,03
<i>zone</i>		0,06	0,06	0,04	0,04	0,03
<i>usage</i>		0,06	0,08	0,27	0,05	0,04
<i>parking</i>		0,03	0,08	0,06	0,05	0,02
<i>Fractionnement</i>		0,71	0,15	0,09	0,09	0,04
<i>Mode de paiement</i>		1	0,18	0,13	0,12	0,02
<i>Periode de sinistre</i>			0,55	0,19	0,4	0,06
<i>Annee</i>			0,12	0,11	0,12	0,23
<i>Coefficient bonus-malus</i>			1	0,23	0,34	0,18
<i>Conduite accompagnée</i>				0,18	0,13	0,03
<i>Age du conducteur</i>				1	0,26	0,05
<i>Nombre d'année au CRM50</i>					0,07	0,17
<i>Ancienneté du véhicule</i>						0,34

- Coefficient V de Cramer supérieur à 0,5
- Coefficient V de Cramer compris entre 0,3 et 0,5
- Coefficient V de Cramer compris entre 0,2 et 0,3

Tableau 8- Matrice de corrélation partielle

B. Modélisation de la fréquence et du coût moyen

Notons Y_{ij} la variable aléatoire décrivant le coût du $j^{\text{ème}}$ sinistre de l'assuré i , X_i celle qui représente la charge annuelle pour l'assuré i et N_i le nombre de sinistres pour cet assuré.

$$\text{On a : } X_i = \sum_{j=1}^{N_i} Y_{ij}$$

Désignons par S la charge de sinistres totale, et n_a le nombre d'assurés.

$$\text{On a : } S = \sum_{i=1}^{n_a} \sum_{j=1}^{N_i} Y_{ij} = \sum_{i=1}^N X_i \quad \text{où } N \text{ est la variable aléatoire du nombre de sinistre annuel.}$$

Sous réserve que les X_i soient indépendants et mutuellement indépendants avec N , on peut déterminer l'espérance et la variance de S (preuve voir annexe 1).

$$\mathbb{E}(S) = \mathbb{E}(X_1)\mathbb{E}(N)$$

$$\mathbb{V}(S) = \mathbb{E}(N)\mathbb{V}(X_1) + \mathbb{V}(N)\mathbb{E}(X_1)^2$$

L'hypothèse de l'indépendance entre la fréquence et le coût n'est pas toujours vérifiée dans la réalité d'où l'intérêt des zones tarifaires qui « décorrèlent » fréquence et coût des sinistres.

1. Modèle coût

❖ Fonction de lien

Nous utiliserons comme fonction lien la fonction logarithmique étant donné le caractère multiplicatif de notre modèle coût x fréquence.

❖ Loi de probabilité du coût des sinistres

Nous allons modéliser la garantie Responsabilité Civile (RC) (indifféremment du type). Nous ne conserverons que les sinistres ayant eu un coût se rapportant à cette garantie. Dans la littérature actuarielle traitant de l'assurance non vie, les modèles proposés pour l'estimation du coût des sinistres font référence à trois lois de probabilité : la loi gamma, la loi de Weibull et la loi log-normale.

Dans notre étude, détachons les sinistres graves des sinistres sous crête. Les sinistres graves seront par la suite repartis sur l'ensemble des sinistres suivant une clé de répartition car le manque de données nous empêche d'y faire une modélisation. Comme le montre le graphique ci-dessous notre coût sinistre ne semble correspondre à aucune loi attendue.

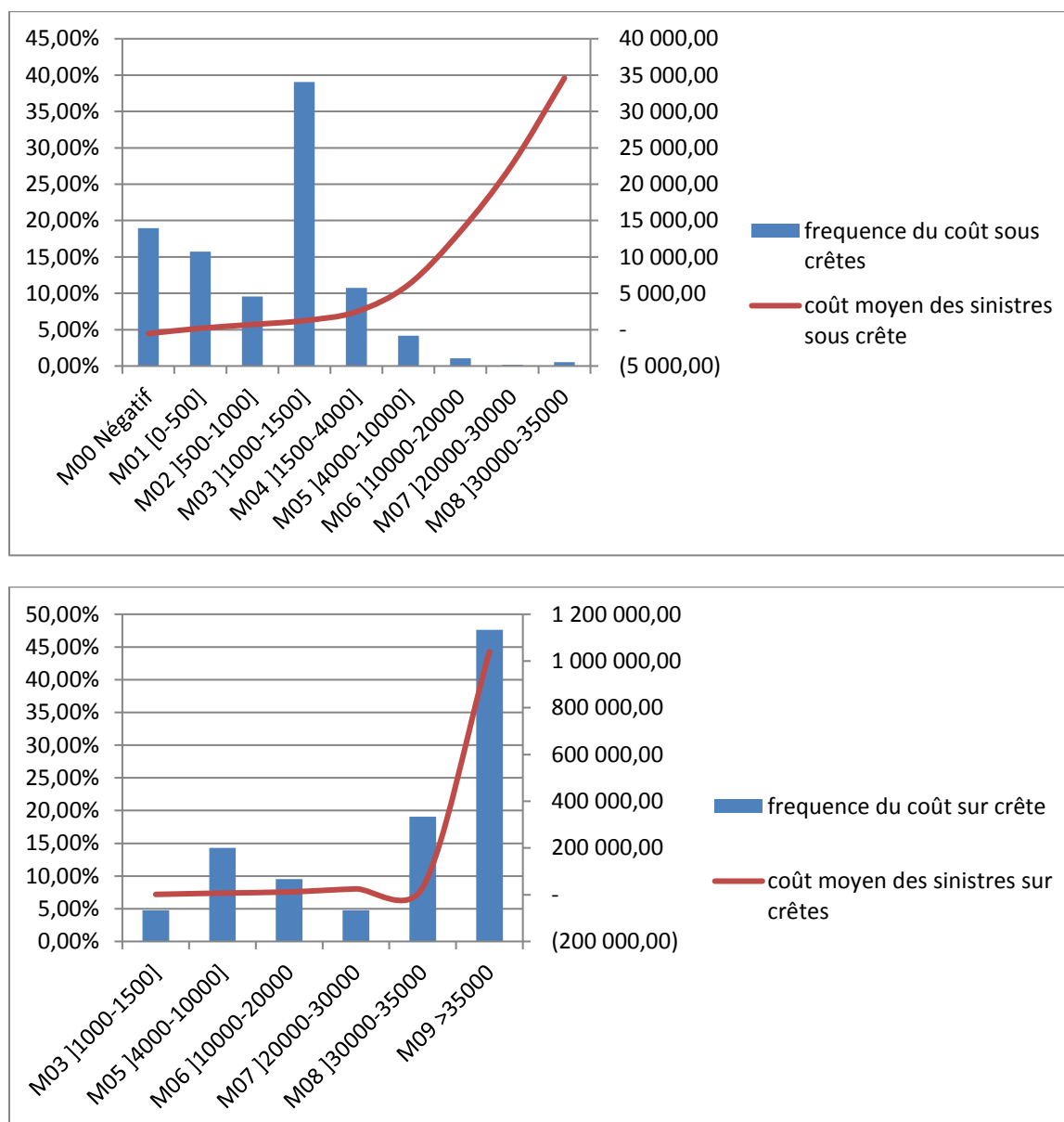


Figure 14 – Représentation de la distribution empirique du coût des sinistres et du coût moyen

On définit les tranches comme suit :

M00 Négatif Tranche des sinistres négatifs

M01 [0-500] Tranche des sinistres compris entre 0 et 500

On constate un pic sur l'histogramme des coûts sous crête, dû à la classe M03.

On remarque la classe M00 des sinistres négatifs, en effet le coût du sinistre n'est pas supposé être négatif.

Le coût moyen quant à lui est continu et croissant suivant les classes.

Comme l'indique le zoom sur la classe M03 dans le graphique ci-dessus, ce pic est lié à une accumulation de montant constant, ce qui est anormal pour une distribution continue.

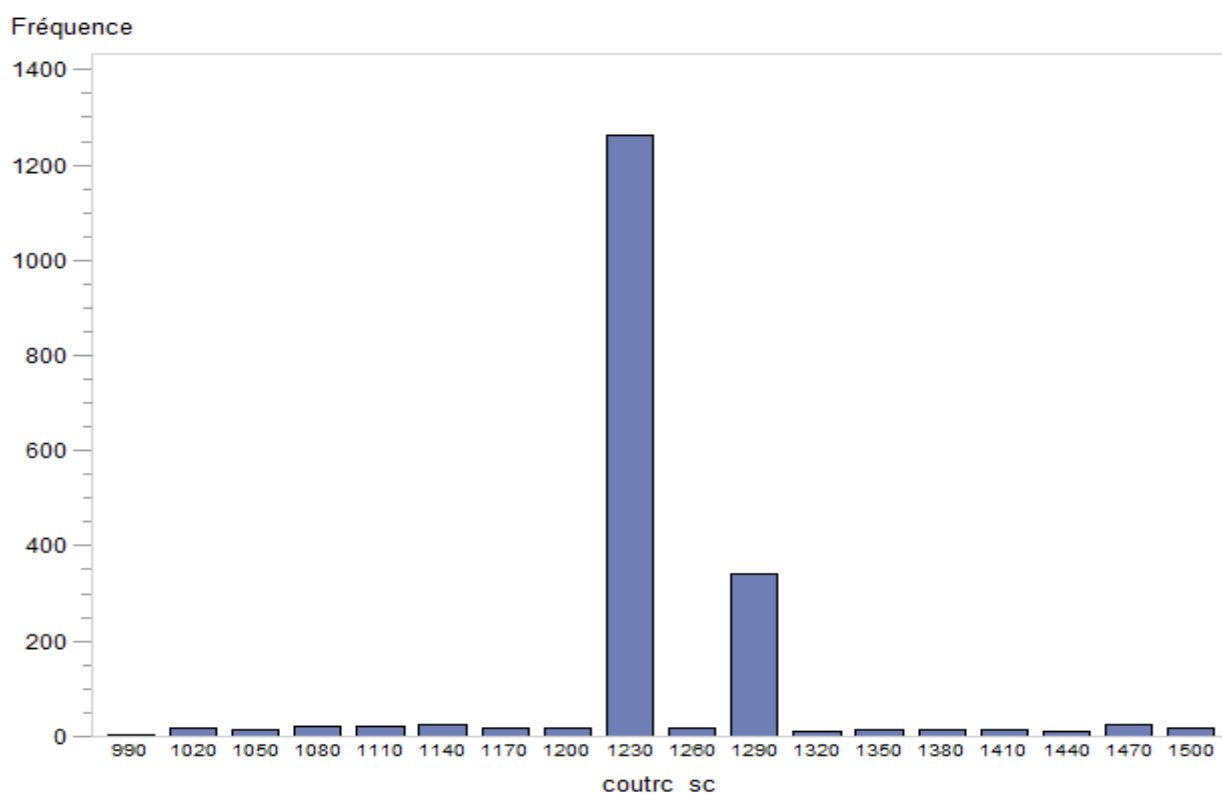


Figure 15 – Distribution du coût RC dans la classe M03

L'accumulation des montants constants et les sinistres négatifs correspond à une indemnisation forfaitaire entre assureur appelé convention IDA (Indemnisation Directe de l'Assuré).

La charge sinistre est tronquée par un principe de compensation entre assureur. Le principe et les exemples nous permettrons de mieux appréhender le problème.

❖ Convention IDA

Le forfait IDA associé à la convention est le montant que s'échangent les assureurs conventionnés en cas de petits sinistres. Si la charge matérielle ne dépasse pas le seuil de 6500€ HT (soit 7774€ TTC), un recours auprès de la compagnie assurant le tiers responsable est effectué à la valeur du

forfait. Si le sinistre dépasse ce seuil, le recours s'effectue à la valeur réelle. Le seuil n'a pas changé depuis 2003 et l'historique du forfait IDA est le suivant :

Période d'application	Montant du forfait
De 2000 à 2001	1 128 €
En 2002	1 172 €
De 2003 à 2011	1 204 €
En 2012	1 236 €
En 2013	1 242 €
En 2014	1 276 €
En 2015	1 308 €
En 2016	1 354 €

Tableau 9- Montant du forfait par année

Suivant le taux de responsabilité de l'assuré, le forfait s'applique différemment. Si notre assuré est 100% responsable, la compagnie tierce peut opposer un recours à hauteur du forfait. S'il n'est pas responsable, on doit demander le forfait IDA à la compagnie adverse.

En cas de responsabilité partagée, le taux de responsabilité s'applique au forfait. La compagnie tierce faisant de même, on considère que la charge recours s'annule.

De plus, le forfait IDA² s'applique à la charge matérielle qu'il y ait la présence ou non d'une charge corporelle. Il faut cependant la présence d'un tiers.

Cette convention influence la modélisation de ma charge sinistre en garantie R.C.

En effet, deux cas de figure se dégagent pour la charge sinistre R.C.

Premier cas : Le sinistre se trouve dans le scope de la convention IDA

Exemple 1: Considérons un sinistre R.C. non responsable d'une valeur de 4000€ dont 3000€ de charge matérielle et 1000€ de charge corporelle, survenu en 2012. L'assureur qui porte le risque paye 4000€ pour la survenance du sinistre et fait recours auprès de la compagnie d'assurance adverse et est remboursé 1000€ pour la garantie corporelle et 1236€ en ce qui concerne la garantie matérielle.

Le coût du sinistre est de 4000€ mais la charge réelle de l'assureur de 1764€.

Exemple 2: Considérons un sinistre R.C. matériel non responsable d'une valeur de 200€, toujours selon cette convention, la charge sinistre serait de -1036€ après recours auprès de la compagnie adverse.

² Indemnisation Directe des Assurés

Le coût du sinistre est de 200€ mais la charge réelle de l'assureur de -1036€

Exemple 3: Considérons un sinistre R.C. responsable survenu en 2012. Le montant du coût du sinistre n'est que partiellement connu car la charge corporelle représenterait la charge réelle, par contre la charge matérielle correspond au montant du forfait dont la valeur peut être au-dessus comme en dessous du coût réel matériel, et nous ne disposons pas de ce coût réel de sinistre adossé au matériel.

Deuxième cas : Le montant du sinistre ne rentre pas dans le cadre de la convention IDA et on peut le modéliser comme présenté dans la littérature actuarielle.

❖ Nouveau modèle

La prise en compte de cette convention entraîne une nouvelle considération de notre modèle de départ. Étant donné la déformation que subit les coûts suite à l'application de la convention IDA, nous ne pouvons modéliser les l'ensemble des coûts de manière identique. Ainsi, le modèle tient compte de l'application ou non de la convention IDA sur un sinistre donné.

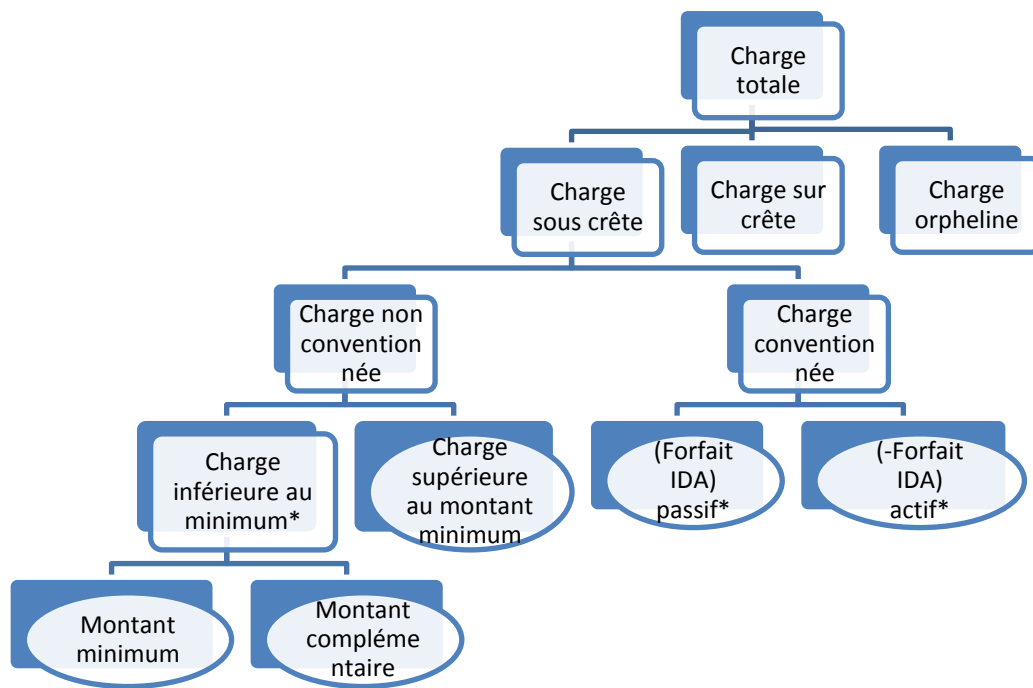
Ainsi nous obtenons :

$$\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^N X_i + \sum_{i=1}^{N_{ida\ passive}} X_i + \sum_{i=1}^{N_{ida\ active}} X_i\right)$$

Étant donné que le montant des forfaits est déterministe et que toute charge comprend une partie non conventionnée (quitte à ce que cette charge soit nulle) on a :

$$\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X_1) + \text{forfait IDA} \times \mathbb{E}(N_{ida\ passive}) - \text{forfait IDA} \times \mathbb{E}(N_{ida\ active})$$

Nous effectuerons la modélisation de la charge totale, en répartissant le montant de la charge sur crête et celui de la charge orpheline, charge à laquelle l'on n'a pu rattacher de contrat, sur l'ensemble de la charge sous crête qui elle-même est décomposée en une charge conventionnée et en une autre qui ne l'est pas. L'organigramme de la charge sinistre nous récapitule les différentes classes de coût sinistre.



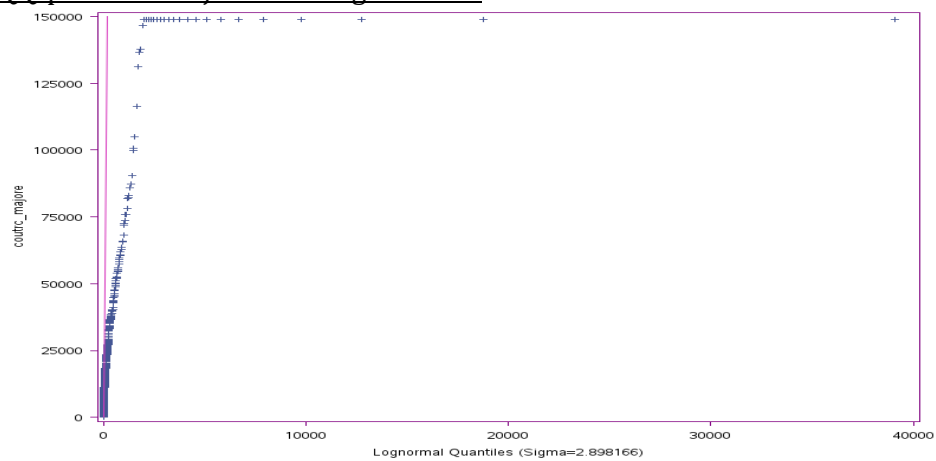
- La charge totale est la somme des coûts RC du portefeuille
- La charge sous crête correspond à la part du montant du sinistre inférieure au seuil d'écèlement
- La charge orpheline représente le coût auxquels l'on a pu rattacher de contrat
- La charge non conventionnée représente la part du coût non associé à la convention IDA, elle vaut pour chaque sinistre : charge sous crête-forfait IDA
 - Afin d'éviter une trop grande pénalisation des gros sinistres et une très faibles pénalisation des faibles sinistres et récréer un problème d'homogénéité l'on fixe un seuil minimum du montant de sinistre où les sinistres aux montants inférieurs à ce seuil se verront fixer à ce minimum et l'on conservera le montant complémentaire dans la clé de répartition
- La charge conventionnée correspond au montant du \mp forfait IDA selon que l'on rembourse une compagnie tierce (IDA passif) ou que l'on se fasse rembourser (IDA actif)

$$\text{clé de répartition} = \frac{\text{charge totale}}{\text{charge sous crête}}$$

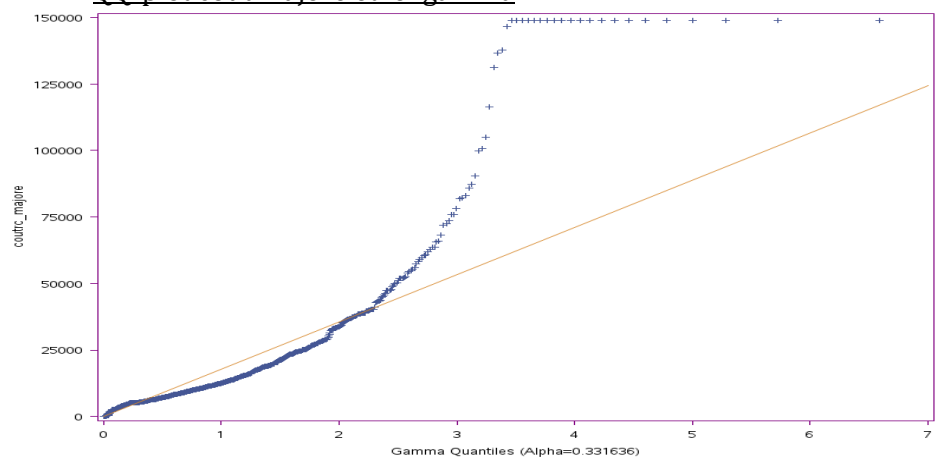
Nous modélisons à présent le coût majoré. Le diagramme quantile-quantile permet de vérifier la qualité de l'ajustement d'une loi à des données empiriques en comparant les quantiles théoriques aux quantiles des données étudiées.

Les graphiques ci-dessous représentent le QQ-plot de la distribution des coûts moyens majorés en fonction de la loi log-normale, gamma et Weibull.

QQ-plot coût majoré et loi log-normale



QQ-plot coût majoré et loi gamma



QQ-plot coût majoré et loi Weibull

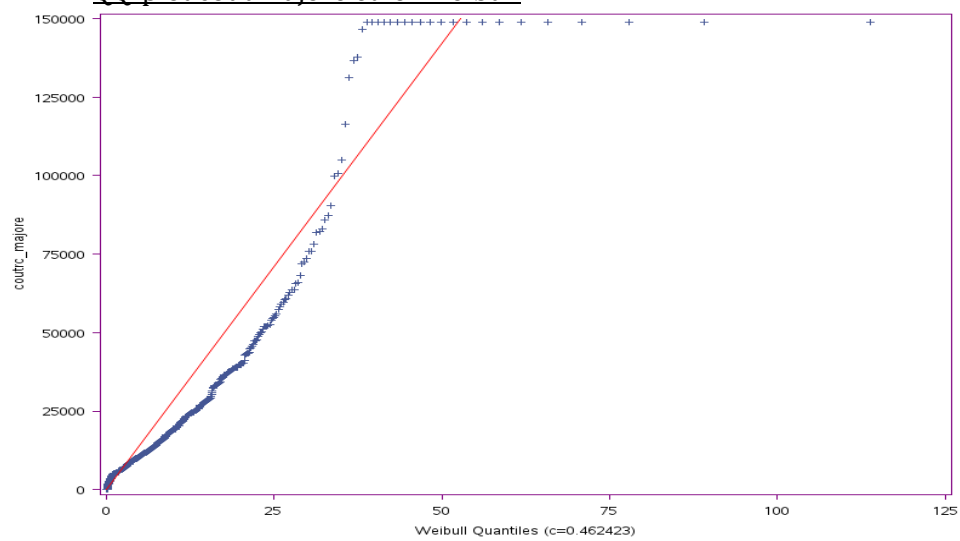


Figure 16 – QQ-plot du coût majoré et des lois comparées

Dans les trois cas l'on constate que la loi gamma est plus adaptée au modèle car les points sont bien mieux alignés que sur les autres graphiques. De plus, en analysant les densités ci-après, nous confirmons cette conjecture.

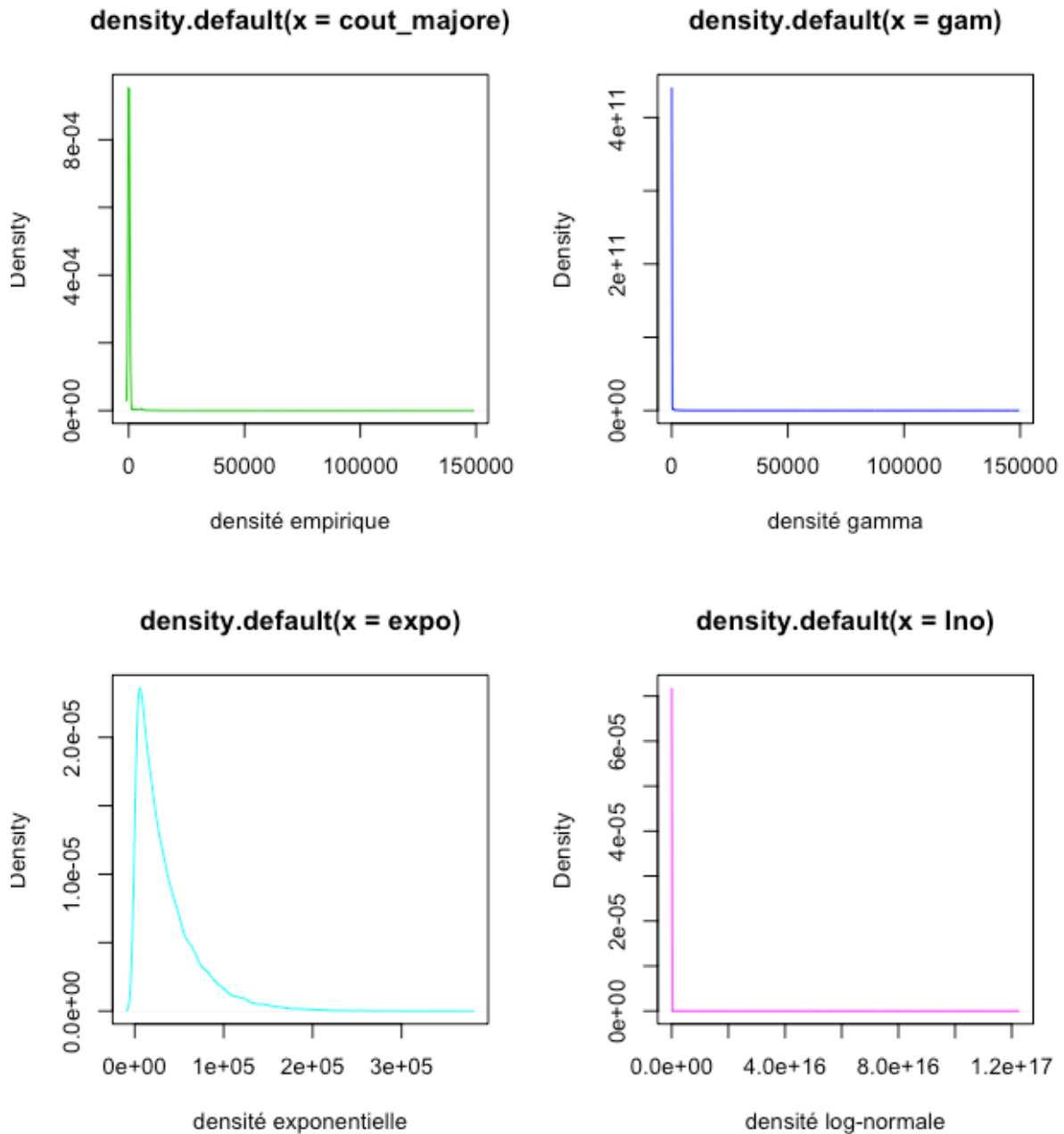


Figure 17 – Densités des lois candidates

Nous traçons également la fonction de répartition pour la loi gamma et celle des coûts moyens majorés. Nous observons que les deux courbes sont assez similaires.

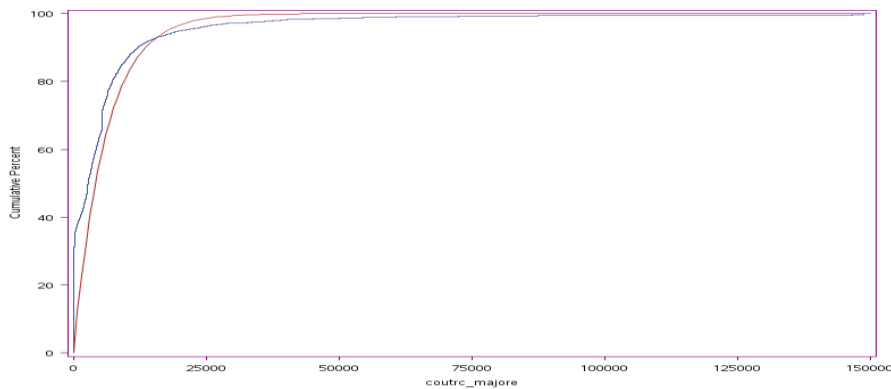


Figure 18 –

Fonction de répartition gamma et empirique

Pour obtenir notre modèle de coût, il nous faut sélectionner les variables qui seront à même d'estimer au mieux le coût moyen.

❖ Sélection des variables et interactions

Pour le modèle coût nous reprenons les 17 variables sélectionnées dans l'analyse univariée.

Pour la sélection des variables nous procédons par la méthode dite pas à pas.

○ Méthode pas à pas (*stepwise*)

Une alternative à la suppression manuelle de variables existe en utilisant les procédures pas à pas (*backward*, *forward*, *alternative*), qui suivant la méthode ne conservent que les variables qui ont un apport significatif à l'explication de la variable modélisée. Cependant, cette méthode ne se base que sur les données, alors il faut espérer et vérifier que les résultats convergent dans le même sens que nos idées.

En résumer les variables retenues pour notre modèle coût se trouve dans le tableau suivant.

1	classe
2	groupe
3	ancienneté du véhicule
4	zone
5	période RI
6	interruption d'assurance

○ Méthode manuelle

Comme la méthode *stepwise* se base sur les données, il est souhaitable de vérifier les résultats manuellement.

On part du modèle contenant toutes les variables explicatives. L'idée est de sélectionner chacune des variables et de comparer le modèle avec et sans la variable sélectionnée. Par

le principe de parcimonie, l'objectif est d'obtenir le modèle avec le moins possible de variables explicatives et de paramètres à estimer. D'un point de vue statistique, nous retenons les variables explicatives qui diminuent significativement la déviance avec une p-value de la statistique du χ^2 inférieure à 5%. Une p-value inférieure à 5% signifie que l'éviction de la variable brûlerait trop d'information. Nous regardons également l'impact de la variable sur les critères AIC, BIC, AICc.

A l'issue de cette méthode nous retiendrons les mêmes variables que pour les méthodes pas à pas.

Nous avons aussi testé toutes les interactions entre les variables fortement corrélées, cependant, aucune d'entre elles n'a été retenue.

Analyse des résidus

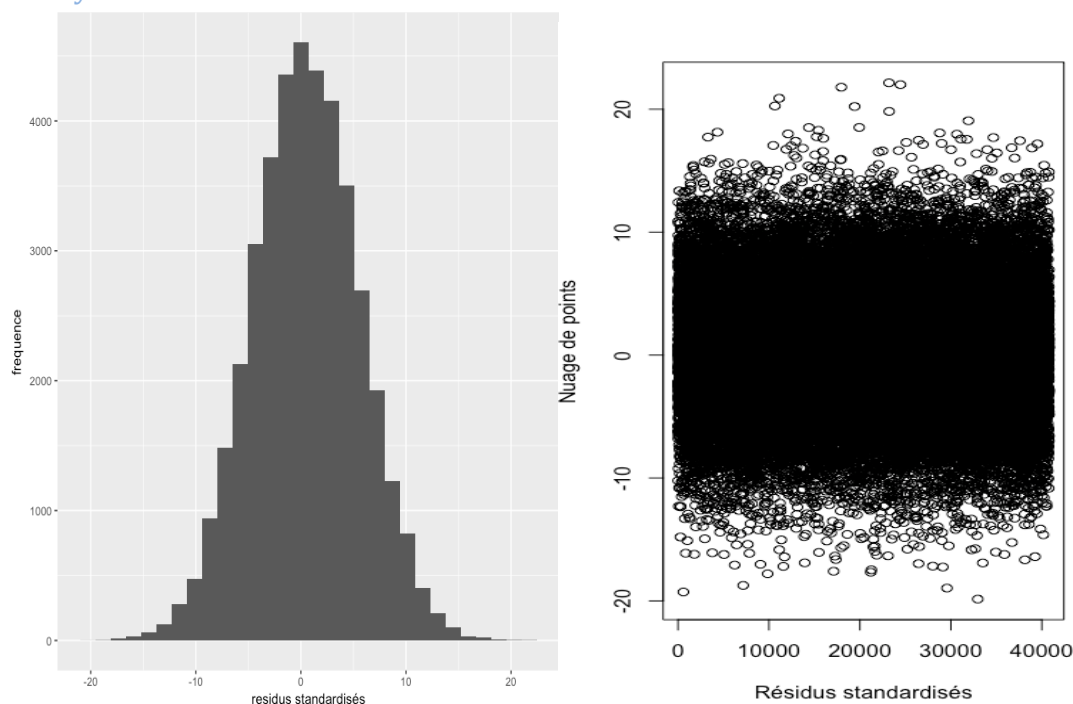


Figure 19 – Résidus

On observe des résidus centrés autour de 0 et qui ont une tendance normale. Les résidus sont répartis de façon homogène. Nous n'observons pas de points atypiques. On validera donc ce modèle.

Nous pouvons observer les coefficients dans l'annexe 4

2. Modèle fréquence

❖ Fonction de lien

Nous utiliserons comme fonction lien la fonction logarithmique étant donné le caractère multiplicatif de notre modèle coût x fréquence.

❖ Loi de probabilité de la fréquence

Nous allons modéliser la fréquence de la garantie Responsabilité Civile (RC) (indifféremment du type).

En actuariat, les principales lois pour la variable aléatoire de fréquence sont les lois de Poisson, binomiale et binomiale négative. Les lois des coûts correspondants sont alors appelées lois Poisson composée, binomiale composée et binomiale négative composée. Ces lois sont au cœur de la modélisation des risques en assurance IARD. Pour cette raison, nous les décrivons.

🚦 Loi de Poisson

La loi de Poisson est fondamentale dans la modélisation du nombre de sinistres pour les risques en assurance IARD. Elle constitue en quelque sorte la loi de base. L'espérance et la variance de la loi de Poisson sont égales. Cette propriété est appelée l'équidispersion.

Lorsque $N \sim \text{Pois}(\lambda)$, il découle que la v.a. S de même loi que les coûts obéissent à une loi Poisson composée avec les paramètres λ et F_X , notée $S \sim \text{PComp}(\lambda; F_X)$.

Les moments de S sont déterminés avec la preuve de l'annexe 2, en prenant :

$$\mathbb{E}(N) = \mathbb{V}(N) = \lambda.$$

$$\mathbb{E}(S) = \lambda \mathbb{E}(X)$$

$$\mathbb{V}(S) = \lambda \mathbb{V}(X) + \lambda \mathbb{E}(X)^2 = \lambda \mathbb{E}(X^2)$$

$$\mathbb{E}(S^3) = \lambda \mathbb{E}(X^3) + 3\lambda^2 \mathbb{E}(X^2) \mathbb{E}(X) = \lambda^3 \mathbb{E}(X)^3.$$

En développant le moment centré d'ordre 3 on obtient :

$$\mathbb{E}[(S - \mathbb{E}(S))^3] = \lambda \mathbb{E}(X^3)$$

À partir duquel on détermine l'expression suivante pour le coefficient d'asymétrie :

$$\gamma(X) = \frac{\mathbb{E}[(S - \mathbb{E}(S))^3]}{\mathbb{V}(S)^{3/2}} = \frac{\mathbb{E}(X^3)}{\sqrt{\lambda \mathbb{E}(X^2)^3}} > 0.$$

On constate que la distribution Poisson composée est toujours positivement asymétrique.

Loi binomiale

Quand $N \sim \text{Bin}(n, p)$, il en résulte que X obéit à une loi binomiale composée avec les paramètres n, q et F_X , notée $S \sim \text{BComp}(n, p; F_X)$.

Les moments de S sont déterminés avec la preuve de l'annexe 2, en prenant :

$$\mathbb{E}(N) = np, \mathbb{V}(N) = np(1 - p).$$

$$\mathbb{E}(S) = np\mathbb{E}(X).$$

$$\mathbb{V}(S) = np\mathbb{V}(X) + np(1 - p)\mathbb{E}(X)^2.$$

La loi binomiale comporte un intérêt plus limité que celle de poisson. En effet, en fixant $np = \lambda$ la loi binomiale tend vers la loi de poisson de paramètre λ lorsque n tend vers ∞ . Ce qui signifie que pour un échantillon assez grand (ce qui est très souvent le cas pour un portefeuille assuré), la loi binomiale peut se substituer en loi de poisson.

Cela se montre assez aisément :

Soit la v.a. $M_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$, on a :

$\lim_{n \rightarrow \infty} P_{M_n}(t) = \lim_{n \rightarrow \infty} \left(1 + \frac{\lambda(t-1)}{n}\right)^n = e^{\lambda(t-1)}$, qui correspond à la fonction génératrice d'une loi de poisson.

Loi binomiale négative

Quand $N \sim \text{Bin}(n, p)$, il en résulte que X obéit à une loi binomiale composée avec les paramètres r, q et F_X , notée $S \sim \text{BNComp}(r, p; F_X)$.

Les moments de S sont déterminés avec la preuve de l'annexe 2, en prenant :

$$\mathbb{E}(N) = r \frac{1-p}{p}, \mathbb{V}(N) = r \frac{1-p}{p^2}.$$

$$\mathbb{E}(S) = r \frac{1-p}{p} \mathbb{E}(X).$$

$$\mathbb{V}(S) = r \frac{1-p}{p} \mathbb{V}(X) + r \frac{1-p}{p^2} \mathbb{E}(X)^2.$$

Le moment centré d'ordre 3 de S est :

$$\mathbb{E}[(S - \mathbb{E}(S))^3] = r \frac{1-p}{p} \mathbb{E}(X^3) + 3r \left(\frac{1-p}{p}\right)^2 \mathbb{E}(X^2) + \mathbb{E}(X) + 2r \left(\frac{1-p}{p}\right)^3 \mathbb{E}(X)^3$$

Puisque le moment centré d'ordre 3 est toujours positif, on constate que la loi binomiale négative composée possède toujours une asymétrie positive.

- ✚ Comparaison des variances des trois lois composées. Ces lois sont caractérisées par les propriétés suivantes :

Loi binomiale	$E(N) > V(N)$
Loi de poisson	$E(N) = V(N)$
Loi binomiale négative	$E(N) < V(N)$

On compare les variances de la loi binomiale composée et de la loi binomiale négative composée à la variance de la loi composée. Pour la binomiale composée, on a :

$$V(S) = np(V(X) + (1 - p)E(X)^2) \leq npE(X^2) = E(X)E(X^2)$$

Pour la loi binomiale négative composée, on a :

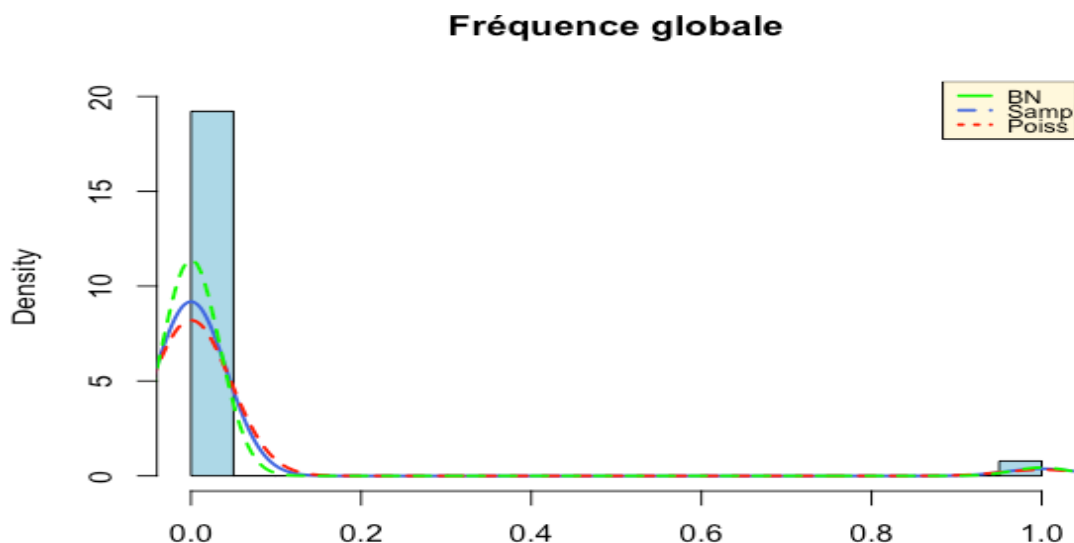
$$V(S) = r \frac{1-p}{p} \left(V(X) + \frac{1}{p} E(X)^2 \right) \geq r \frac{1-p}{p} E(X^2) = E(X)E(X^2)$$

Par conséquent, si les paramètres des trois lois de fréquence pour N sont fixés de telle sorte que leur espérance est identique et si la loi de X est identique pour les trois lois composées, on constate que :

$$\begin{cases} E(X^{BComp}) = E(X^{PComp}) = E(X^{BNComp}) \\ V(X^{BComp}) \leq V(X^{PComp}) \leq V(X^{BNComp}) \end{cases}$$

Nous ne conserverons que les lois, binomiale négative (BN) et de poisson (Pois) car comme démontré dans la section précédente la loi binomiale peut être remplacée par celle de Poisson, pour une taille d'échantillon importante.

Observons graphiquement les distributions théoriques de ces lois avec nos fréquences empiriques.



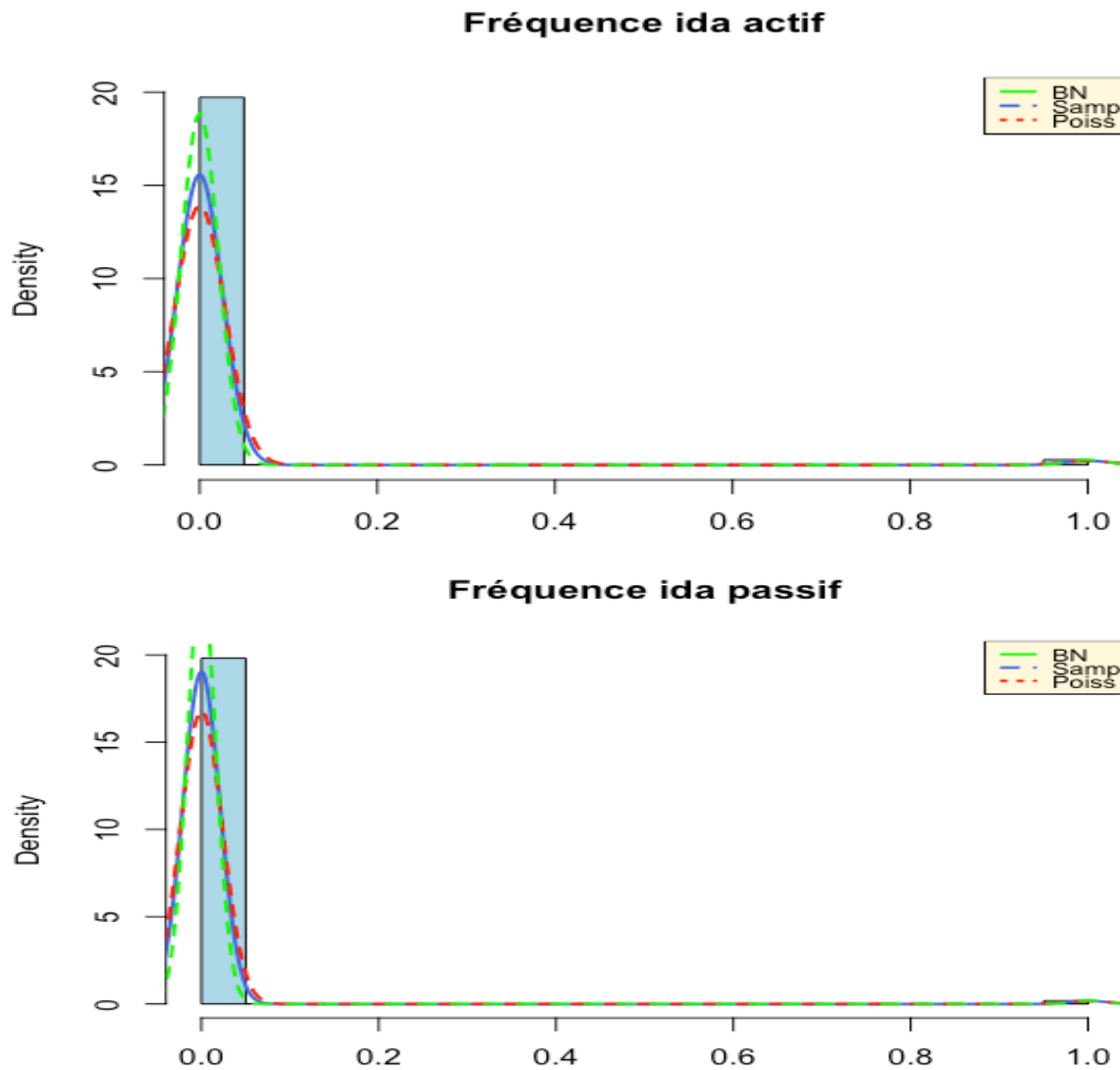


Figure 20 – Densités des fréquences empiriques et candidates

De plus la statistique du chi2 entre distribution observée et théorique nous pousse à considérer la loi de Poisson au vu des résultats du tableau ci-dessus.

	Chi-square	DF	Pr>ChiSq
Loi binomiale	23,4	3	<0,001
Loi négative			
Loi de poisson	15,07	3	<0,001

En sélectionnant les variables de la même manière analogue au modèle coût et en testant les interactions, pour chaque fréquence nous sélectionnons les variables suivantes :

Variables candidates	fréquence totale	Fréquence Ida passif	Fréquence Ida actif
age du conducteur	✓	✓	✓
ancienneté du permis	✓	✓	✓
ancienneté du véhicule	✓	✓	✓
classe	✓	✓	✗
fractionnement des paiements	✓	✓	✓
crm	✓	✓	✗
groupe	✓	✓	✓
parking	✓	✓	✓
zone	✓	✓	✓
nature*responsabilité	✓	✓	✓

Dans cette partie, nous nous sommes confronté à des problématiques liées à l'adéquation des lois de la fréquence et du coût du sinistre. Cette hypothèse forte nous a poussé à faire des ajustements, nous permettant de respecter des hypothèses de continuité. Nous avons modélisé quatre modèles dont un modèle de coût et trois modèles de fréquence, permettant de tenir compte de la particularité de certains coûts. Dans la partie suivante, nous modéliserons notre prime pure à l'aide d'un modèle ne nécessitant pas d'hypothèses de lois.

TROISIEME PARTIE

TARIFICATION AUTOMOBILE ET RESEAU DE NEURONES

Nous modélisons la prime pure par les algorithmes des réseaux de neurones. Ces algorithmes n'utilisent aucune hypothèse de loi sur les variables à expliquer. De plus, ceux-ci offrent la possibilité de modéliser directement la prime pure.

I. Tarification selon les réseaux de neurones

L'année 1943, marque le début des travaux concernant les réseaux neuronaux, avec l'article de McCulloch et Pitts intitulé *A logical calculus of the ideas imminent in nervous activity* (McCulloch et Pitts, 1943). A cette époque les recherches furent motivées par la modélisation des fonctions cérébrales. Elles réunissaient des biologistes, des psychologues, des neuroscientifiques. (Une application des réseaux de neurones artificiels MLP à la prévision du prix d'une action négociable-Antonio Fiordalisio).

Toutefois ce n'est qu'en 1980 qu'ils connurent un grand succès dans la modélisation mathématique avec un prestige venant de leur structure réputée imiter le cerveau humain, organiser en neurones et synapses. Leur réglage est délicat, cependant leur pouvoir prédictif parfois très élevé.

A. Introduction aux réseaux de neurones

Les réseaux de neurones se présentent comme des ensembles d'unités (encore appelées neurones formels ou nœuds) connectés entre elles par des synapses. Chaque unité étant activée en recevant une donnée et en renvoyant une autre. Ces ensembles d'unités sont organisés en niveaux appelés couches, et chaque couche envoie ses données à la couche suivante. Ils tirent leur mécanisme du fonctionnement du cerveau humain. En effet, le cerveau humain est organisé en neurones biologiques et en synapses. Le neurone biologique est composé d'un corps, d'un axone et de dendrites. Les synapses servent de liens entre deux neurones biologiques. Le tout forme un réseau interconnecté de neurones qui communiquent. L'information se transmet d'un neurone à l'autre par l'intermédiaire des synapses, qui ajustent leur comportement afin d'assurer le transfert.

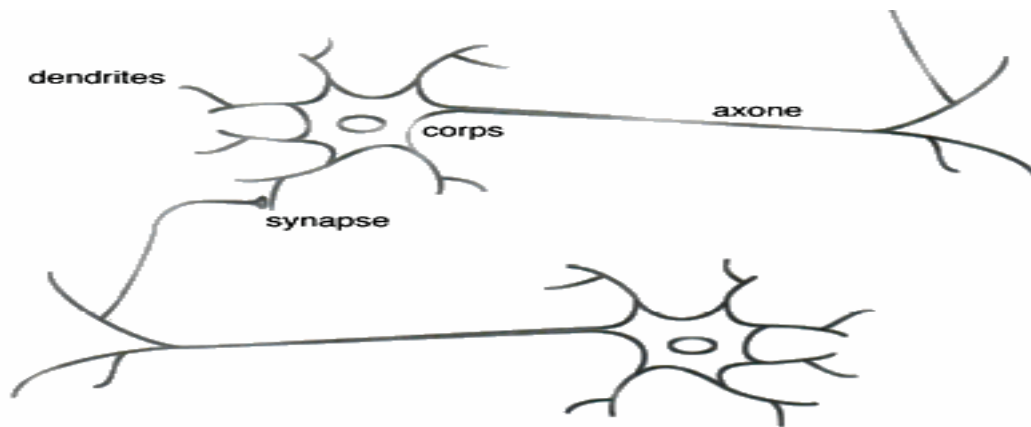


Figure 21 - Réseau de neurone biologique

Les réseaux de neurones connaissent un essor dans différents domaines, favorisés par le développement d'outils informatiques de plus en plus puissant. Les domaines d'application sont multiples, allant de l'industrie, à la télécommunication en passant par la finance.

Les réseaux de neurones peuvent être utilisés à des fins différentes. On a d'une part l'utilisation de ceux-ci en classification avec les réseaux de Kohonen et d'autre part, on les retrouve en méthodes de prédictions (perceptrons, réseaux à fonctions radiales de base). Les réseaux descriptifs sont dits à apprentissage non supervisé tandis que les réseaux prédictifs sont dits à apprentissage supervisé.

Nous étudions dans ce mémoire, les réseaux de neurones dans un but de prédictions.

Dans la suite, nous aborderons les notions théoriques et les fondamentaux nous permettant de mieux appréhender cet outil, nous verrons les différents types de réseaux de neurones prédictifs et enfin nous étudierons les différentes méthodes d'apprentissage d'un réseau de neurones.

B. Notions théoriques des réseaux de neurones

Un réseau de neurones est un ensemble complexe d'objets élémentaires, neurones formels, qui interagissent entre eux. Il est caractérisé par son type d'architecture et son mode d'apprentissage. Chaque composante joue un rôle essentiel dans le fonctionnement du réseau.

1. Neurone formel

Le neurone formel est caractérisé par des valeurs d'entrées, des poids associés à chacune de ces valeurs, une fonction de combinaison des poids et de leur valeur, une fonction d'activation (appelée aussi, fonction de transfert) et enfin une valeur de sortie.

Dans la plupart des ouvrages traitant des réseaux de neurones, on relie les valeurs d'entrées à leur poids par le biais d'une combinaison linéaire, toutefois, il en existe d'autres.

La valeur d'entrée du neurone formel est égale à la combinaison linéaire des valeurs d'entrées et des poids de connexion. Le neurone peut lui ajouter une valeur d'entrée seuil (biais) et lui applique une fonction d'activation. Enfin, il en résulte de cette dernière valeur, la valeur de sortie du neurone formel.

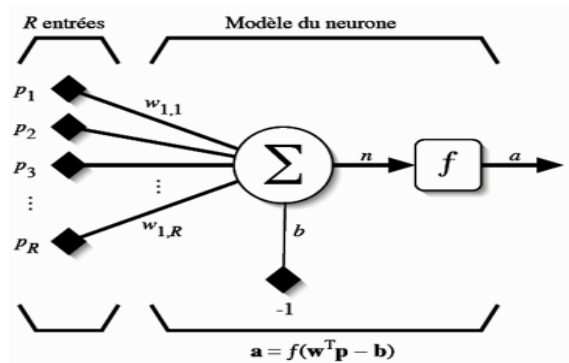


Figure 22 - Réseau de neurone formel

On considère un neurone formel à n entrées n_1, n_2, \dots, n_k , et p_1, p_2, \dots, p_k , les poids associés à chaque valeur d'entrée et α le biais du neurone. La valeur d'entrée dans le neurone, en considérant la fonction combinaison linéaire, est la suivante :

$$\alpha + \sum_{i=1}^k n_i p_i$$

On applique ensuite à la valeur d'entrée, la fonction d'activation f , ce qui nous donne comme valeur de sortie pour le neurone formel :


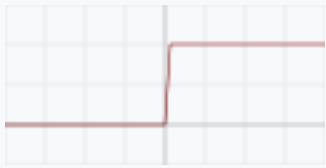
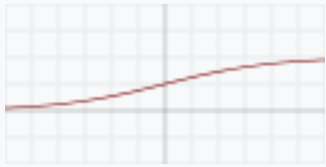
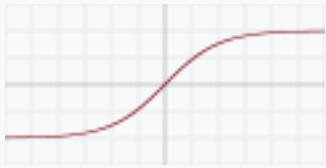
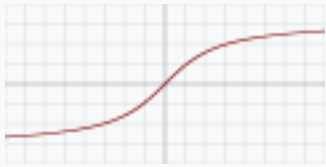
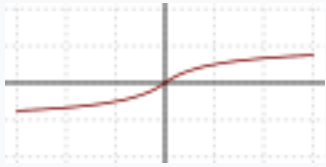

$$f\left(\alpha + \sum_{i=1}^k n_i p_i\right)$$

2. Fonction d'activation

La fonction d'activation joue un rôle essentiel dans la caractérisation du neurone formel. Elle permet la transformation d'une combinaison affine des valeurs d'entrées et des poids associés.

La fonction d'activation doit être régulière et respecter certaines conditions de dérivabilité, car on le verra par la suite, l'apprentissage nécessite une recherche d'optimum. Généralement, elle est strictement croissante et bornée.

Il en existe plusieurs types, on peut citer : le pas unitaire, identité, la linéaire seuillée, la sigmoïde (nommée la courbe en S), la normale, l'inverse, l'exponentielle, la carrée, les stochastiques ...

Nom	Graphe	Équation
Identité		$f(x) = x$
Marche/ Heaviside		$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$
Logistique (ou marche douce ou sigmoïde)		$f(x) = \frac{1}{1+e^{-x}}$
Tangente Hyperbolique (TanH)		$f(x) = \frac{2}{1+e^{-2x}} - 1$
Arc Tangente (ArcTan ou Tan ⁻¹)		$f(x) = \tan^{-1}(x)$
Signe doux ⁶		$f(x) = \frac{x}{1+ x }$
Unité de Rectification Linéaire (ReLU) ⁷		$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases}$


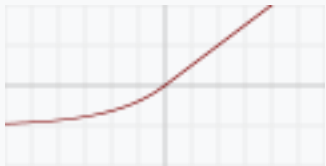
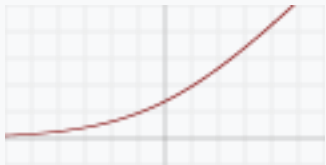

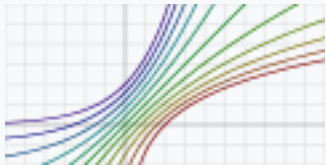
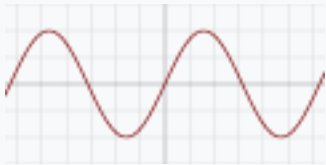
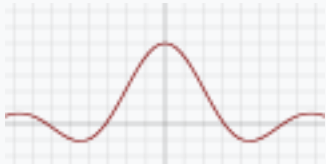
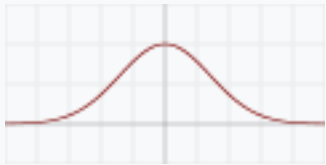
Unité de Rectification Linéaire Paramétrique (PReLU) ⁸		$f(x) = \begin{cases} \alpha x & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases}$
Unité Exponentielle Linéaire(ELU) ⁹		$f(x) = \begin{cases} \alpha(e^x - 1), & x < 0 \\ x, & \text{sinon} \end{cases}$
Unité de Rectification Linéaire Douce (SoftPlus) ¹⁰		$f(x) = \log_e(e^x + 1)$
Identité courbée		$f(x) = \frac{\sqrt{(1+x^2)}-1}{2} + 1$
Exponentielle douce paramétrique (soft exponential) ¹¹		$f(x, \alpha) = \begin{cases} -\frac{\log_e(1-\alpha(x+\alpha))}{\alpha}, & \alpha < 0 \\ x, & \alpha = 0 \\ \frac{e^x+1}{\alpha} + \alpha, & \alpha > 0 \end{cases}$
<u>Sinusoïde</u>		$f(x) = \sin(x)$
<u>Sinus cardinal</u> (Sinc)		$f(x) = \begin{cases} 1 & \text{pour } x = 0 \\ \frac{\sin(x)}{x} & \text{pour } x \neq 0 \end{cases}$
Fonction <u>Gaussienne</u>		$f(x) = e^{-x^2}$

Tableau 10- Représentation de quelques fonctions d'activation

Source : Wikipédia

3. Réseau de neurones

Le réseau de neurones est un ensemble de neurones formels reliés entre eux par des poids et organisés en couches successives. Il est constitué d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie. L'information se déplace de manière séquentielle couche par couche, entre chaque neurone de chaque couche. Par l'intermédiaire des couches cachées. Le type de propagation de cette information est lié à l'architecture du neurone.

C. Type d'architecture des réseaux de neurones

Il existe plusieurs types de réseau de neurones. On peut les classer en deux grandes familles :

- Les réseaux non bouclés ou « feed forward »
- Les réseaux récurrents

1. Réseau non bouclé

Les réseaux non bouclés sont des réseaux à propagation directe. L'information se propage de la couche i vers la couche j , où $i < j$. Nous nous intéresserons à quelques-uns.

❖ Perceptron simple

Le perceptron simple ou monocouche est un réseau composé d'un seul neurone et agit par la fonction d'activation suivante :

$$f = \begin{cases} 1 & , si \ y > 0 \\ 0 \text{ (ou } -1), & sinon \end{cases}$$

❖ Perceptron multicouche PMC

Le perceptron multicouche (PMC) est un ensemble de perceptrons repartis en couches successives [figure 11].

Tous les neurones des couches cachées sont munis de la fonction d'activation σ sigmoïde.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Le neurone de la couche cachée (Ils peuvent être plusieurs) est muni quant à lui (chacun en sera muni) de la fonction d'activation identité.

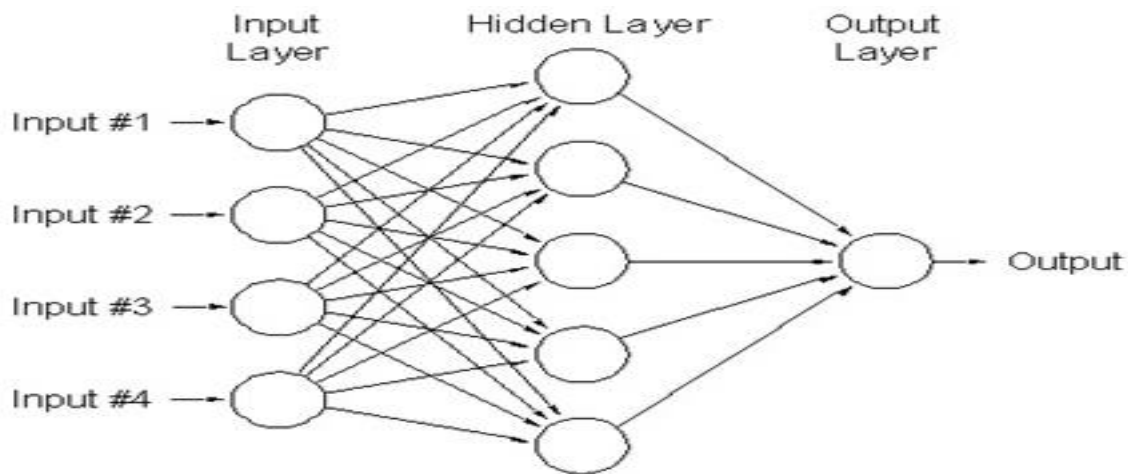


Figure 23 - Un perceptron à une couche cachée

❖ Réseau « random vector functional »

Random vector functional link neural network (RVFLNN) est un modèle de perceptron multicouche où il existe des connexions supplémentaires entre la couche d'entrée et la couche de sortie. La fonction utilisée pour les couches cachées et la couche de sortie est la fonction sigmoïde.

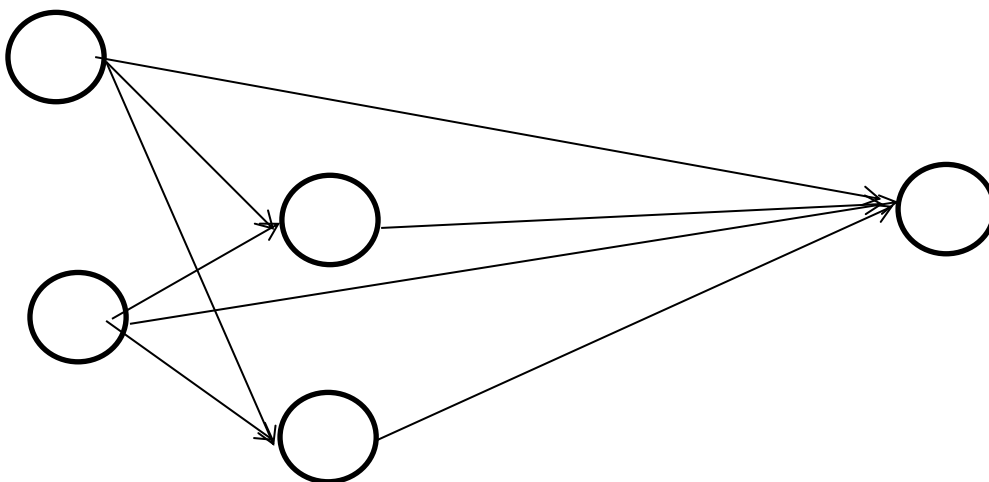


Figure 24 - RVFLNN à une couche cachée

❖ Réseau à base de fonction radiale (RBF)

Le réseau de neurone à base de fonction radiale est un réseau à trois couches. Une couche d'entrée, une couche cachée et une couche de sortie. La particularité de ce réseau est que les

neurones de la couche cachée sont munis d'une fonction de transfert H_i à base radiale. Il y'a plusieurs formes de fonctions à base radiale. La plus connus est la fonction gaussienne.

$$H_i = \exp\left(-\frac{\|x_i - \mu_i\|^2}{2\sigma_i}\right)$$

où $\| \cdot \|$ représente la division euclidienne, x_i, μ_i, σ_i , représentent respectivement le vecteur d'entrée, le centre et le rayon d'influence du neurone i sur la couche cachée.

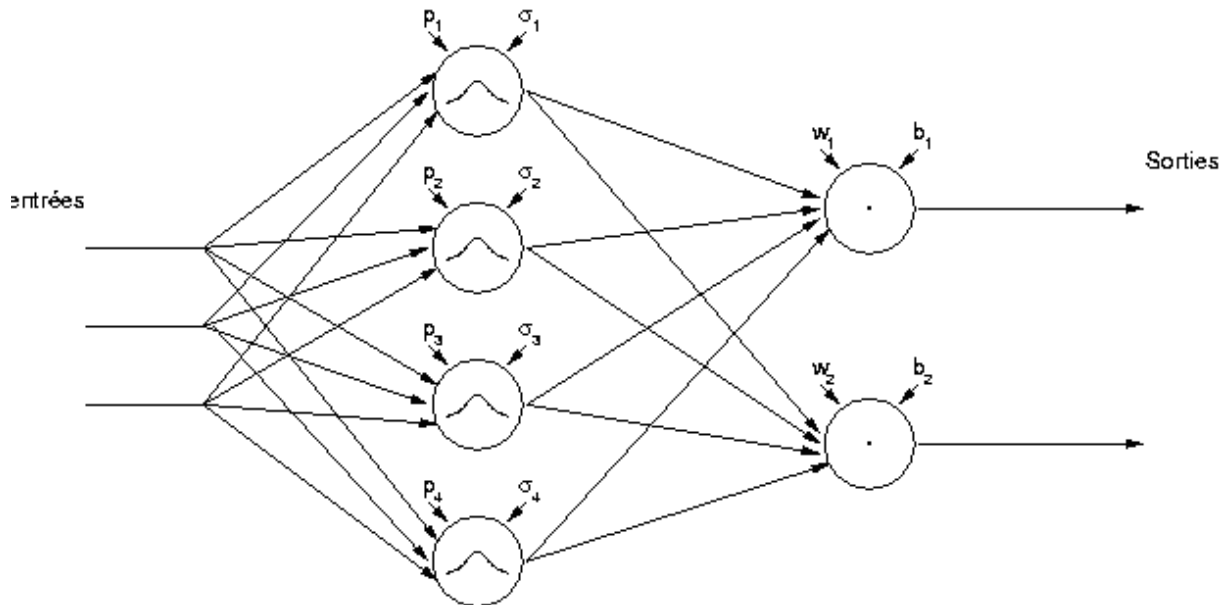


Figure 25 - Réseau à base de fonction radiale

❖ Carte de Kohonen

Une carte de Kohonen est une carte d'auto-organisatrice en deux couches, une couche d'entrée et une couche de sortie. La deuxième couche aussi appelée couche compétitive est en deux dimensions. Chaque neurone d'entrées est connecté à l'ensemble des neurones de sortie par le poids w_{ij} . La carte de Kohonen est une carte à apprentissage non-supervisé, à partir d'une base d'apprentissage constituée seulement des éléments d'entrées, elle permet de regrouper ces entrées en classe.

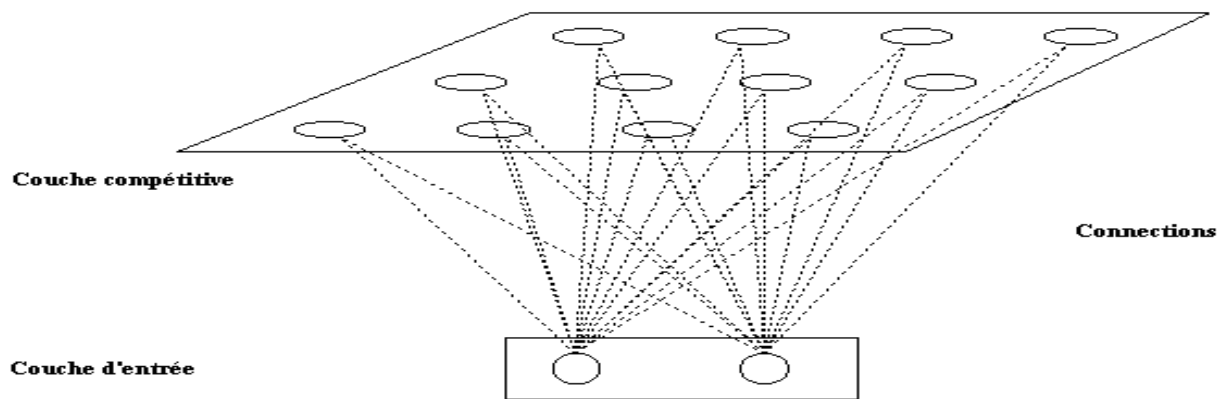


Figure 26 - Architecture d'une carte de Kohonen

2. Réseau bouclé

Les réseaux de neurones bouclés ou récurrents sont caractérisés par une topologie qui autorise toutes les connexions et par le fait que chaque neurone possède sa propre dynamique de connexion. Tous les neurones sont liés par une liaison pondérée et également par une liaison vers eux-mêmes.

Ces réseaux ont une architecture circulaire, il n'est pas possible de distinguer les couches successives pour la propagation de l'information.

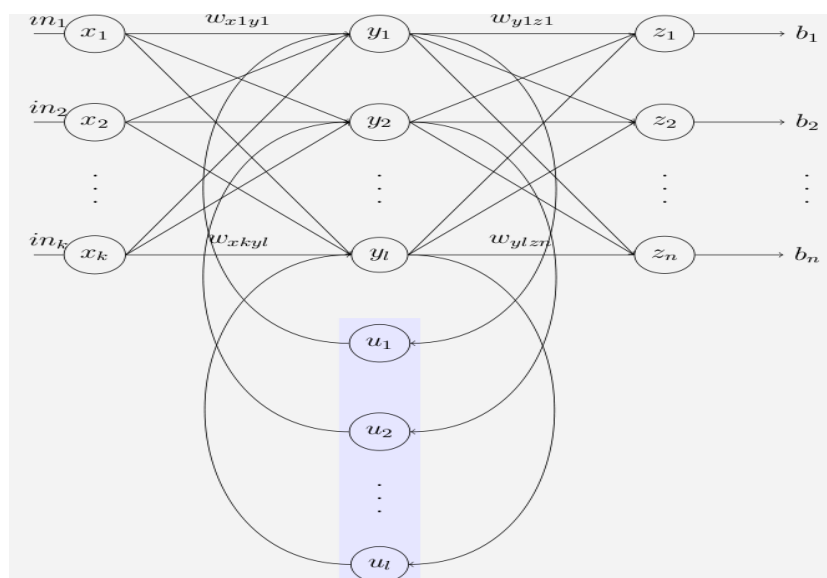


Figure 27- Exemple de réseau de neurones récurrent

L'évolution de l'état d'activation des neurones dans le temps est régie par l'équation différentielle suivante :

$$\tau_i \frac{\partial y_i}{\partial t} = -y_i + \sum_{j=1}^k \omega_{j,i} \sigma_j + I_i$$

Avec τ est une constante de temps, y_i l'activation du neurone i , k le nombre total de neurones $\omega_{j,i}$ le poids de l'arc (j, i) , σ_j l'activation du neurone j modifié pour suivre une forme de sigmoïde et I_i la valeur d'entrée appliquée au neurone i .

L'équation de l'activation des neurones suivant une forme de sigmoïde :

$$\sigma_i = \frac{1}{1 + e^{-(y_i + b_i)}}$$

Avec respectivement y_i , b_i , l'activation du neurone i et le biais.

Chaque neurone possède sa propre constante de temps τ , ils évoluent selon une fréquence spécifique.

Nous constatons qu'il existe diverses architectures pour décrire un réseau de neurone. Ces architectures plus ou moins complexes répondent à des besoins différents. En matière de prédiction statistique l'idée étant de reproduire une dépendance fonctionnelle entre les variables explicatives et la ou les variable(s) expliquée(s), nous pouvons limiter notre étude à un perceptron multicouche à une seule couche cachée comme l'indique la propriété d'approximation universelle.

Propriété d'approximation universelle: Un perceptron multicouche doté d'une couche cachée, avec un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision voulue.

Dans la suite de ce mémoire, nous focaliserons notre attention sur les méthodes adaptées aux perceptrons.

D. Algorithme d'apprentissage

Il existe plusieurs types d'algorithmes d'apprentissage. Ils ont pour rôle, dans le cadre d'un réseau de neurones, d'ajuster les poids de connexion de sorte à ce qu'au fil des exemples qui lui sont donnés, le réseau affine ses prédictions. Nous parlerons des premiers algorithmes d'apprentissage, qui sont appliqués sur un perceptron simple (perceptron linéaire). Ceux-ci, nous permettront de mieux appréhender les algorithmes utilisés pour le Perceptron Multi Couche.

1. Algorithme d'apprentissage par correction d'erreur

Etant donné un échantillon d'apprentissage A de $E \times S$, avec $E = \mathbb{R}^n$ ou $\{0,1\}^n$ et $S = \mathbb{R}$ ou $\{0,1\}$, c'est-à-dire un ensemble d'exemples dont les descriptions sont sur n attributs réels

(respectivement binaires) et la cible réelle (respectivement binaire). Il s'agit de trouver un algorithme qui infère à partir de S un perceptron qui prédit correctement les éléments de A au vu de leurs descriptions, et au mieux sinon.

L'algorithme d'apprentissage peut être décrit succinctement de la manière suivante :

On initialise le poids du perceptron à des valeurs quelconques. A chaque fois que l'on présente un exemple, on ajuste les poids selon que le perceptron l'ait correctement classé ou non. L'algorithme s'arrête lorsque tous les exemples lui ont été présentés sans aucune modification des poids.

Dans la suite, nous désignerons par \vec{x} une description qui sera un élément de E . La $i^{\text{ème}}$ composante de \vec{x} sera notée x_i . Un échantillon A est un ensemble de couples (\vec{x}, c) où c est la valeur à prédire de \vec{x} . Lorsqu'il sera utile de désigner un élément de A , nous noterons (\vec{x}^j, c^j) le $j^{\text{ème}}$ élément de A . x_i^j désigne donc la $i^{\text{ème}}$ composante du vecteur \vec{x}^j . Si \vec{x}^j est présenté en entrée au perceptron, nous noterons o^j la sortie calculée par le perceptron.

L'algorithme d'apprentissage par correction d'erreur du perceptron linéaire à seuil est :

Algorithme par correction d'erreur :

Entrée : Un échantillon A de $E \times S$,
initialisation aléatoire des poids p_i . Pour i entre 0 et n

Répéter

Pr

endre un exemple (\vec{x}, c) dans A
calculer la sortie o du perceptron pour l'entrée \vec{x}
-- Mise à jour des poids--

Pour i de 0 à n

$$p_i \leftarrow p_i + (c - o)x_i$$

FinPour

FinRépéter

Sortie : Un perceptron P défini par (p_1, p_2, \dots, p_n)

Le processus d'apprentissage du perceptron, est un processus de correction d'erreur, puisque les poids ne sont pas modifiés lorsque la sortie attendue c est égale à la sortie calculée o par le perceptron courant. Etudions les modifications apportées au poids pour un perceptron linéaire à sortie binaire.

- $o = 0$ et $c = 1$, cela signifie que le perceptron n'a pas assez pris en compte les neurones actifs de l'entrée (c'est-à-dire les neurones ayant une entrée à 1) ; dans ce cas $p_i \leftarrow p_i + x_i$; l'algorithme ajoute de la valeur de l'entrée au poids synaptique (renforcement).
- $o = 1$ et $c = 0$, alors $p_i \leftarrow p_i - x_i$, l'algorithme retranche la valeur de l'entrée au poids synaptique (inhibition).

Certains éléments importants ont été volontairement imprécis. En premier lieu, il faut préciser comment se fait le choix des éléments de A : le sont-ils tirés aléatoirement ? Dans un ordre précis ? Doivent-ils être tous présentés ? Le critère d'arrêt de la boucle principale n'est pas défini : après

un certain nombre d'étapes? Lorsque tous les exemples ont été présentés ? Lorsque les poids ne sont plus modifiés pendant un certain nombre de temps ? Nous reviendrons sur toutes ces questions par la suite. Tout d'abord examinons cet algorithme à l'aide d'un exemple.

Exemple : Apprentissage du OU

Les descriptions appartiennent à $\{0,1\}^2$, les entrées du perceptron appartiennent à $\{0,1\}^3$, la première composante correspond à l'entrée x_0 et vaut toujours 1, les deux composantes suivantes correspondent aux variables x_1 et x_2 .

On suppose qu'à l'initialisation les poids suivant ont été choisis : $p_0 = 0, p_1 = 1, p_2 = -1$

Etape	p0	p1	p2	Entrée	y	o	c	p0	p1	p2
init								0	1	-1
1	0	1	-1	100	0	0	0	$0+0 \times 1$	$1+0 \times 0$	$-1+0 \times 0$
2	0	1	-1	101	-1	0	1	$0+1 \times 1$	$1+1 \times 0$	$-1+1 \times 1$
3	1	1	0	110	2	1	1	1	1	0
4	1	1	0	111	2	1	1	1	1	0
5	1	1	0	100	1	1	0	$1+(-1) \times 1$	$1+(-1) \times 0$	$0+(-1) \times 0$
6	0	1	0	101	0	0	1	$0+1 \times 0$	$1+1 \times 0$	$0+1 \times 1$
7	1	1	1	110	2	1	1	1	1	1
8	1	1	1	111	3	1	1	1	1	1
9	1	1	1	100	1	1	0	$1+(-1) \times 1$	$1+(-1) \times 0$	$1+(-1) \times 0$
10	0	1	1	101	1	1	1	0	1	1

Tableau 11- valeurs des poids à chaque itération

Critiques: L'algorithme par correction d'erreur converge à condition que l'échantillon d'apprentissage soit linéairement séparable. De plus, il n'existe aucun moyen de savoir que celui-ci n'est pas convergent.

Le but des sections suivantes est de présenter des algorithmes qui conduisent à des solutions plus ou moins robustes selon que l'échantillon d'apprentissage soit linéairement séparable ou non.

2. Algorithme par descente du gradient

Plutôt que d'avoir une solution qui prédit avec exactitude tous les exemples, il s'agira de calculer une erreur et de la minimiser.

Définition : Soit un perceptron linéaire P ayant pour entrée un vecteur \vec{x} de n valeurs (x_1, x_2, \dots, x_n) et calcule une sortie o . Un perceptron est défini par la donnée d'un vecteur \vec{w} de n constantes : les coefficients synaptiques (w_1, w_2, \dots, w_n) . La sortie o est définie par :

$$o = \vec{x} \cdot \vec{w} = \sum_{i=1}^n x_i w_i$$

L'erreur du perceptron P, défini par \vec{w} sur l'échantillon d'apprentissage A, est définie en utilisant la fonction d'erreur quadratique par :

$$E(\vec{w}) = \frac{1}{2} \sum_{(\vec{x}, c) \in A} (c - o)^2$$

L'erreur mesure donc l'écart entre la sortie attendue et la sortie estimée sur l'échantillon complet. On remarque que $E(\vec{w}) = 0$ si et seulement si le perceptron prédit correctement l'échantillon complet.

Pour A fixé, le problème est de trouver le vecteur w qui minimise E . Un moyen de minimiser une fonction est d'utiliser la méthode du gradient. Celle-ci est rappelée ci-dessous.

Méthode du gradient

Soit f une fonction d'une variable réelle à valeur réelle, suffisamment dérivable dont on cherche le minimum. La méthode du gradient cherche à construire une suite réelle qui tend vers le minimum telle que $f(x_{n+1}) \leq f(x_n)$. Pour cela on part d'une valeur quelconque x_0 , et l'on construit une suite récurrente par : $\forall n > 0, x_{n+1} = x_n + Dx_n$ où $Dx_n = -ef'(x_n)$ où e est une valeur « bien » choisie. D'après la formule de Taylor, on a : $f(x_{n+1}) = f(x_n - ef'(x_n)) \approx f(x_n) - e(f'(x_n))^2$. On voit que sous réserve de la correction de l'approximation : $f(x_{n+1}) \leq f(x_n)$

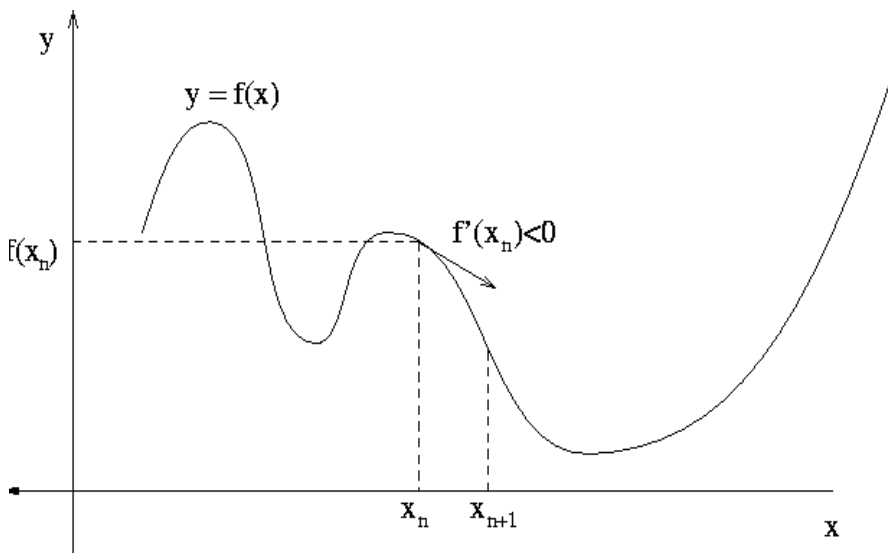


Figure 28 - Méthode du gradient

On constate que, plus la pente est grande, plus l'écart entre x_{n+1} et x_n est important. On peut décider d'arrêter l'itération lorsque cette pente est « suffisamment » faible.

Critiques :

- * Le choix de e est empirique.

- * Si e est trop petit le nombre d'itérations peut être élevé.
- * Si e est trop grand, les valeurs de la suite risquent d'osciller autour du minimum sans jamais converger.
- * Rien ne garantit que le minimum trouvé soit un minimum global.

Algorithme

E est une fonction de n variables w_i . La méthode du gradient a été rappelée pour une fonction à une seule variable réelle mais, elle peut être étendue à une fonction à plusieurs variables réelles. Pour mettre en œuvre cette méthode appliquée à la fonction erreur quadratique, nous évaluons tout d'abord la dérivée partielle de E par rapport à w_i . On a :

$$\begin{aligned}\frac{\partial E}{\partial w_i} &= \frac{1}{\partial w_i} \frac{1}{2} \sum_{(\vec{x}, c) \in A} (c - o)^2 \\ &= \frac{1}{2} \sum_{(\vec{x}, c) \in A} 2(c - o) \frac{1}{\partial w_i} (c - o) \\ &= \sum_{(\vec{x}, c) \in A} (c - o) \frac{1}{\partial w_i} (c - \vec{x} \cdot \vec{w}) \\ \frac{\partial E}{\partial w_i} &= \sum_{(\vec{x}, c) \in A} (c - o)(-x_i)\end{aligned}$$

L'application de la méthode du gradient nous invite à modifier les poids w_i après une présentation complète de A , d'une quantité Dw_i définie par :

$$Dw_i = -e \frac{\partial E}{\partial w_i} = e \sum_{(\vec{x}, c) \in A} (c - o)x_i$$

Algorithme par descente du gradient :

Entrée : Un échantillon A de $E \times S$; e
initialisation aléatoire des poids w_i pour i entre 1 et n

Répéter

Pour tout i , $Dw_i \leftarrow 0$ **FinPour**

Pour tout exemple (\vec{x}^j, c^j) de A calculer la sortie o^j

Pour tout i $Dw_i \leftarrow Dw_i + e(c^j - o^j)x_i^j$ **FinPour**

FinPour

Pour tout i $w_i \leftarrow w_i + Dw_i$

FinRépéter

Sortie : Un perceptron P défini par (w_1, w_2, \dots, w_n)

La fonction erreur quadratique ne possède qu'un minimum (la surface est un paraboloïde figure 17). L'algorithme précédent est assuré de converger vers un minimum pour un e bien choisi et suffisamment petit, même si l'échantillon d'entrée n'est pas linéairement séparable. Si e est trop grand on risque d'osciller autour du minimum. Pour cette raison, une modification classique est de diminuer graduellement le e en fonction du nombre d'itérations.

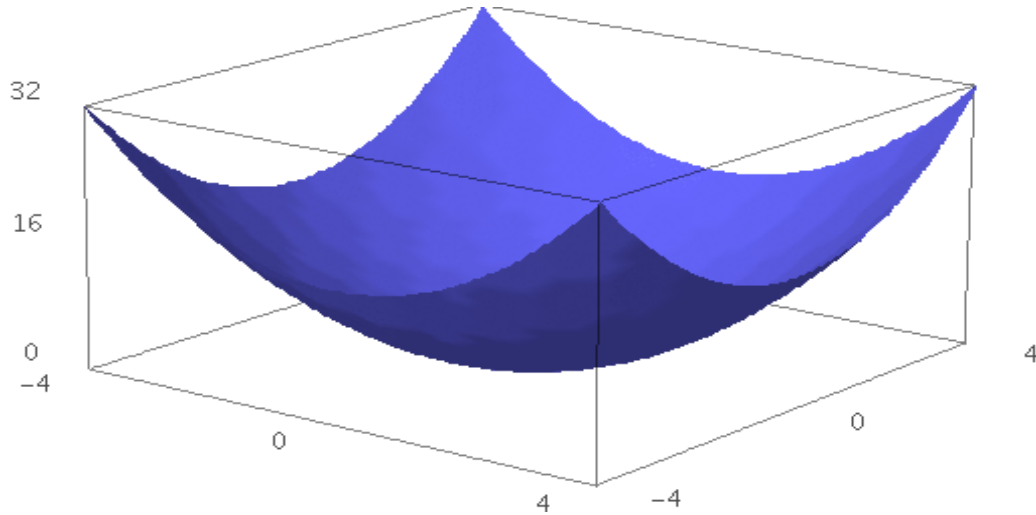


Figure 29- paraboloïde en dimension 3

Critique : La convergence peut-être très lente et que chaque étape nécessite le calcul sur tout l'apprentissage.

Pour cette dernière raison que l'algorithme suivant a été mis au point.

3. Algorithme d'apprentissage de Widrow-Hoff

Cet algorithme est une variante de l'algorithme précédent. Au lieu de calculer les variations des poids en sommant sur tous les exemples de A , Il faut modifier les poids à chaque présentation d'exemples. La règle de modification des poids devient :

$$Dw_i = e(c^j - o^j)x_i^j$$

Cette règle est appelée règle delta, ou règle Adaline ou encore règle Widrow-Hoff d'après le nom de ses inventeurs. L'algorithme s'écrit alors :

Algorithme de Widrow-Hoff :

Entrée : Un échantillon A de $E \times S$; e ,
initialisation aléatoire des poids w_i . Pour i entre 0 et n

Répéter

Prendre un exemple (\vec{x}, c) dans A
calculer la sortie o du perceptron pour l'entrée \vec{x}

-- Mise à jour des poids--

Pour i de 0 à n

$$w_i \leftarrow w_i + e(c - o)x_i$$

FinPour

FinRépéter

Sortie : Un perceptron P défini par (w_1, w_2, \dots, w_n)

En général, on parcourt l'échantillon dans un ordre prédéfini. Le critère d'arrêt généralement choisi est : pour un passage complet de l'échantillon, toutes les modifications de poids sont en dessous d'un seuil prédéfini.

Au coefficient e près dans la règle de modification des poids, on retrouve l'algorithme d'apprentissage par correction d'erreur.

Pour l'algorithme de Widrow-Hoff, il y'a correction chaque fois que la sortie totale est différente de la sortie attendue. L'avantage de cet algorithme par rapport à l'algorithme par correction d'erreur, est qu'il réussit à prédire même si l'échantillon d'apprentissage n'est pas linéairement séparable.

L'algorithme de Widrow-Hoff s'écarte de l'algorithme du gradient sur un point important : la modification des poids se fait après présentation de chaque exemple en fonction de l'erreur locale et non de l'erreur globale. Néanmoins, la diminution de l'erreur en un point, pourrait être compensée par une augmentation de l'erreur pour les autres points. La justification empirique de cette manière de procéder est commune à toutes les méthodes adaptatives. Le champ d'application des méthodes adaptatives est justement l'ensemble des problèmes pour lesquels des ajustements locaux vont finir par converger vers une solution globale.

L'algorithme de Widrow-Hoff est très souvent utilisé en pratique et donne de bons résultats. La convergence est, en général, plus rapide que par la méthode du gradient. Il est fréquent pour cet algorithme de faire diminuer la valeur de e en fonction du nombre d'itérations comme pour l'algorithme du gradient.

Il est bien évident que la plupart des problèmes d'apprentissage qui se posent naturellement ne peuvent pas être résolus par des méthodes aussi simples : il n'y a que très peu d'espoir que les exemples « naturels » se répartissent « sagement » de part et d'autre d'un hyperplan. Une manière de résoudre cette difficulté serait, soit de mettre au point des séparateurs non-linéaires, soit, ce qui revient à peu près au même, de complexifier l'espace de représentation de manière à linéariser le problème initial. C'est ce que permettent de faire les réseaux multicouches que nous étudions maintenant par le biais de l'algorithme de rétropropagation du gradient.

4. Algorithme de rétropropagation du gradient

Le principe de l'algorithme est, comme dans le cas du perceptron linéaire, de minimiser une fonction d'erreur. Il s'agit par la suite de calculer la contribution à cette erreur de chacun des poids synaptiques. C'est l'étape la plus difficile de l'algorithme. En effet, chacun des poids influe sur le neurone correspondant, mais, la modification pour ce neurone va influencer sur tous les neurones des couches suivantes. Ce problème est parfois désigné sous le nom de « Credit Assignment Problem ».

Soit un perceptron multicouche défini par une architecture à n entrées et p neurones de sorties. Soit \vec{w} le vecteur poids synaptiques associés à tous les liens du réseau. L'erreur du PMC sur un échantillon d'apprentissage A d'exemple (\vec{x}, c) est définie par :

$$E(\vec{w}) = \frac{1}{2} \sum_{(\vec{x}, c) \in A} \sum_{k=1}^p (c_k - o_k)^2$$

o_k est la $k^{ième}$ composante du vecteur sortie o (contient tous les neurones de la couche cachée, ainsi que la sortie pour un PMC à une seule couche cachée)

L'idée est donc de trouver le vecteur \vec{w} qui minimise E . Cependant de la même façon que la règle de Widrow-Hoff, plutôt que de chercher à minimiser l'erreur globale sur l'échantillon A , nous cherchons à minimiser l'erreur à chaque présentation d'un exemple de l'échantillon.

L'erreur pour un exemple est définie par :

$$E(\vec{w}) = \frac{1}{2} \sum_{k=1}^p (c_k - o_k)^2$$

E est une fonction de poids synaptique, sa dérivée doit être exprimée en fonction de ces poids. Les calculs qui suivent sont faciles, la seule complexité réside dans la notation et des indices utilisés.

Nous utilisons la notation suivante :

- * Chaque cellule est définie par un indice,
- * le réseau comprend p cellules de sorties,
- * si i est l'indice d'une cellule de sortie, c_i est la sortie attendue pour cette cellule sur l'entrée \vec{x} ,
- * w_{ij} est le lien synaptique entre la cellule j vers la cellule i , ce qui signifie qu'elles se trouvent sur deux couches successives d'après l'architecture,
- * x_{ij} est l'entrée associée au lien entre la cellule j vers la cellule i ,
- * $Pred(i)$ est l'ensemble des cellules dont la sortie est une entrée de la cellule i ; ceci implique que la cellule n'est pas une cellule d'entrée et que tous les éléments de $Pred(i)$ appartiennent à la couche précédente de celle à laquelle appartient la cellule i ,
- * $y_i = \sum_{j \in Pred(i)} w_{ij} x_{ij}$, est l'entrée totale de la cellule i ,
- * $o_i = \sigma(y_i)$ est la sortie de la cellule i où σ est la fonction lien.

- * $Succ(i)$ est l'ensemble des cellules qui prennent comme entrée la sortie de la cellule i , ceci implique la cellule n'est pas une cellule de sortie et que tous les éléments de $Succ(i)$ appartiennent à la couche suivante de celle à laquelle appartient la cellule i .

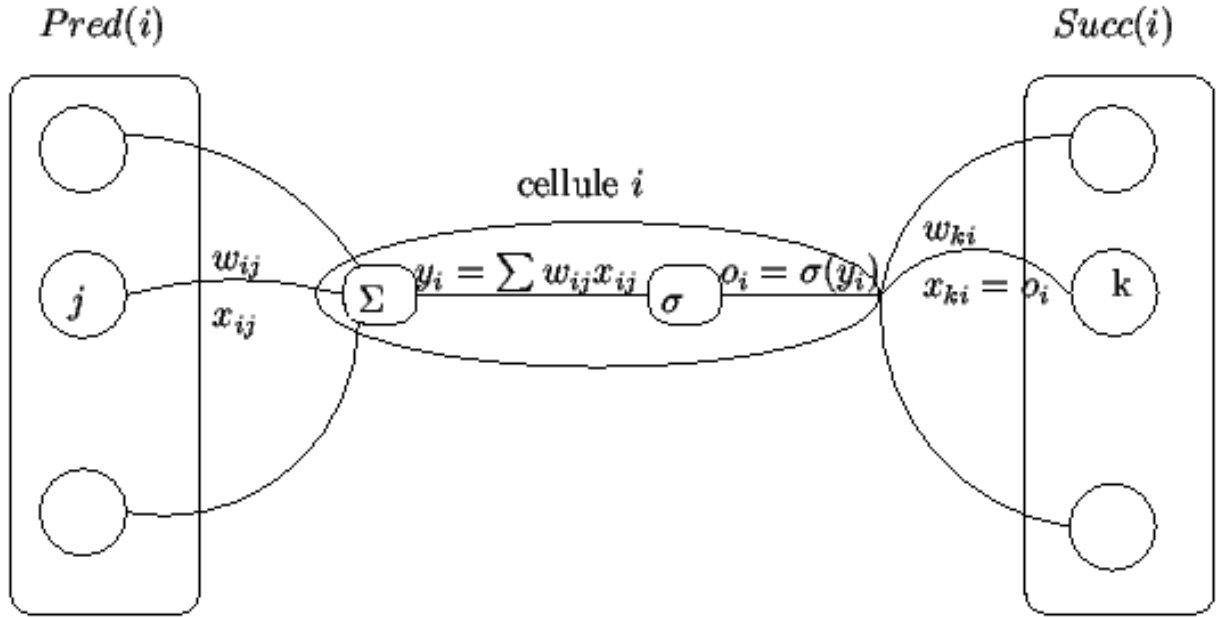


Figure 30 - Notations illustrées

Il nous reste à évaluer la quantité $\frac{\partial E}{\partial w_{ij}}(\vec{w})$ que nous noterons plus simplement $\frac{\partial E}{\partial w_i}$.

Nous remarquons que w_{ij} ne peut influencer la sortie que par le biais de y_i . On écrit donc :

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} x_{ij}$$

La quantité $\frac{\partial E}{\partial y_i}$ dépend de la couche à laquelle appartient la cellule i . On distingue donc deux cas.

1^{er} cas : La cellule i appartient à la couche de sortie.

La quantité y_i ne peut influencer le réseau que par l'intermédiaire de o_i .

On a :

$$\frac{\partial E}{\partial y_i} = \frac{\partial E}{\partial o_i} \frac{\partial o_i}{\partial y_i}$$

Nous allons maintenant calculer chacune des deux dérivées partielles.

Pour la première on a :

$$\frac{\partial E}{\partial o_i} = \frac{\partial}{\partial o_i} \left(\frac{1}{2} \sum_{k=1}^p (c_k - o_k)^2 \right)$$

$$\frac{\partial E}{\partial o_i} = -(c_i - o_i)$$

Ensuite :

$$\frac{\partial o_i}{\partial y_i} = \frac{\partial \sigma}{\partial y_i}(y_i)$$

En rappelant la définition de la fonction sigmoïde, $\sigma(y) = \frac{1}{1+e^{-y}}$, on peut aisément montrer que sa dérivée est telle que : $\frac{\partial \sigma}{\partial y}(y) = \sigma(y)(1 - \sigma(y))$.

Ce qui implique que :

$$\frac{\partial o_i}{\partial y_i} = \sigma(y_i)(1 - \sigma(y_i)) = o_i(1 - o_i)$$

Nous obtenons des résultats obtenus des deux dérivées partielles que :

$$\frac{\partial E}{\partial y_i} = -o_i(1 - o_i)(c_i - o_i)$$

2^e cas : La cellule i appartient à une couche cachée.

Dans ce cas-là, la cellule va influencer le réseau par tous les calculs des cellules appartenant à $Succ(i)$.

Nous avons alors :

$$\frac{\partial E}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial o_i} \frac{\partial o_i}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} w_{ki} o_i(1 - o_i)$$

Soit encore:

$$\frac{\partial E}{\partial y_i} = o_i(1 - o_i) \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k}$$

Par l'étude de ces deux cas nous obtenons deux équations permettant de calculer les dérivées partielles $\frac{\partial E}{\partial y_i}$. Ce calcul devra se faire de la couche de sortie vers la première couche cachée. D'où le terme « rétropropagation » dans le nom de l'algorithme. Nous pouvons ainsi déterminer toutes les dérivées partielles $\frac{\partial E}{\partial w_{ij}}$.

Enfin, pour en déduire la modification à effectuer sur les poids synaptiques, il nous reste simplement à rappeler que la méthode du gradient nous indique que :

$$Dw_{ij} = -e \frac{\partial E(\vec{w})}{\partial w_{ij}}$$

Notons d_i le terme $\frac{\partial E}{\partial y_i}$, En utilisant les équations précédentes, on a :

- Si la cellule i appartient à la couche de sortie :

$$d_i = o_i(1 - o_i)(c_i - o_i)$$

- Si la cellule i appartient à une couche cachée :

$$d_i = o_i(1 - o_i) \sum_{k \in \text{Succ}(i)} d_k w_{ki}$$

La modification du poids est alors définie par :

$$Dw_{ij} = ex_{ij}d_i$$

On peut désormais écrire l'algorithme :

Algorithme de rétropropagation du gradient :

Entrée : Un échantillon A de $\mathbb{R}^n \times \mathbb{R}^p$; e ,

Un perceptron PMC avec une couche d'entrée C_0 , $q - 1$ couches cachées C_1, \dots, C_{q-1} et une couche de sortie C_q , n cellules

initialisation aléatoire des poids w_{ij} dans $[-0,5; 0,5]$, pour i entre 0 et n

Répéter

Prendre un exemple (\vec{x}, c) dans A

calculer la sortie o du perceptron pour l'entrée \vec{x}

--Calcul des d_i par rétropropagation--

Pour toute cellule de sortie i $d_i \leftarrow o_i(1 - o_i)(c_i - o_i)$ **FinPour**

Pour chaque couche de $q - 1$ à 1

Pour chaque cellule i de la couche courante

$$d_i \leftarrow o_i(1 - o_i) \sum_{k \in \text{Succ}(i)} d_k w_{ki}$$

FinPour

FinPour

-- Mise à jour des poids--

Pour tout poids, $w_{ij} \leftarrow w_{ij} + ed_ix_{ij}$

FinPour

FinRépéter

Sortie : Un PMC défini par la structure initiale choisie et les w_{ij} .

Remarque :

- L'algorithme de rétropropagation du gradient est une extension de l'algorithme de Widrow-Hoff. En effet, dans les deux cas, les poids sont mis à jour à chaque présentation d'exemple et donc on tend à minimiser l'erreur calculée pour chaque exemple et pas l'erreur globale.
- La méthode donne de bons résultats pratiques. Dans la plupart des cas, on rencontre très peu de problèmes dus aux minima locaux, peut être grâce au fait que l'on minimise une erreur locale.

- Cependant, les problèmes de minimas locaux existent. Pour améliorer l'algorithme vis à vis de ce problème, mais aussi pour essayer d'améliorer la vitesse de convergence, une variante couramment utilisée consiste à pondérer la modification des poids en fonction du nombre d'itérations déjà effectué. Plus formellement, on fixe une constante appartenant à $[0,1[$, appelée moment (momentum) ; soit t un compteur du nombre d'itérations de la boucle principale, la règle de modification des poids devient :

$$w_{ij} \leftarrow w_{ij} + Dw_{ij}(t)$$

$$Dw_{ij}(t) = ed_i x_{ij} + aDw_{ij}(t-1)$$

- L'intérêt de cette règle est de prendre en compte les modifications antérieures des poids dans le but d'éviter des oscillations perpétuelles.
- Le mode de présentation des exemples est également absent de notre algorithme. En règle générale, on choisit une présentation équitable.
- Le choix de l'architecture initiale du réseau reste un problème difficile. Ce choix peut être fait par l'expérience. Des méthodes dites « autoconstructives » existent : il s'agit d'ajouter des cellules au cours de l'apprentissage pour que l'apprentissage se fasse bien. Mais ces méthodes rencontrent souvent le problème de la sur-spécialisation. L'architecture peut aussi être choisie à l'aide de méthodes basées sur les algorithmes génétiques.
- Le modèle et l'architecture étant choisie, un problème est également de choisir les « bonnes » valeurs pour les paramètres e et a . Pour cela, on découpe l'ensemble d'apprentissage en un ensemble d'apprentissage, un ensemble de validation et un ensemble test. Lors de la phase d'apprentissage, on arrête périodiquement l'apprentissage, on estime l'erreur réelle sur l'ensemble de validation, on met à jour les paramètres e et a , en fonction de la variation de cette erreur estimée. L'ensemble test sert à estimer l'erreur réelle à la fin de l'apprentissage.
- L'algorithme de rétropropagation du gradient a pu être étendu à certaines classes de réseaux récurrents pour lesquels il y a rétroaction. Ces réseaux sont très utiles pour la prédiction de séries temporelles.
- La qualité d'un PMC plus généralement du réseau de neurones nécessite l'ajustement d'un certain nombre de paramètres.
- L'algorithme peut nous estimer le « modèle parfait » c'est à dire qui est telle que l'erreur est nul. Un tel modèle est issu d'un apprentissage par cœur sur l'ensemble d'apprentissage. Le pouvoir prédictif en cas de sur-apprentissage est faible lorsque les exemples sont hors de la base d'apprentissage. Autrement dit : Etant donné l'ensemble d'apprentissage A , un ensemble test T L'hypothèse h sur-apprend si et seulement si il existe une hypothèse h' telle que : $E_A(h) < E_A(h')$ et $E_T(h) > E_T(h')$
- La qualité d'un PMC plus généralement du réseau de neurones nécessite l'ajustement d'un certain nombre de paramètres.

5. Réglage du réseaux et critère d'arrêt

Les réglages des réseaux de neurones sont un peu délicats, et l'on verra plus loin sur leur sensibilité à certains paramètres. Deux de ces paramètres sont particulièrement importants : le nombre d'unités dans la couche cachée, et le terme de pénalisation « weight decay » qui ajoute à l'erreur un terme $\lambda \sum w_{ij}^2$, les w_{ij} étant les poids du réseau. L'augmentation du decay et la diminution du nombre de couches sont les deux principaux moyens pour diminuer le sur-apprentissage. Certains préconisent d'ailleurs d'augmenter le nombre d'unités de la couche cachée mais de compenser ce nombre d'unités par un decay accru. La principale difficulté du réglage de ces paramètres réside dans l'absence de critères théoriques pour leur choix. Il est aussi possible de considérer un weight decay en norme L^1 , qui ajoute à l'erreur un terme $\lambda \sum |w_{ij}|$ et permet d'annuler certains poids et simplifie le réseau contrairement à celui en norme L^2 vu précédemment.

Le nombre d'itérations constitue également un paramètre très sensible dans la performance du réseau de neurones. Le nombre d'itérations de l'apprentissage, qu'il est peut être bon de limiter à une valeur assez faible, peut être basse pour trouver le minimum global mais suffisamment élevé pour que le minimum local trouvé ne soit pas trop éloigné de ce dernier.

Le critère d'arrêt dépend de l'erreur observée mesurée sur l'ensemble d'apprentissage et non de l'erreur réelle. Une technique bien connue s'appelle le « early stopping » qui consiste à arrêter la prématurément la présentation des exemples afin d'éviter le sur-apprentissage. Cette technique consiste à arrêter l'algorithme lorsque l'on obtient l'erreur minimum sur l'ensemble test sans pour autant avoir atteint le minimum sur l'ensemble d'apprentissage.

Il faut donc, si la taille de l'ensemble le permet, découper l'ensemble d'apprentissage en trois :

- Ensemble d'entraînement
- Ensemble de validation
- Ensemble test

L'ensemble d'entraînement sert à l'apprentissage des coefficients, celui de validation est utilisé pour déterminer le nombre d'itérations optimal et enfin l'ensemble test permet de vérifier la performance du réseau ainsi que sa capacité de généralisation.

Soient N le nombre d'exemples de l'ensemble d'apprentissage, la k -ième valeur attendue y_k et la k -ième valeur prédite $f(x_k)$, l'Erreur Quadratique Moyenne (E.Q.M) peut-être définie comme suit :

$$EQM(i) = \frac{1}{N} \sum_{k=1}^N (y_k - f(x_k))^2$$

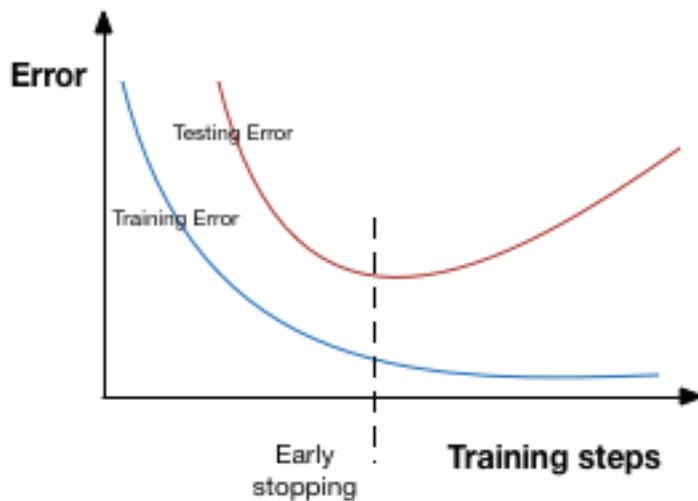


Figure 31- Illustration early stopping

II. Application à tarification automobile selon les réseaux de neurones

D'après la propriété d'approximation universelle du PMC à une seule couche cachée, on peut modéliser directement la prime pure sans passer par un modèle coût x fréquence à l'aide de ce réseau de neurones.

Notre portefeuille de données est constitué de 110 112 contrats, avec un historique allant de 2011 à 2014. On le divise en trois échantillons :

- Une base d'apprentissage, composée de 1/2 du portefeuille,
- Une base de validation, qui permet de déterminer le nombre d'opérations, composée de 1/4 du portefeuille,
- Une base test, qui permet de vérifier la performance du réseau.

Le modèle est construit à l'aide de la base d'apprentissage et la base de validation. La base test sert à comparer de manière objective les trois modèles que nous construirons.

Nous articulerons notre démarche en trois axes.

Premièrement, nous identifierons nos variables d'entrées et de sortie. Ensuite, nous retraiterons nos données car les méthodes de machine learning, prédisent mal les variables numériques à fortes valeurs et modélisent chaque modalité des variables qualitatives comme une variable binaire. Enfin, nous déterminerons notre modèle neuronal optimal à l'aide du package nnet sous le logiciel R.

A. Identification des données en entrées et en sortie

Les réseaux de neurones ne disposants pas de critère de sélection de variables nous conservons les variables utilisées lors de la modélisation à l'aide du GLM. De plus, les réseaux de neurones permettent d'utiliser les variables continues sans les discrétiser en amont. Par conséquent, nous utiliserons les variables continues et les variables qualitatives.

La variable de sortie correspond à la variable à expliquer. Nous modéliserons directement la prime pure. Il s'agit ici de la prime pure par profil de risque. On la défini par :

$$Prime\ Pure_{\theta} = \mathbb{E} \left(\frac{Montant\ des\ sinistres}{Exposition\ du\ contrat} \middle| \theta \right)$$

Les variables d'entrées correspondent aux variables explicatives et la variable explicative correspond à la variable de sortie. Nous rappelons variables retenues pour le GLM :

Variables d'entrées	Nature
age du conducteur	Continue
ancienneté du permis	Continue
ancienneté du véhicule	Continue
classe	Ordinale
fractionnement des paiements	Nominale
crm	Discrète
groupe	Ordinale
parking	Nominale
zone	Ordinale
nature	Nominale
responsabilité	Nominale

B. Retraitement des données

Dans un réseau de neurones, les variables d'entrées et de sortie doivent d'être numériques. Cela nécessite un retraitement des variables à notre disposition.

1. Retraitement des variables qualitatives

Pour rendre numérique une variable qualitative, il est nécessaire de décomposer chaque modalité en variable binaire ce qui accroît considérablement le nombre de variables d'entrées.

2. Retraitement des variables continues

Il s'agit ici d'une normalisation des variables continues pour qu'elles soient comprises dans l'intervalle $[0,1]$. Il existe plusieurs méthodes.

Méthode 1 (écart-type):

La première méthode est la plus utilisée. On centre et on réduit la variable. Notons x_i l'observation d'origine, la moyenne de la variable μ_x et l'écart-type de la variable σ_x . Chaque observation est transformée de la manière suivante :

$$\tilde{x}_i = \frac{x_i - \mu_x}{\sigma_x}$$

Méthode 2 (plage) :

La méthode suivante est la remise à l'échelle. On note le minimum de la variable $\min(x)$ et $\max(x)$ respectivement le minimum et le maximum de la variable.

$$\tilde{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Cependant, en présence de forts extrêmes, cette normalisation peut présenter des problèmes d'écrasement des valeurs normales par ces extrêmes.

Méthode 3 (intermédiaire):

Cette méthode dite intermédiaire est procédée par la transformation suivante :

$$\tilde{x}_i = \frac{x_i - m}{r}$$

Avec $m = \frac{\max(x) + \min(x)}{2}$ et $r = \frac{\max(x) - \min(x)}{2}$

Nous avons fait le choix de la méthode la plus standard, la méthode 1.

Nous normalisons toutes nos variables continues y compris la variable de sortie prime pure.

Ce qui signifie que nous devons la « dénormaliser » après la modélisation.

C. Modélisation de la prime pure à l'aide du perceptron à une couche cachée

Le perceptron à une couche cachée est implémenté dans le package nnet de R.

On l'utilise simplement en spécifiant l'échantillon d'apprentissage, le nombre d'itérations, qui peut ne pas être atteint si l'algorithme de recherche de l'erreur minimale converge avant.

La fonction `nnet`, nous permet de construire notre perceptron en fixant le nombre d'itérations maximum.

Nombre d'itérations	EQM	
	Base d'apprentissage	Base de validation
100	0,034935	0,038648
300	0,034684	0,037834
500	0,033548	0,037271
600	0,032673	0,036841
700	0,032645	0,037847
800	0,031874	0,037981
900	0,031945	0,039452
1000	0,033846	0,039634

Tableau 12- EQM en fonction des itérations

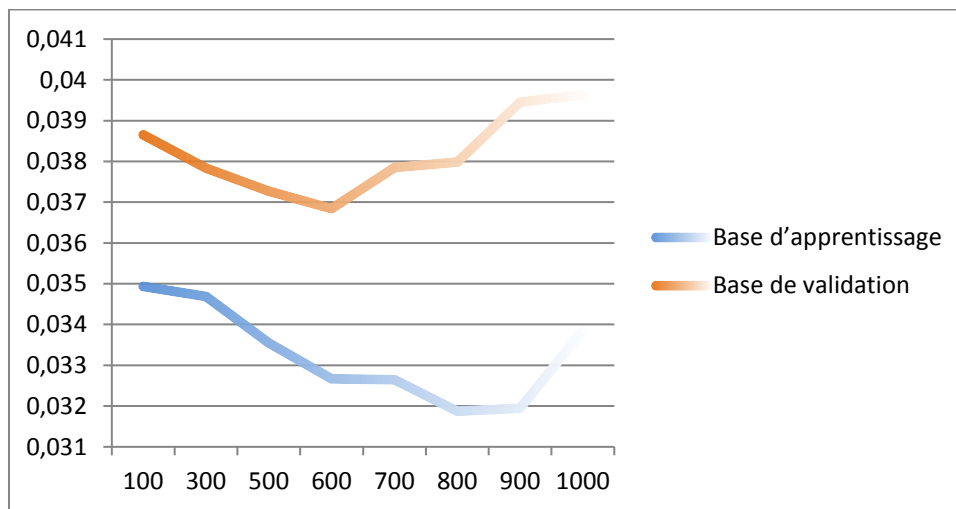


Figure 30- EQM en fonction du nombre d'itérations

Ces simulations nous suggèrent un nombre d'itérations optimal, pour éviter le sur-apprentissage, compris entre 600 et 700 simulations.

Nous utilisons la fonction `best.nnet` du package `e1071` pour avoir le nombre de neurones optimal de la couche cachée.

Nous recensons les sorties dans le tableau suivant :

Nombre de couche d'entrées	Nombre de neurones de la couche cachée	Nombre de sortie	Nombre de poids
21	20	1	461

Tableau 13- Modèle issu de la fonction de R `best.nnet`

Voir la commande R et une partie des coefficients de pondération en annexe.

Construisons à présent notre propre réseau de neurones à l'aide de l'algorithme de rétropropagation du gradient tel qu'on énoncé dans la section V.B.5.d

On choisit comme modèle initial :

- Architecture du réseau de neurones : Perceptron Multicouches à une seule couche cachée
- Variables de la couche d'entrée : Toutes les variables retenues pour le GLM
- Nombre de neurones de la couche cachée : 20
- Distribution des poids de connexion : Normal $N(0,1)$ pris uniquement dans $[-0,5 ; +0,5]$
- Fonction de combinaison de la couche cachée : linéaire
- Fonction d'activation : sigmoïde
- Fonction d'activation de la couche de sortie : linéaire
- Nombre de simulations: 650

Après « dénormalisation » de la prime pure, nous recensons les résultats de la comparaison entre notre algorithme et celui issu de la fonction `best.nnet` dans le tableau suivant :

	EQM	PP Min	PP Max	PP Moy
best.nnet	75 426,00	54,70	3 050,20	407,60
own model	85 438,00	-87,10	3 548,30	501,40

Tableau 14- comparaison des deux modèles

Ces données sont obtenues sur la base test.

Nous retenons naturellement le modèle issu de la commande `best.nnet`.

D. Comparaison entre le GLM et le réseau de neurones

Dans cette section les comparaisons vont être faites sur la base des prédictions de prime pure issues de la base test. Une première analyse globale sera faite et ensuite nous analyserons les primes obtenues par profil de risque.

1. Comparaison globale

à l'aide de l'Erreur Quadratique Moyenne.

	EQM	PP Min	PP Max	PP Moy
Réseau de Neurones	75 426,00	54,70	3 050,20	836,60
GLM	77 220,12	10,23	6 230,54	1 200,78
Prime Pure observée	x	10,00	4 703,76	900,34

Tableau 15- Comparaison GLM et Réseau de neurones

Le modèle linéaire généralisé surestime la prime pure, alors que le réseau de neurones la sous-estime. De plus, l'écart quadratique moyen entre les prédictions et le réel est plus faible pour le réseau de neurones que pour le GLM. Le réseau de neurones prédit globalement plutôt bien.

Dans cette section, nous avons décidé de créer un modèle agrégé entre le GLM et le Réseau de neurones sur la base test.

2. Comparaison par profil de risque

Nous composons les profils de risque en fonction des différents croisements possibles entre les différentes modalités des variables explicatives (variables continues discrétisées y comprises) quand ceux-ci sont possibles. Nous obtenons ainsi 124 000 profils de risques potentiels. Nous comparons les primes pures sur le top 15 des profils de risques les plus récurrents sur les 27 528 contrats de la base test.

Profil	Nombre d'observations	PP Moy observée	PP Moy GLM	PP Moy RN	Ecart GLM	Ecart RN
Profil1	735	847	898	1292	6%	53%
Profil2	669	1362	1454	432	7%	-68%
Profil3	662	635	756	929	19%	46%
Profil4	554	358	1170	394	227%	10%
Profil5	544	1156	1643	1374	42%	19%
Profil6	497	1376	365	755	-73%	-45%
Profil7	493	847	1381	1544	63%	82%
Profil8	474	1244	1494	1176	20%	-5%
Profil9	443	1088	480	802	-56%	-26%
Profil10	317	595	706	1056	19%	77%
Profil11	300	621	1152	973	86%	57%
Profil12	275	881	695	308	-21%	-65%
Profil13	271	405	647	1470	60%	263%
Profil14	223	976	1518	245	56%	-75%
Profil15	221	1001	1466	747	46%	-25%

Tableau 16- Primes pures sur les 15 profils les plus représentés

Le modèle linéaire généralisé nous donne de meilleurs résultats dans 8 cas contre 7, pour le réseau de neurones. L'écart relatif moyen relatif entre le GLM et les observations sur ces 8 cas est de 26%, tandis que l'écart moyen relatif entre le réseau de neurones et les observations sur les cas où ce dernier estime mieux la prime pure est de -2%. Par ailleurs, l'écart global entre le GLM et les observations est de 33% contre 20% entre le réseau de neurones et les observées de la base test.

3. Conclusion de la comparaison

Nous avons constaté que le modèle linéaire généralisé est plus prudent que le réseau de neurone, dans le sens où il surestime la prime pure.

De plus, nous avons vu que le modèle généralisé à obtenu un écart moins important sur plus de profils.

Toutefois, en matière d'EQM le réseau de neurone possède un taux d'erreur plus faible que le GLM. Aussi, le réseau de neurones à un écart relatif très faible sur les profils où il prédit mieux que le GLM.

Nous pouvons émettre l'hypothèse que chaque modèle à une préférence en matière de profil à prédire. Cette hypothèse nous conduit à construire un modèle qui tiendrait compte des deux modèles et qui serait construit sur a base de leur préférence.

E. Modèle agrégé

Dans cette section, nous allons construire un estimateur qui se construira à l'aide des deux modèles en utilisant la méthode du bagging.

1. La méthode du bagging

La méthode du bagging, bootstrap aggregating (Breiman 1996), repose sur une construction aléatoire d'une famille de modèles. Cet algorithme est développé à la frontière entre machine learning et statistique.

❖ Principe

Soit Y une variable à expliquer quantitative ou qualitative, X_1, \dots, X_p les variables explicatives et $\phi(x)$ est un modèle fonction de $x = (x_1, \dots, x_p) \in \mathbb{R}^p$.

On note n le nombre d'observations et $z = \{(x_1, y_1), \dots, (x_p, y_p)\}$ un échantillon de loi F .

L'espérance $\phi(\cdot) = \mathbb{E}_F(\hat{\phi}_z)$ de l'estimateur définie sur l'échantillon z . Considérons M échantillons indépendants notés $\{z_m\}_{m=1, \dots, M}$ et construisons une agrégation des modèles dans le cas où la variable à expliquer Y est :

- * Quantitative : $\hat{\phi}_M(\cdot) = \frac{1}{M} \sum_{m=1}^M \hat{\phi}_{z_m}(\cdot)$,
- * Qualitative : $\hat{\phi}_M(\cdot) = \operatorname{argmax}_j \operatorname{card}\{m | \hat{\phi}_{z_m}(\cdot) = j\}$

Nous nous intéressons qu'aux variables quantitatives dans le reste de notre étude.

Dans le cas des variables quantitatives moyenner les prévisions de plusieurs modèles indépendants permet de réduire la variance et donc réduire l'erreur de prévision.

Cependant, l'hypothèse des M échantillons indépendants n'est pas réaliste. On les remplace donc par une réplcation de M échantillons bootstrap obtenus chacun par n tirages aléatoires. Nous voyons par la suite l'algorithme classique de bagging et une variante que nous avons apportés.

❖ Algorithmes

L'algorithme ci-dessous nous décrit la méthode d'obtention d'un estimateur agrégé.

Algorithme Bagging

Entrée : x_0 à prévoir et $z = \{(x_1, y_1), \dots, (x_p, y_p)\}$ un échantillon

Pour m de 1 à M

Tirer un échantillon bootstrap z_m^* .

Estimer $\hat{\phi}_{z_m}(x_0)$ sur l'échantillon bootstrap.

FinPour

Calculer l'estimation moyenne $\hat{\phi}_M(x_0) = \frac{1}{M} \sum_{m=1}^M \hat{\phi}_{z_m}(x_0)$

Sortie : $\hat{\phi}_M$

Cet algorithme est appliqué pour agréger les modèles de même famille. Nous apportons une variante qui nous permet de faire la sélection du meilleur modèle à chaque tirage.

Algorithme Bagging adapté

Entrée : x_0 à prévoir et $z = \{(x_1, y_1), \dots, (x_p, y_p)\}$ un échantillon

Pour m de 1 à M

Tirer un échantillon bootstrap z_m^* .

Estimer $\phi_{z_m}^{GLM}(x_0)$ et $\phi_{z_m}^{RN}(x_0)$ sur l'échantillon bootstrap.

$$\hat{\phi}_{z_m}(x_0) = \min[\phi_{z_m}^{GLM}(x_0), \phi_{z_m}^{RN}(x_0)]$$

FinPour

Calculer l'estimation moyenne $\hat{\phi}_M(x_0) = \frac{1}{M} \sum_{m=1}^M \hat{\phi}_{z_m}(x_0) = \frac{\alpha}{M} \sum \phi_{z_m}^{GLM}(x_0) + \frac{\beta}{M} \sum \phi_{z_m}^{RN}(x_0)$

Sortie : $\hat{\phi}_M$

Où α est le nombre de fois où le GLM a été plus performant et β le nombre de fois où le réseau de neurones est plus le performant.

2. Comparaisons avec les autres modèles

Nous rajoutons à la comparaison précédente une ligne supplémentaire pour le modèle agrégé

	EQM	PP Min	PP Max	PP Moy
Modèle agrégé	70 674,36	34,95	4 286,67	829,85
Réseau de Neurones	75 426,00	54,70	3 050,20	836,60
GLM	77 220,12	10,23	6 230,54	1 200,78
Prime Pure observée	x	10,00	4 703,76	900,34

Tableau 17- Comparatif des trois modèles

Le modèle agrégé est globalement meilleur que les deux autres au sens de l'erreur quadratique moyenne. Le graphique ci-dessous présente côte à côte la prime pure moyenne, du top 15 des profils les plus présents dans la base test, entre modèle agrégé et les observations.

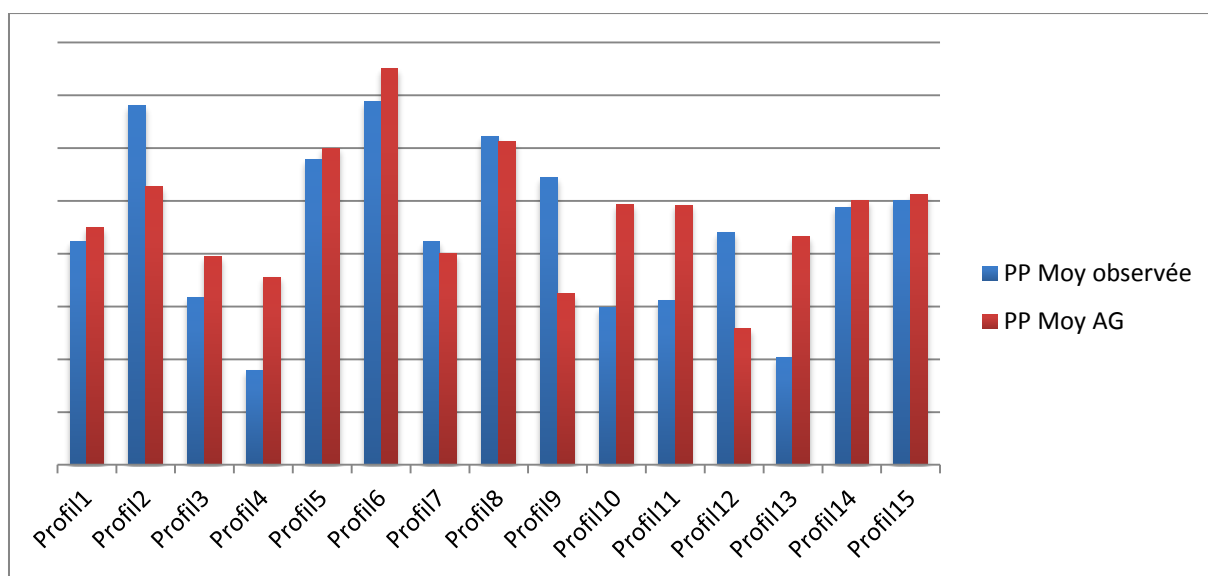


Figure 31 - prime pure observée et modèle agrégé

III. Limites et avantages de chaque modèle

Dans le cadre de la tarification de la prime pure, nous avons réalisé une modélisation avec 3 modèles. L'objet de cette étude était de comparer ces modèles.

Chacune des méthodes présente un certain nombre d'avantages:

- ✓ Le modèle linéaire généralisé est la méthode de tarification en assurance automobile la plus utilisée.
 - ✓ Le modèle linéaire généralisé est facile à mettre en œuvre.
 - ✓ L'implémentation du modèle linéaire généralisé se fait via tout logiciel statistique.
 - ✓ La prédiction du GLM est plutôt satisfaisante.
 - ✓ Aucune normalisation des données n'est nécessaire.
-
- ✓ Les réseaux de neurones par leur propriété d'approximation universelle peuvent prédire toute forme de liaison assez régulière entre des variables explicatives et expliquées.

- ✓ La prime pure est modélisée directement par le réseau de neurone et nécessite donc pas de passer par deux modèles de manière général voir quatre pour notre modèle tenant compte des montants de forfait IDA.
- ✓ Utilisation de variables continues
- ✓ Le pouvoir prédictif du réseau de neurones est plus fort que celui du GLM.

Ces méthodes comme toute méthode statistique présente des inconvénients :

- ✓ Le GLM a pour principal inconvénient de faire appel à des hypothèses de loi très fortes,
- ✓ Le GLM ne peut pas utiliser les variables continues pour la prédiction,
- ✓ Les réseaux de neurones ont une structure très complexe,
- ✓ Les réseaux de neurones sont considérés comme une boîte noire car difficile à interpréter,
- ✓ Nécessite un choix de modalités réduite pour les variables qualitatives car les transforment binaire.
- ✓ Le risque de sur-ajustement très peu maîtrisé.

CONCLUSION GÉNÉRALE

Le présent mémoire avait pour but de comparer un référentiel de tarification historique, le GLM, et une des nouvelles méthodes de tarification en assurance, les réseaux de neurones.

Afin d'atteindre cet objectif, nous avons mis en œuvre pour chaque contrainte, nombre de solutions.

En ce qui concerne la méthode du GLM :

Tout d'abord, nous avons constaté des pics du coût moyen sur certaines modalités. Ces pics s'expliquaient par la présence de montants de sinistre exceptionnellement élevés apportant des anomalies de distribution. Pour résoudre ce problème et obtenir une homogénéité dans la distribution, nous avons écrêté nos montants de sinistre. Nous avons évoqué plusieurs méthodes de détermination du seuil d'écrêtement. Le montant du seuil de l'écrêtement a été fixé à partir de la moyenne des seuils obtenus dans chaque méthode.

S'est, ensuite, posée la question de la distribution écrêtée. Nous avons constaté une accumulation de points à certaines valeurs correspondant au montant de forfait IDA. Ce qui contredisait une des hypothèses de continuité de la distribution du coût du sinistre. Nous avons opté pour une construction de quatre modèles. Un modèle de coût de sinistre, qui ne prend pas en compte les forfaits IDA et auquel on applique une majoration sur chaque tête. Cet écrêtement est lié à la mutualisation de la charge sinistre exceptionnelle. Le forfait IDA étant un montant connu d'avance, on choisit le montant de l'année en cours. On modélise trois fréquences, la première non concernée par la convention, la deuxième, des contrats pour lesquels la charge est un forfait IDA actif et le dernier modèle IDA passif.

La modélisation du GLM nous apporte des prédictions plutôt satisfaisantes sur l'ensemble du portefeuille.

Pour les réseaux de neurones :

La mise en œuvre a été moins contraignante mais l'interprétation des résultats obtenus est plus complexe.

La prédiction globale est plus satisfaisante que le modèle GLM mais l'implémentation informatique est coûteuse en matière de temps.

Par ailleurs, nous avons sélectionné les profils les plus représentés dans la base test, puis nous avons réalisé une comparaison des prédictions issues des deux modèles suivant ces profils. Nous obtenons ainsi plus de prédictions exactes avec le modèle GLM qu'avec les réseaux de neurones. Toutefois, l'écart relatif global sur l'ensemble des profils reste largement plus faible pour les réseaux de neurones, que pour le GLM.

Sur la base des résultats précédemment obtenus, la prédiction globale nous conduirait à opter pour la méthode des réseaux de neurones. Cependant, on constate également qu'il existe des classes de risque plus adaptées au GLM. Nous faisons donc le choix de construire un modèle agrégé à l'aide de la méthode du bagging, que nous avons adaptée et qui prend en compte les préférences de profils

Bibliographies

- [1] AOUIZERATE J.M. [2010]. « Alternative neuronal en tarification santé Mémoire d'actuariat », *Mémoire Actuariat, C.N.A.M.*
- [2] A. CHARPENTIER, M. DENUIT. [2005] « Mathématiques de l'assurance non-vie, Tome II : tarification et provisionnement », *Economica, Paris.*
- [3] A. CHARPENTIER. [2011], « Statistique de l'assurance, STT 6705V, Statistique de l'assurance II », *Cours université de Rennes 1.*
- [4] CORNILLON P-A .et al. [2012], « Statistiques avec R », *Presses Universitaires de Rennes, 3^{ème} éd., 2012.*
- [5] HABERMEHL P. et. KESNER D. [2015] « Cours de programmation logique et IA », *cours 18 IRIF.*
- [6] LAMURE M. [2013]. « Méthode de classification par réseau de neurones », *cours, ISFA*
- [7] MASIELLO E. [2015]. « Modèles linéaires Généralisés » *Cours, ISFA*
- [8] MAXWELL C. [2013]. « Écrêtement des sinistres en assurance non-vie », Publication Galea
- [9] PAGLIA A. [2010]. « Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique », *Mémoire d'actuariat, EURIA.*
- [10] TUFFERY S. [2015]. « Modélisation prédictive et Apprentissage statistique avec R ». *Edition Technip*
- [11] ZIADI F. [2015] « Tarification en plaisance : GLM vs Réseaux de neurones », *Mémoire ISFA*
- [12] « Agrégation de modèles », *cours wikisat université de Toulouse.*
- [13] « Apprentissage automatique des réseaux de neurones », *cours chapitre 3, université de Lille.*
- [14] Chiffres Clés Fédération Française de l'Assurance [En ligne]
<http://www.ffa-assurance.fr/chiffre-cle>
- [15] Principe des courbes ROC [en ligne]
www.xlstat.com/fr/solutions/fonctionnalites/courbes-roc
- [16] Support SAS, the GENMOD Procedure
<http://support.sas.com/documentation>

Tables des figures

Figure 1- Nombre de contrats et chiffre d'affaire 2013 des principaux acteurs de l'assurance Auto	10
Figure 2 - Histogramme de l'âge du conducteur	24
Figure 3- Histogramme de l'ancienneté du véhicule.....	25
Figure 4 - Histogramme de l'ancienneté du permis	26
Figure 5 – Taux de sinistres en fonction du découpage des variables.....	28
Figure 6 – Taux de sinistres en fonction du découpage des variables.....	29
Figure 7– Découpage de la variable age du conducteur	33
Figure 8 – Découpage de la variable ancienneté du permis	34
Figure 9- AUC maximale par nombre de classes	35
Figure 10 – Découpage ancienneté du véhicule	35
Figure 11 - Boxplot de la charge sinistre RC.....	36
Figure 12 - Répartition empirique de la charge RC.....	38
Figure 13 - Fonction moyenne des excès des coûts sinistre.....	42
Figure 14 – Représentation de la distribution empirique du coût des sinistres et du coût moyen	53
Figure 15 – Distribution du coût RC dans la classe M03.....	54
Figure 16 – QQ-plot du coût majoré et des lois comparées	58
Figure 17 – Densités des lois candidates.....	59
Figure 18 – Fonction de répartition gamma et empirique	60
Figure 19 – Résidus.....	61
Figure 20 – Densités des fréquences empiriques et candidates	65
Figure 21 - Réseau de neurone biologique	69
Figure 22 - Réseau de neurone formel	70
Figure 23 - Un perceptron à une couche cachée.....	74
Figure 24 - RVFLNN à une couche cachée.....	74

Figure 25 - Réseau à base de fonction radiale	75
Figure 26 - Architecture d'une carte de Kohonen.....	76
Figure 27- Exemple de réseau de neurones récurrent.....	76
Figure 28 - Méthode du gradient	80
Figure 29- paraboloïde en dimension 3.....	82
Figure 30- EQM en fonction du nombre d'itérations.....	93
Figure 31 - prime pure observée et modèle agrégé.....	98

Liste des tableaux

Tableau 1 -Cotisations par type d'assurance.....	9
Tableau 2- Nombre d'intervalle en fonction du nombre d'observations.....	27
Tableau 3 - Découpage des variables continues.....	28
Tableau 4-Discrétisations retenues.....	29
Tableau 5- AUC maximale par nombre de classes	33
Tableau 6- AUC maximale par nombre de classes	34
Tableau 7- Paramètres obtenus sur notre base	44
Tableau 8- Matrice de corrélation partielle	51
Tableau 9- Montant du forfait par année	55
Tableau 10- Représentation de quelques fonctions d'activation	72
Tableau 11- valeurs des poids à chaque itération.....	79
Tableau 12- EQM en fonction des itérations.....	93
Tableau 13- Modèle issu de la fonction de R best.nnet.....	93
Tableau 14- comparaison des deux modèles.....	94
Tableau 15- Comparaison GLM et Réseau de neurones	94
Tableau 16- Primes pures sur les 15 profils les plus représentés	95
Tableau 17- Comparatif des trois modèles	98

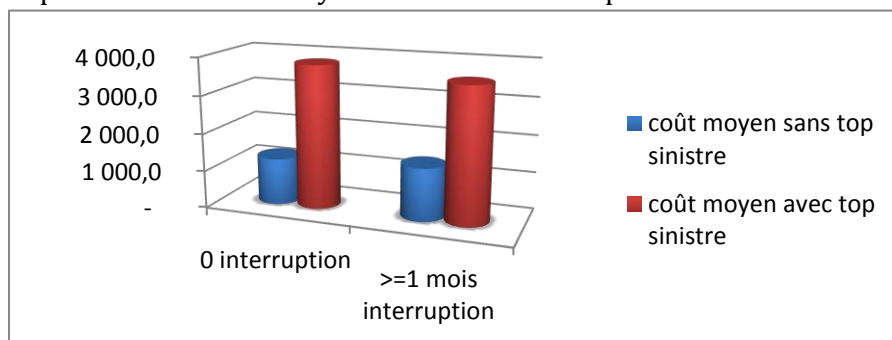
ANNEXES

Annexe 0

Comparaison du coût moyen avec et sans les 15 sinistres les plus élevés

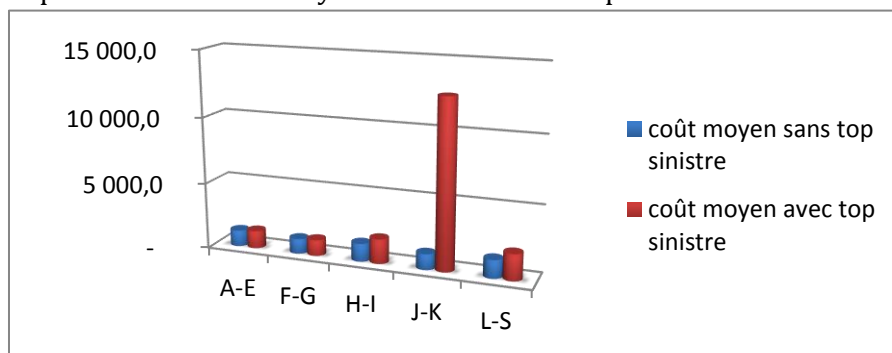
Interruption d'assurance

Expression des coûts moyens avec et sans les tops sinistres



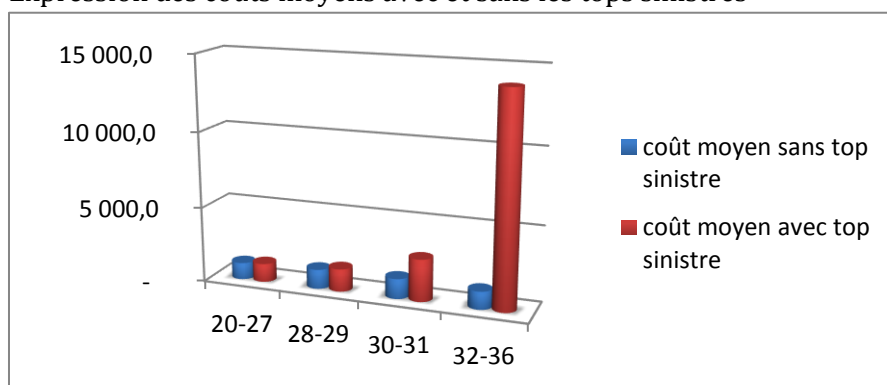
Classe

Expression des coûts moyens avec et sans les tops sinistres



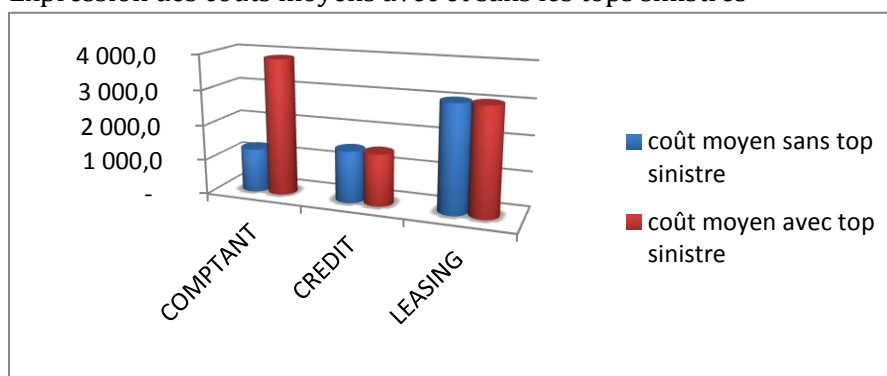
Groupe

Expression des coûts moyens avec et sans les tops sinistres



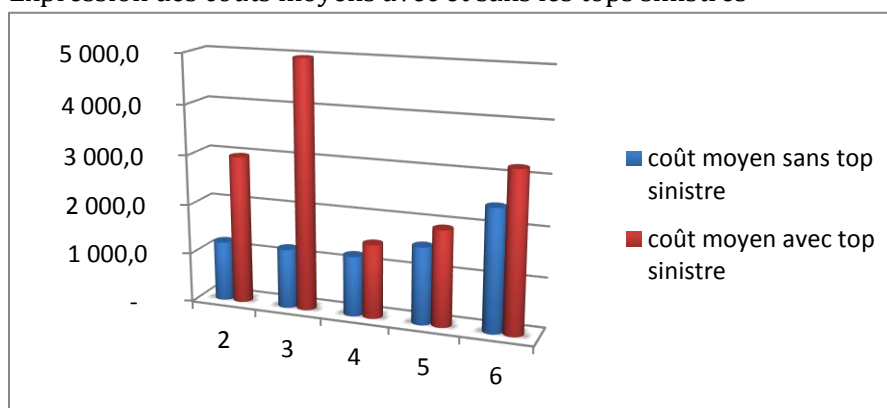
Mode d'acquisition

Expression des coûts moyens avec et sans les tops sinistres



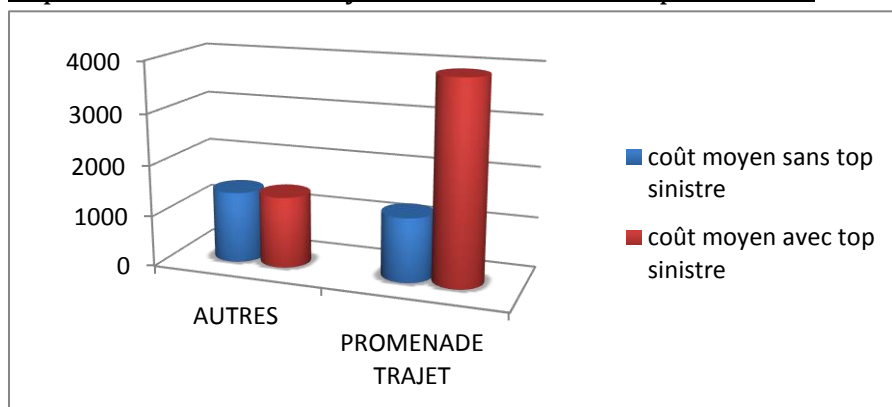
Zone

Expression des coûts moyens avec et sans les tops sinistres



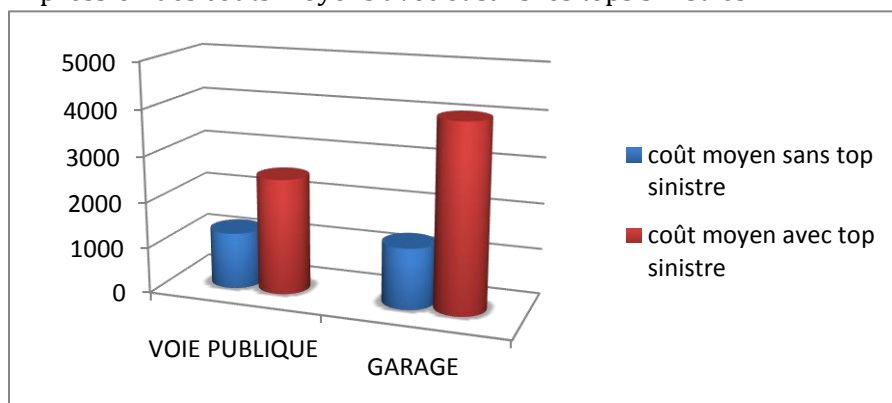
Usage

Expression des coûts moyens avec et sans les tops sinistres



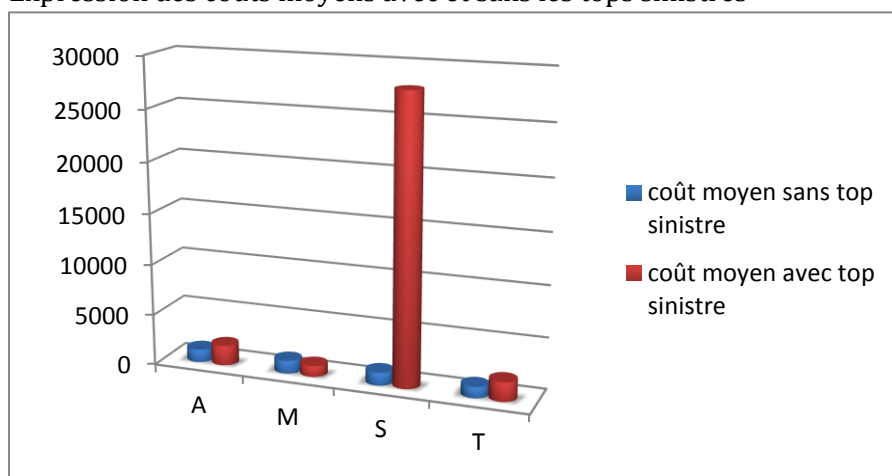
Parking

Expression des coûts moyens avec et sans les tops sinistres



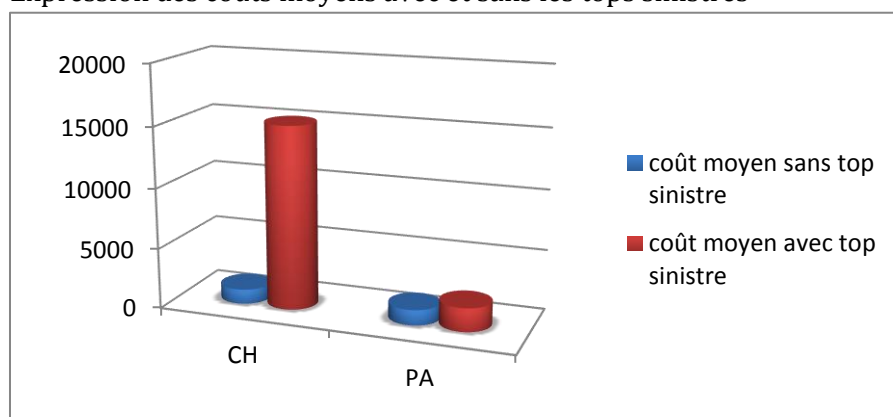
Fractionnement

Expression des coûts moyens avec et sans les tops sinistres



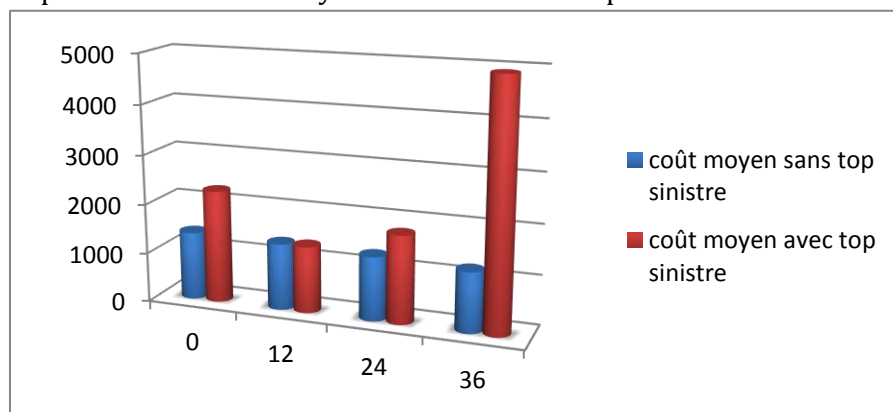
Mode de paiement

Expression des coûts moyens avec et sans les tops sinistres



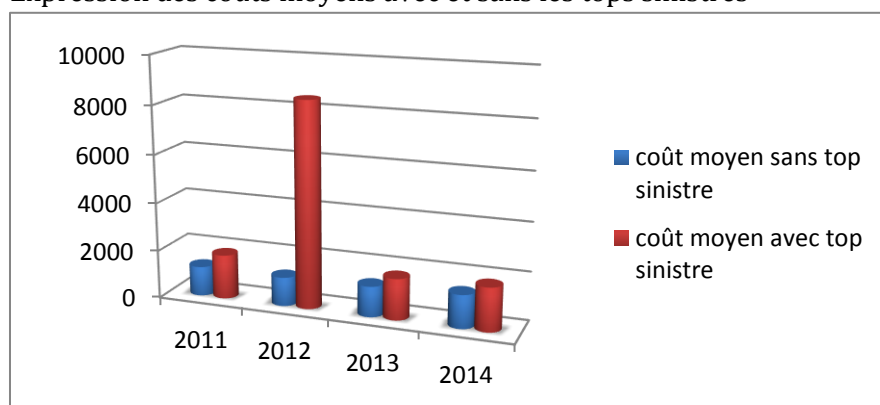
Période sinistre

Expression des coûts moyens avec et sans les tops sinistres



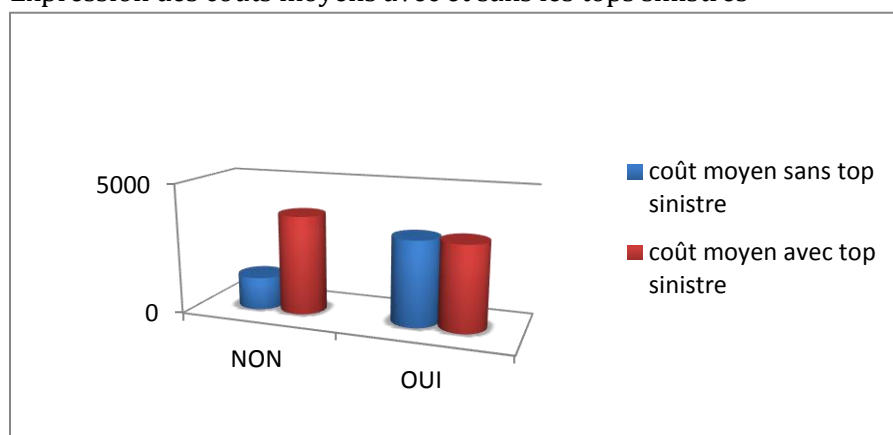
Année

Expression des coûts moyens avec et sans les tops sinistres



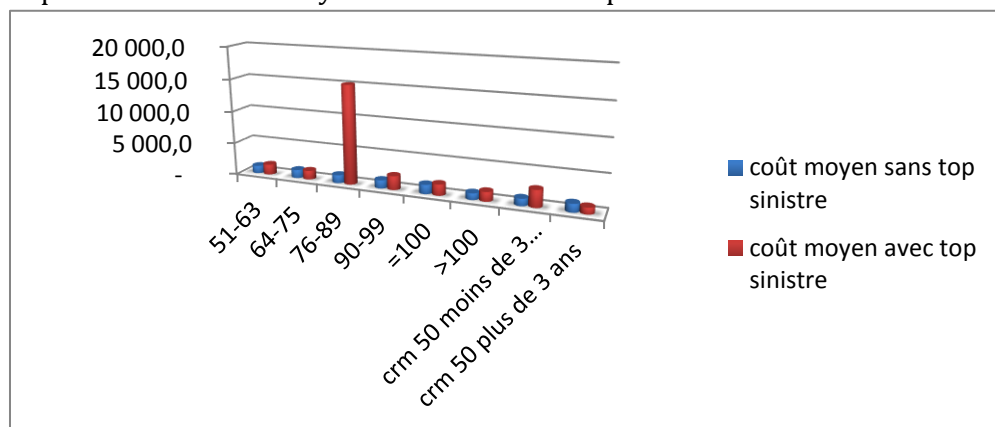
Conduite accompagnée

Expression des coûts moyens avec et sans les tops sinistres



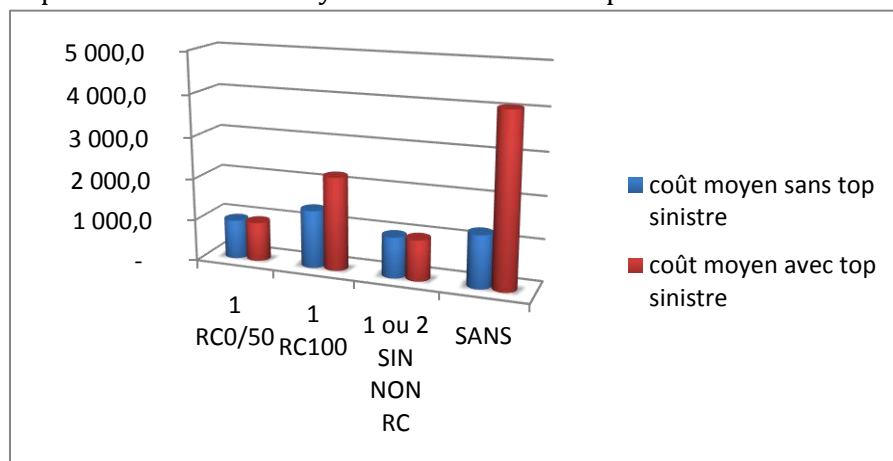
CRM

Expression des coûts moyens avec et sans les tops sinistres



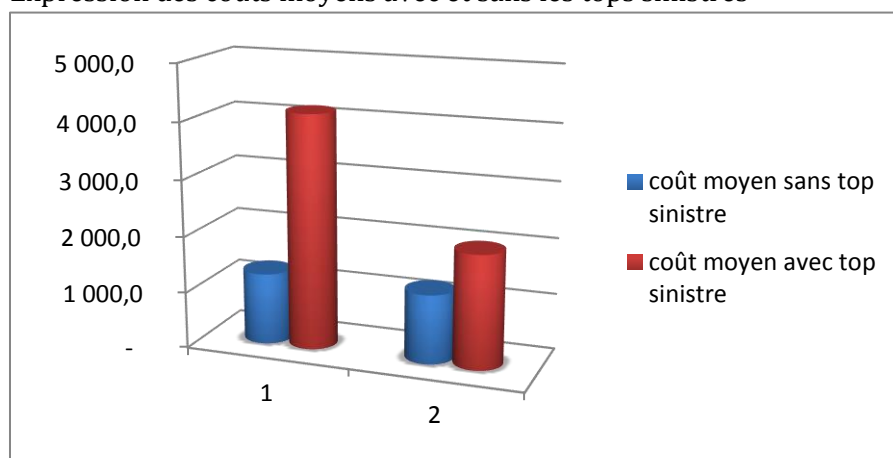
Antécédent & responsabilité

Expression des coûts moyens avec et sans les tops sinistres



Nombre de conducteur

Expression des coûts moyens avec et sans les tops sinistres



Annexe 1

Test d'indépendance du χ^2

Une première indication de l'impact que pourrait avoir nos variables explicatives sur la sinistralité peut être déterminée en considérant la variable top_sinistre qui indique si un sinistre a été déclaré (oui) ou non. Pour ce faire, nous testons l'indépendance entre les variables explicatives et cette dernière, à l'aide du test d'indépendance du chi-deux où nous présentons les résultats dans le tableau suivant :

VARIABLE	DDL	CHI-SQUARE	P-VALUE
<i>Interrup_assur</i>	2	13,7070	0,0011
<i>classe</i>	4	679,4645	< 0,0001
<i>Groupe</i>	5	556,0963	<0,0001
<i>mode_acq</i>	2	281,9643	<0,0001
<i>zone</i>	4	204,9483	<0,0001
<i>usage</i>	6	28,9155	<0,0001
<i>parking</i>	2	4,1008	0,1287
<i>fract</i>	3	83,0752	<0,0001
<i>mod_paie</i>	1	16,4045	<0,0001
<i>periode_sin</i>	3	631,6452	<0,0001
<i>annee</i>	3	796,9270	<0,0001
<i>coef_bm</i>	6	2 264,0552	<0,0001
<i>cond_ac</i>	1	2,2995	0,1294

<i>age_cond</i>	6	204,5682	<0,0001
<i>nbanbm50</i>	10	318,3176	<0,0001
<i>anc_perm</i>	10	196,8408	<0,0001
<i>nbcond</i>	5	127,0219	<0,0001
<i>antécédent assurance</i>	60	1 065,2801	<0,0001
<i>anc_veh</i>	10	1 702,3018	<0,0001
<i>anc_contra</i>	4	4 795,6709	<0,0001

Nous constatons que les variables *parking* et *cond_ac* ont une *p_value* plus grande que le seuil fixé à 5%. L'hypothèse H_0 (la variable observée est indépendante de la variable *top_sinistre*) n'est donc pas rejetée pour ces deux variables. Par conséquent, nous nous séparerons de ces deux variables dans la suite de notre étude.

Annexe 2

Modèle collectif

Le calcul de la prime pure nécessite la connaissance du premier moment de la sinistralité S . Néanmoins, un chargement de sécurité basé sur une mesure de risque peu justifier la connaissance de moments d'ordre supérieurs ou bien des quantiles de la distribution de S . On détermine ici les trois moments de S sans préciser de loi pour N .

On considère le modèle suivant :

$$S = \sum_{i=1}^N X_i$$

où : S représente la charge de sinistre totale

N le nombre total de sinistres survenus

X_i le montant du sinistre numéro i

X_1, X_2, \dots, X_n forment un échantillon i.i.d.

X_i et N sont des v.a . Indépendantes

On a :

$$\mathbb{E}(S|N = n) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mathbb{E}(X_1) \Rightarrow \mathbb{E}(S|N) = N\mathbb{E}(X_1)$$

$$\mathbb{V}(S|N = n) = \mathbb{V}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{V}(X_i) = n\mathbb{V}(X_1) \Rightarrow \mathbb{V}(S|N) = N\mathbb{V}(X_1)$$

On en déduit que :

$$\begin{aligned}
\mathbb{E}(S) &= \mathbb{E}(\mathbb{E}(S|N)) \\
&= \mathbb{E}(N\mathbb{E}(X_1)) \\
\mathbb{E}(S) &= \mathbb{E}(N)\mathbb{E}(X_1) \\
\mathbb{V}(S) &= \mathbb{E}(\mathbb{V}(S|N)) + \mathbb{V}(\mathbb{E}(S|N)) \\
&= \mathbb{E}(N\mathbb{V}(X_1)) + \mathbb{V}(N\mathbb{E}(X_1)) \\
\mathbb{V}(S) &= \mathbb{E}(N)\mathbb{V}(X_1) + \mathbb{E}(X_1)^2\mathbb{V}(N)
\end{aligned}$$

Il est possible de procéder par la fonction génératrice afin d'obtenir les moments de S .

Rappel : Soit X une variable aléatoire admettant des moments d'ordre k .

$$\begin{aligned}
M_X(t) &= \mathbb{E}(e^{tX}) \\
M_X(t) &= \int e^{tx} dF_X(x) \\
\Rightarrow \frac{d^k M_X(t)}{dt^k} &= \int x^k e^{tx} dF_X(x) \\
\frac{d^k M_X(t)}{dt^k} &= \mathbb{E}(X^k e^{tX}) \\
\frac{d^k M_X(t)}{dt^k}(0) &= \mathbb{E}(X^k)
\end{aligned}$$

On a :

$$\begin{aligned}
M_S(t) &= \mathbb{E}(e^{tS}) \\
M_S(t) &= \mathbb{E}\left(\mathbb{E}\left(e^{t\sum_{i=1}^N X_i} \middle| N\right)\right) \\
M_S(t) &= \mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^N e^{tX_i} \middle| N\right)\right) \\
M_S(t) &= \mathbb{E}(\mathbb{E}(e^{tX_1})^N) \\
&= \mathbb{E}(M_X(t)^N) \\
&= \mathbb{E}(e^{N \ln(M_X(t))}) \\
M_S(t) &= M_N(\ln(M_X(t)))
\end{aligned}$$

On peut d'alors évaluer les dérivées successives de M_S en posant $f(t) = M_X(t)$ et $g(t) = M_N(t)$

$$\begin{aligned}
M_S(t) &= g(\ln(f(t))) \\
\frac{dM_S(t)}{dt} &= \frac{f'(t)}{f(t)} \times g'(\ln(f(t)))
\end{aligned}$$

$$\frac{d^2 M_S(t)}{dt^2} = \frac{f''(t)f(t) - f'(t)^2}{f(t)^2} \times g'(\ln(f(t))) + \left(\frac{f'(t)}{f(t)} \right)^2 g''(\ln(f(t)))$$

Posons $A_1 = \frac{f''(t)f(t) - f'(t)^2}{f(t)^2}$, $A_2 = g'(\ln(f(t)))$, $A_3 = \left(\frac{f'(t)}{f(t)} \right)^2$, $A_4 = g''(\ln(f(t)))$

$$\frac{d^3 M_S(t)}{dt^3} = A'_1 A_2 + A_1 A'_2 + A'_3 A_4 + A_3 A'_4$$

$$A'_1 = \frac{f'''f - 3f''f'f + 2f'^3}{f^3}$$

$$A'_2 = \frac{f'}{f} g''(\ln(f))$$

$$A'_3 = 2 \frac{f'}{f} \times \frac{f''f - f'^2}{f^2}$$

$$A'_4 = \frac{f'}{f} g'''(\ln(f))$$

Comme $f(0) = 1$, on obtient :

$$\mathbb{E}(S) = \frac{dM_S(t)}{dt}(0) = \mathbb{E}(X)\mathbb{E}(N)$$

$$\mathbb{E}(S^2) = \frac{d^2 M_S(t)}{dt^2}(0) = \mathbb{E}(X)^2 (\mathbb{E}(N^2) - \mathbb{E}(N)) \mathbb{E}(X^2) \mathbb{E}(N)$$

$$\mathbb{E}(S^3) = \frac{d^3 M_S(t)}{dt^3}(0)$$

$$\mathbb{E}(S^3) = \mathbb{E}(X)^3 (2\mathbb{E}(N) - 3\mathbb{E}(N^2) + \mathbb{E}(N^3)) + 3\mathbb{E}(X)\mathbb{E}(X^2)(\mathbb{E}(N^2) - \mathbb{E}(N)) + \mathbb{E}(X^3)\mathbb{E}(N)$$

Annexe 3

Distribution de Pareto généralisée

Sa distribution se définit comme suite:

$$p(x) = \begin{cases} \frac{\frac{h\nu}{\sigma}}{\sigma} (1 - \alpha(x - \theta)/\sigma)^{1/\alpha-1} & \text{if } \alpha \neq 0 \\ \frac{h\nu}{\sigma} \exp(-x/\sigma) & \text{if } \alpha = 0 \end{cases}$$

où

θ =seuil

α =paramètre de forme

σ =paramètre de forme ($\sigma > 0$)

h =amplitude de l'intervalle

v =paramètre d'échelle

Annexe 4

Modèle et coefficients du GLM

Modèle coût

PARAMETER	LEVEL1	ESTIMATE	STDERR	CHISQ	EXP(ESTIMATE)
Intercept		8,8045	0,1566	3160,97	6 664,12
classe	A-I	-0,2718	0,1033	6,93	0,76
	J-L	-0,2295	0,0977	5,52	0,79
	M-S	0,0000	0,0000	0,00	1,00
groupe	20-27	-0,2275	0,1278	3,17	0,80
	28-31	0,0082	0,0920	0,01	1,01
	32-36	0,0000	0,0000	0,00	1,00
anc_veh	<=2ans	0,0000	0,0000	0,00	1,00
]2-10]	0,0921	0,1016	0,82	1,10
]10-14]	-0,0998	0,1111	0,81	0,91
]14-20]	-0,0878	0,1156	0,58	0,92
zone	>20	-0,2975	0,1675	3,16	0,74
	2	-0,2707	0,1119	5,85	0,76
	3	-0,2763	0,0852	10,51	0,76
	4	-0,3435	0,1009	11,60	0,71
	>=5	0,0000	0,0000	0,00	1,00
periodexinter	0	0,0000	0,0000	0,00	1,00
	>12 et 0	0,3481	0,0682	26,07	1,42
	>12 et >12	0,5829	0,1086	28,81	1,79
Scale		0,3176	0,0000		1,37

Modèle fréquence globale

PARAMETER	LEVEL1	ESTIMATE	STDERR	CHISQ	EXP(ESTIMATE)
Intercept		-1,09	0,05	402,12	0,33
	18-26ans	-	-	-	1,00
	27-49ans	0,05	0,02	4,11	1,05

age_cond	>=50ans	0,10	0,04	6,01	1,11
	<=5ans	-	-	-	1,00
anc_perm]5-20]	-0,25	0,02	111,70	0,78
]20-30]	-0,26	0,03	65,39	0,77
	>30	-0,43	0,05	73,56	0,65
anc_veh	<=2ans	-	-	-	1,00
	>20	-0,68	0,05	203,13	0,51
]10-14]	-0,27	0,03	72,68	0,76
]14-20]	-0,34	0,03	101,38	0,71
]2-10]	-0,15	0,03	24,60	0,86
classe	A-I	0,05	0,03	2,37	1,05
	J-L	0,07	0,03	6,77	1,08
	M-S	-	-	-	1,00
fract	A	-	-	-	1,00
	AUTRES	0,12	0,02	25,95	1,13
coef_bm	50	-0,43	0,03	161,89	0,65
	100	-	-	-	1,00
	>100	-0,25	0,08	10,47	0,78
]51-100[-0,34	0,02	221,88	0,71
groupe	20-27	-0,35	0,04	90,06	0,70
	28-31	-0,24	0,03	80,37	0,78
	32-36	-	-	-	1,00
parking	VOIE PUBLIQUE	-	-	-	1,00
	GARAGE	-0,12	0,02	49,24	0,89
zone	2	-0,60	0,03	344,50	0,55
	3	-0,35	0,02	203,82	0,70
	4	-0,17	0,03	34,25	0,84
	>=5	-	-	-	1,00
natxresp	SANS	-	-	-	1,00
	1 RC	0,05	0,03	3,46	1,05
	1 ou 2 SIN NON RC	0,22	0,05	20,48	1,25
Scale		0,51	-		1,25

Modèle fréquence IDA Passif

PARAMETER	LEVEL1	ESTIMATE	STDERR	CHISQ	EXP(ESTIMATE)
Intercept		-2,61	0,07	1 342,25	0,07
age_cond	18-26ans	0,01	0,03	0,05	1,01
	27-49ans	-	-	-	1,00
	>=50ans	-0,08	0,04	2,93	0,93

anc_perm	<=5ans	-	-	-	1,00
]5-20]	-0,33	0,03	146,32	0,72
]20-30]	-0,27	0,04	50,01	0,76
	>30	-0,24	0,06	16,39	0,79
anc_veh	<=2ans	-0,66	0,06	107,49	0,52
]2-10]	-0,15	0,04	12,90	0,86
]10-14]	0,15	0,04	13,47	1,16
]14-20]	0,33	0,04	66,31	1,39
coef_bm	>20	-	-	-	1,00
	50	-0,60	0,04	240,81	0,55
]51-100[-0,55	0,03	470,70	0,58
	100	-	-	-	1,00
classe	>100	-0,30	0,09	11,53	0,74
	A-I	0,09	0,04	5,88	1,10
	J-L	0,22	0,04	35,87	1,24
	M-S	-	-	-	1,00
fract	A	-	-	-	1,00
	AUTRES	0,28	0,03	72,90	1,32
	20-27	-0,16	0,04	13,32	0,85
	28-31	-0,18	0,03	27,17	0,84
groupe	32-36	-	-	-	1,00
	GARAGE	-0,15	0,02	46,42	0,86
	VOIE PUBLIQUE	-	-	-	1,00
parking	2	-0,83	0,04	478,09	0,44
	3	-0,52	0,03	330,83	0,60
	4	-0,33	0,03	95,14	0,72
	>=5	-	-	-	1,00
zone	SANS	-	-	-	1,00
	1 RC	-0,23	0,04	37,47	0,80
	1 ou 2 SIN NON RC	0,13	0,06	4,12	1,14
natxresp					
Scale		0,31	-		1,36

Modèle fréquence IDA Actif

PARAMETER	LEVEL1	ESTIMATE	STDERR	CHISQ	EXP(ESTIMATE)
Intercept		-3,7151	0,0788	2223,58	0,02
age_cond	18-26ans	0,0000	0,0000	0,00	1,00
	27-49ans	0,0062	0,0276	0,05	1,01
	>=50ans	-0,1640	0,0520	9,94	0,85

anc_perm	>30	-0,4435	0,0592	56,10	0,64
	<=5ans	0,0000	0,0000	0,00	1,00
]5-20]	-0,2451	0,0263	87,01	0,78
]20-30]	-0,1998	0,0349	32,83	0,82
anc_veh	<=2ans	1,3102	0,0648	408,88	3,71
]2-10]	1,1005	0,0593	344,34	3,01
]10-14]	0,8154	0,0611	178,10	2,26
]14-20]	0,5374	0,0623	74,49	1,71
fract	>20	0,0000	0,0000	0,00	1,00
	A	0,0000	0,0000	0,00	1,00
	AUTRES	0,1791	0,0274	42,57	1,20
groupe	20-27	-0,4122	0,0366	126,50	0,66
	28-31	-0,2878	0,0250	132,77	0,75
	32-36	0,0000	0,0000	0,00	1,00
parking	VOIE PUBLIQUE	0,0000	0,0000	0,00	1,00
	GARAGE	-0,1101	0,0196	31,71	0,90
	2	-0,4542	0,0391	135,01	0,63
zone	3	-0,1573	0,0298	27,77	0,85
	4	-0,0012	0,0349	0,00	1,00
	>=5	0,0000	0,0000	0,00	1,00
	1 RC	0,1593	0,0284	31,49	1,17
natxresp	1 ou 2 SIN NON RC	0,3915	0,0504	60,43	1,48
	SANS	0,0000	0,0000	0,00	1,00
Scale		0,3447	0,0000		1,00

Modèle global

PARAMETER	LEVEL1	ESTIMATE	STDERR	CHISQ	EXP(ESTIMATE)
INTERCEPT		0,33	2 613,38	0,07	0,02
	18-26ans	1,00	1,00	1,01	1,00
	27-49ans	1,05	1,00	1,00	1,01
	>=50ans	1,11	1,00	0,93	0,85
age_cond	<=5ans	1,00	1,00	1,00	1,00
]5-20]	0,78	1,00	0,72	0,78
]20-30]	0,77	1,00	0,76	0,82
	>30	0,65	1,00	0,79	0,64
anc_perm	<=2ans	1,00	1,00	0,52	3,71
]2-10]	0,86	1,10	0,86	3,01
]10-14]	0,76	0,91	1,16	2,26
]14-20]	0,71	0,92	1,39	1,71

classe	>20	0,51	0,74	1,00	1,00
	A-I	1,05	0,76	1,10	1,00
	J-L	1,08	0,79	1,24	1,00
	M-S	1,00	1,00	1,00	1,00
fract	A	1,00	1,00	1,00	1,00
	AUTRES	1,13	1,00	1,32	1,20
coef_bm	50	0,65	1,00	0,55	1,00
	[51-100[0,71	1,00	0,58	1,00
	100	1,00	1,00	1,00	1,00
	>100	0,78	1,00	0,74	1,00
groupe	20-27	0,70	0,80	0,85	0,66
	28-31	0,78	1,01	0,84	0,75
	32-36	1,00	1,00	1,00	1,00
	VOIE PUBLIQUE	1,00	1,00	1,00	1,00
parking	GARAGE	0,89	1,00	0,86	0,90
	2	0,55	0,76	0,44	0,63
zone	3	0,70	0,76	0,60	0,85
	4	0,84	0,71	0,72	1,00
	>=5	1,00	1,00	1,00	1,00
	SANS	1,00	1,00	1,00	1,17
natxresp	1 RC	1,05	1,00	0,80	1,48
	1 ou 2 SIN NON RC	1,25	1,00	1,14	1,00
	0	1,00	0,81	1,00	1,00
periodexinter	>12 et 0	1,00	1,00	1,00	1,00
	>12 et >12	1,00	1,00	1,00	1,00
Scale		1,25	1,37	1,36	1,00

Annexe 5

Pondération du modèle issu du réseau de neurones

```

> library(e1071)
> optimrbis=best.nnet(Primern~., data=train, tunecontrol = tune.control(nrepeat = 5), size=1:20, decay=1:3)
> summary(optimrbis)
a 21-20-1 network with 461 weights
options were - decay=1
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1  i10->h1
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h1 i12->h1 i13->h1 i14->h1 i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2  i9->h2  i10->h2
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h2 i12->h2 i13->h2 i14->h2 i15->h2 i16->h2 i17->h2 i18->h2 i19->h2 i20->h2 i21->h2
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h3  i1->h3  i2->h3  i3->h3  i4->h3  i5->h3  i6->h3  i7->h3  i8->h3  i9->h3  i10->h3
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h3 i12->h3 i13->h3 i14->h3 i15->h3 i16->h3 i17->h3 i18->h3 i19->h3 i20->h3 i21->h3
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h4  i1->h4  i2->h4  i3->h4  i4->h4  i5->h4  i6->h4  i7->h4  i8->h4  i9->h4  i10->h4
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h4 i12->h4 i13->h4 i14->h4 i15->h4 i16->h4 i17->h4 i18->h4 i19->h4 i20->h4 i21->h4
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h5  i1->h5  i2->h5  i3->h5  i4->h5  i5->h5  i6->h5  i7->h5  i8->h5  i9->h5  i10->h5
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h5 i12->h5 i13->h5 i14->h5 i15->h5 i16->h5 i17->h5 i18->h5 i19->h5 i20->h5 i21->h5
    0.13    0.10    0.02    0.03    0.09   -0.05   -0.06    0.02    0.00    0.11    0.05
  b->h6  i1->h6  i2->h6  i3->h6  i4->h6  i5->h6  i6->h6  i7->h6  i8->h6  i9->h6  i10->h6
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h6 i12->h6 i13->h6 i14->h6 i15->h6 i16->h6 i17->h6 i18->h6 i19->h6 i20->h6 i21->h6
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h7  i1->h7  i2->h7  i3->h7  i4->h7  i5->h7  i6->h7  i7->h7  i8->h7  i9->h7  i10->h7
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h7 i12->h7 i13->h7 i14->h7 i15->h7 i16->h7 i17->h7 i18->h7 i19->h7 i20->h7 i21->h7
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h8  i1->h8  i2->h8  i3->h8  i4->h8  i5->h8  i6->h8  i7->h8  i8->h8  i9->h8  i10->h8
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00
i11->h8 i12->h8 i13->h8 i14->h8 i15->h8 i16->h8 i17->h8 i18->h8 i19->h8 i20->h8 i21->h8
    0.13    0.10    0.02    0.03    0.09   -0.04   -0.06    0.02    0.00    0.11    0.05
  b->h9  i1->h9  i2->h9  i3->h9  i4->h9  i5->h9  i6->h9  i7->h9  i8->h9  i9->h9  i10->h9
    0.13    0.06    0.00    0.00    0.00    0.13    0.00    0.00    0.13    0.00    0.00

```
