Mémoire présenté le :

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

Par : Julie Lavenu

Titre **Les méthodes de Machine Learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ?**

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 Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : Run Services/ Axa France

Signature :

Directeur de mémoire en entreprise :

Nom : Vanessa Roger

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 :

Institut de Science Financiere et d'Assurances

Les méthodes de Machine Learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ?

Julie Lavenu

11/04/2016

Résumé

Le marché IARD a subi ces dernières années de nombreuses transformations. Les changements réglementaires au travers de la loi Hamon et le développement de comparateur sur internet exacerbent la concurrence entre assureurs. Dans un marché qui devient de plus en plus arbitré par les prix, ces derniers doivent s'adapter en segmentant leur tarif.

Pour autant, une segmentation trop fine pourrait remettre en cause le principe de mutualisation des risques et donc la notion même d'assurance. La question étant ainsi de définir le niveau de segmentation le plus adapté.

En assurance automobile, les acteurs du marché ont recours à un principe de regroupement de véhicules en risques homogènes nommé véhiculier. Le véhiculier constitue aujourd'hui l'un des fondements de la tarification *a priori*. Pourtant, sa théorie a été peu développée dans la littérature actuarielle, tout particulièrement pour les véhicules de type "motos", qui constituent le périmètre de ce mémoire. Ceci s'explique en partie par le fait que les véhiculiers sont aujourd'hui souvent construits sur l'avis d'experts pour ce type de véhicules.

De manière à anticiper le développement du Big Data, et dans la perspective d'identifier une segmentation au plus juste, sur la base de données que l'expert ne saurait, à lui seul, exploiter, ce mémoire a pour objectif d'élaborer un véhiculier non pas à dire d'experts mais à dire de machines.

Dans un premier temps, deux approches pour établir un véhiculier à dire de machines sont développées en s'appuyant sur des travaux initiateurs existants. La première approche consiste à construire un arbre CART à l'appui d'une utilisation singulière de la théorie de la crédibilité. La seconde approche associe un autre algorithme de Machine, le Randon Forest, à un lissage spatial.

Ensuite, les données utilisées sont présentées et fiabilisées. A partir de celles-ci des modèles GLM sont construits en extrayant la part du risque liée au véhicule.

Enfin, à partir de l'effet véhicule extrait, des véhiculiers à dire de machines issus des deux approches sont construits puis comparés aux véhiculiers à dire d'experts, nous permettant alors de conclure quant au degré d'intervention pertinent des experts dans l'élaboration d'un véhiculier et donc d'une segmentation au plus juste du risque.

Abstract

In the recent years, the P& C market has passed through many changes. The competition between insurance companies is enhanced by regulatory changes (through the Hamon law for instance) and the development of online comparers. In an increasingly price arbitrated market, insurers have to adapt by segmenting their tariff.

However, too much segmentation may call into question the principle of pooling risk and therefore the insurance concept. So insurers must determine the most appropriated level of segmentation.

Cars and motorcycles insurances create a homogeneous risk pooling by clustering vehicles. This vehicle classification is one of a priori pricing foundations. However, it should be noted that vehicle classifications were poorly developed in the actuarial literature, especially for " motorcycles " type vehicles, which are the main subject of this study. This is partly explained by the fact that vehicle classification is now often built on expert judgement.

In order to anticipate the development of the Big Data, and with a purpose to identify accurate segmentation and given big datasets available that the expert opinion can't operate by his own, this paper aims to develop a vehicle classification not according to an expert opinion but according to machine learning.

First, two approaches of vehicle classification by machines are built on previous studies results. The first approach consists on building a CART decision tree with the support of credibility theory. The second approach combines random forest and spatial smoothing.

Then, data processing enables the use of a GLM model in order to extract claim costs related vehicle only.

Finally, from the effect vehicle extract, vehicle classifications according machines were built with the two methods, then compared with vehicle classifications according to expert opinions. It allows us to conclude on the appropriate level of expert intervention in the development of vehicle classification and, therefore, to define a risk-adapted segmentation.

Remerciement

Je tiens à remercier Hervé Gicquel, directeur général de Club 14 et Delphine Maisonneuve, directeur du Marché IARD d'Axa France particuliers, pour m'avoir donné l'opportunité de réaliser ce mémoire.

Je remercie l'ensemble de l'équipe actuariat Auto pour leur bonne humeur et le partage de leur expertise et tout particulièrement Thomas Gauthron et Vanessa Roger, tutrice en entreprise.

J'adresse mes remerciements à Pierre Thérond, le tuteur pédagogique de ce mémoire, pour son encadrement ainsi qu'à Frédéric Planchet pour ses avis techniques éclairants et sa disponibilité.

Je voudrais également dire merci à Denis Bourgeois et Anne-Sophie Musset, pour m'avoir aidé à finir ce mémoire, pour leurs précieux conseils et leur bienveillance.

Sans oublier les violets qui m'ont soutenus pendant la dernière ligne droite, merci pour votre implication, vos avis avertis et vos relectures.

Enfin, j'ai une pensée particulière pour mes proches et Adrien, merci pour votre compréhension et vos encouragements.

Table des matières

Introduction

4 à 1. C'est en mars 2016 que la machine a battu l'homme au jeu de Go, jeu pourtant reconnu comme étant impossible à programmer pour être en situation d'égaler le cerveau humain et encore moins celui des spécialistes mondiaux de ce jeu.

Depuis le « test de Turing » [1] en 1950 qui soulève la question de la capacité des machines à imiter une conversation humaine, l'Intelligence Artificielle trouve des applications dans de nombreux domaines, allant de la robotique, à l'imagerie en passant par l'art où des algorithmes sont aujourd'hui capables de transformer n'importe quelle photo selon le style et l'art des plus grands peintres.

L'émergence du Big Data incite de plus en plus à l'intervention du Machine Learning dans les modèles pour exploiter la multitude de données associées et ainsi capter leurs richesses. Le monde de l'actuariat compte bien sûr profiter d'une telle opportunité.

La segmentation du risque devient aujourd'hui incontournable dans l'environnement concurrentiel induit par une course aux prix toujours plus attractifs et une profitabilité également mise à mal par la recrudescence de réformes réglementaires. En particulier, la mise en place de la loi Hamon et la possibilité de résilier à tout moment après un an de souscription viennent perturber l'équilibre et la stabilité tarifaire notamment en assurance IARD (Incendie Accidents et Risques Divers).

Pour autant, une segmentation trop importante pourrait remettre en cause les principes de mutualisation et donc la notion même d'assurance. Il s'agit donc pour les assureurs d'identifier la segmentation la plus pertinente possible.

Pour assurer des automobiles ou des motos, les acteurs du marché de l'assurance ont recours à un principe de regroupement de véhicules en risques homogènes qui se matérialise par un véhiculier. L'approche de classification par le véhiculier est, en effet, un des fondements de la tarification dite *a priori*. Il est à noter que les véhiculiers ont été peu développés dans la littérature actuarielle, en particulier en ce qui concerne les véhicules de type "motos", qui constituent le périmètre même de ce mémoire.

La classification des véhicules et la méthodologie qui l'accompagne sont considérées comme des éléments concurrentiels importants par les assureurs qui souhaitent les préserver confidentielles. Ceci s'explique notamment par le fait que jusqu'à présent les véhiculiers ont été majoritairement réalisés sur avis d'experts.

1. Alan Turing

D'ailleurs, un assureur auto/moto peut choisir aujourd'hui d'utiliser tout simplement le véhiculier SRA [2], réalisé à partir des données de place en concertation avec les experts du secteur.

Récemment, des approches introduisant les machine learning ont vu le jour, notamment avec les travaux SIPULSKYTE. Cependant l'intervention de l'avis d'experts est très fréquemment requise pour finaliser la constitution du véhiculier et ainsi palier le manque d'apprentissage de la méthode.

De manière à anticiper le développement du Big Data, et dans la perspective d'identifier une segmentation au plus juste sur la base d'un nombre de données que l'avis d'experts ne saurait, à lui seul, exploiter, ce mémoire a pour objectif d'élaborer un véhiculier non pas à dire d'experts mais à dire de machines. Il s'agira plus précisément de définir un modèle qui permette de réaliser un véhiculier, et donc d'identifier les groupes de risques homogènes ou segmentations les plus pertinents en regard des garanties couvertes, sans l'intervention d'avis d'experts dans la constitution des résultats. C'est en tout cas la problématique initiée par Axa et le courtier Run Services disposant d'un véhiculier construit exclusivement à dire d'experts à l'appui de Club14 (C14), une association de passionnés de moto.

La ***première partie*** permettra de présenter les deux approches construites à l'appui des travaux initiateurs existants pour élaborer un véhiculier à dire de machines. La première approche consiste à appliquer un arbre CART et la théorie de la crédibilité et la seconde approche, un Random Forest puis un lissage spatial.

Nous verrons dans une ***deuxième partie*** que ces approches s'appuient de manière équivalente sur une exploitation poussée de données, une analyse approfondie des variables explicatives, et surtout sur un isolement de la part véhicule par le biais d'un modèle GLM. La qualité des données étant primordiale au bon aboutissement d'une telle étude, la base utilisée issue des données du produit moto co-proposé par Axa France et Run Services a, au préalable, fait l'objet de nombreux retraitements pour assurer sa fiabilité.

Il est bien entendu que l'objectif poursuivi de constituer des véhiculiers à dire de machines exclusivement peut être considéré comme extrême. Pour autant, cette approche permettra dans une ***troisième partie*** d'analyser non seulement les deux véhiculiers à dire de machines mais de les comparer aux véhiculiers à dire d'experts, tant celui de SRA que celui de C14, et d'en conclure quant au degré d'intervention pertinent des experts dans l'élaboration d'un véhiculier et donc d'une segmentation au plus juste du risque.

2. SRA (Sécurité et Réparation Automobiles) est une association créée en 1977 dont toutes les entreprises d'assurances automobiles sont adhérentes

Chapitre 1

Comment établir un véhiculier à dire de machines ?

En assurance auto/moto, il n'est pas rare de constater des écarts tarifaires importants dû à un zonier ou à un véhiculier obsolète ou insuffisamment segmenté. Le zonier, homologue du véhiculier dans le cas de regroupement de zones géographiques, a fait l'objet ces dernières années de diverses études qui ont permis de développer de nouvelles méthodes et d'affiner la segmentation géographique. A contrario, peu d'études ont été menées jusqu'à présent sur le véhiculier, ce qui explique en partie la prédominance de l'avis d'experts dans cette classification. Cette première partie se propose d'expliciter des méthodes pour élaborer un véhiculier à dire de machines, c'est-à-dire avec un minimum d'intervention d'avis d'experts.

1.1 Les enjeux de la segmentation

1.1.1 Le principe de l'assurance : entre mutualisation et segmentation du risque

A l'époque antique, pour se prémunir des caprices de la mer, les Phéniciens développèrent le *prêt à la grosse aventure* : un particulier consentait alors à un marchand maritime une avance dont le remboursement avec intérêt était conditionné à l'arrivée à bon port de la marchandise. Afin d'éviter de perdre la totalité des montants prêtés, les investisseurs particuliers s'engageaient dans des transactions relatives à différentes expéditions. Des siècles plus tard, la **mutualisation** du risque constitue toujours la clé de voute de l'assurance.

Cependant, « la mutualisation qui peut être vue comme une relecture actuarielle de la loi des grands nombres n'a de sens qu'au sein d'une population de risque homogène », comme le constate A. Charpentier (2015), professeur d'actuariat à Montréal reconnu sur ce domaine.

En effet, la mutualisation tout azimut engendre le **risque d'anti-sélection**. G. Akerlof, prix nobel de l'économie en 2011, a été l'un des premiers à populariser ce principe dans les années

70 avec son étude[1] sur le marché de l'automobile d'occasion aux Etats Unis, composé de voitures de qualités inégales. Il a pu démontrer qu'in fine, un acheteur sur ce marché, focalisant principalement son choix sur le prix, était amené à favoriser les voitures de mauvaises qualités que les vendeurs bradaient délibérément pour se prémunir de vices cachés. Une telle situation entrainera alors le retrait de vendeurs de voitures de bonnes qualités.

Afin de lutter contre ce type d'anti-sélection des « bons » risques résultant d'ailleurs de l'asymétrie de l'information entre l'assuré et l'assureur, les assureurs affinent l'appréhension du risque et adaptent la prime au profil de risque des assurés. C'est ce qu'on appelle la **segmentation du risque**.

1.1.2 L'intérêt de la segmentation

Dans un contexte concurrentiel, une segmentation adaptée au portefeuille est nécessaire pour assurer une rentabilité suffisante à toute société d'assurance. En effet, l'assureur est chargé d'apprécier et de quantifier le risque auquel l'assuré est exposé, de décider du niveau et type de couverture consentis, ainsi que du montant de primes à facturer pour faire face à son engagement. En ignorant tout phénomène d'anti-sélection, proposer une prime (prime commerciale) unique aux assurés est moins rentable que de distinguer les "bons" et "mauvais" risques.

Le marché IARD a subi ces dernières années de profondes transformations ; le marché étant de plus en plus dicté par "l'effet prix". Un des facteurs à l'origine de ce phénomène est le développement de comparateurs sur internet qui accentuent la concurrence entre les assureurs. Cette concurrence est également exacerbée par de récents changements règlementaires et notamment la loi Consommation, dite loi Hamon, qui permet aux assurés, depuis le 1er janvier 2015, de résilier leur assurance automobile, moto ou habitation à n'importe quel moment après un an d'engagement.

Face à un marché qui repose d'abord sur le prix, la segmentation devient un véritable enjeu. Non seulement, la segmentation permet à un assureur d'ajuster ses prix au plus juste d'un groupe homogène considéré mais elle le prémunit du risque d'anti-sélection, à savoir des mauvais risques que la concurrence aura été capable de détecter par une segmentation avisée.

Enfin, sur ce type de marché, l'approche moyenne (ajustement du niveau global des ressources) ne suffit pas à assurer la profitabilité. La sélection positive protège la profitabilité à court terme et génère de la croissance à moyen et long terme.

1. The Market for Lemons : Quality Uncertainty and the Market Mechanism

1.1.3 Quel niveau de segmentation pertinent ?

L'assurance connectée, le Big Data et le développement des données en open source participent à la modification de l'environnement assurantiel. En effet, l'assureur développe ou utilise de nouveaux canaux (applications smartphones et tablettes numériques, réseaux sociaux, etc.) afin de récolter de nouvelles données sur les assurés. Les offres telles que les assurances « Pay as you drive », ou leur équivalent pour les assurances multirisques habitation « Pay as you live» mettent en évidence l'implantation de ce processus dans la stratégie des assureurs qui devrait s'accroitre ces prochaines années.

Il existe un paradoxe entre la tendance actuelle qui favorise, avec le déploiement du Big Data, un degré de segmentation de plus en plus fin et la mutualisation qui est le fondement même de l'assurance. En effet, le gain en segmentation entraine inévitablement une diminution du niveau de mutualisation.

Le niveau de segmentation doit, par ailleurs, tenir compte des contraintes réglementaires [2] qui pourraient s'intensifier dans les années à venir.

Enfin, il faut souligner qu'un niveau de segmentation élevé augmente le risque d'erreur de modélisation, ou de manière plus large ce qu'on appelle le risque opérationnel. Alors qu'un modèle relativement simple sera, au contraire plus fiable. Le choix de complexification d'un modèle doit donc tenir compte de l'augmentation du risque d'instabilité et de la difficulté à le gérer dans le temps.

Dans un environnement concurrentiel, le choix du degré de segmentation et celui de mutualisation est une question délicate au cœur du métier d'actuaire. Dans le cadre de l'assurance Auto/Moto, cette segmentation passe notamment par l'élaboration d'un véhiculier.

2. Une réglementation existe en matière de segmentation : l'interdiction de tarifer selon le sexe de l'assuré suite à l'arrêt de la Cour de Justice de l'Union européenne (CJUE) en mars 2011 est un exemple de contrainte légale qui oblige les assureurs pour des raisons éthiques et déontologiques à utiliser des proxys alors que la variable sexe s'est pourtant révélée discriminante en assurance auto/moto dans diverses études.

1.2 Les critères imposés dans l'élaboration du véhiculier à dire de machines

Avant de présenter dans le détail l'objectif qui est de refondre le véhiculier en place avec une intervention la plus limitée possible d'experts, il apparait important d'introduire dès à présent le produit étudié ainsi que les principes détaillés d'un véhiculier.

1.2.1 Le produit étudié et ses spécificités

Sur le marché de l'assurance deux roues, obligatoire depuis 1958, la concurrence est d'autant plus intense que le parc moto augmente en moyenne plus rapidement que son grand frère, le marché de l'assurance automobile (+1,8% contre +1,1% en 2013 [3]). Par ailleurs, le parc assurable des deux roues a subi ces dernières années de fortes mutations. Composé de plus de 50% de cyclomoteurs en 1994, le parc est aujourd'hui majoritairement constitué de motos de cylindrée supérieure à 125cm^3. Du reste, le marché des deux roues est en constante évolution avec l'arrivée annuelle de nouveaux modèles. Afin que le véhiculier soit en adéquation avec l'expérience et la composition du portefeuille, la segmentation des véhicules doit être suivie et contrôlée.

Cette étude est basée sur le produit co-proposé par Axa France et le courtier Run Services qui ont noué un partenariat avec Club 14 (C14), une association de passionnés de moto. L'offre proposée par Axa/C14, première société d'assurance et deuxième acteur du marché avec 15,5% de part de marché en 2013, s'articule autour de trois produits : le produit moto, le produit véhicule de collection et le produit cyclomoteur. Nous nous intéressons, dans le cadre de ce mémoire, au produit moto.

FIGURE 1.1 – Les différents véhicules couverts dans le produit moto, Source C14

3. FFSA-GEMA (2015) "Étude - l'assurance des deux roues en 2013

Contrairement à ce que son appellation pourrait laisser supposer, le produit moto ne concerne pas seulement les véhicules à deux roues, mais également les trois roues (telles que les scooters à trois roues et les trikes) et certaines catégories de quatre roues à moteurs (de type quad, ssv et buggy). Les différentes catégories de véhicules couvertes dans le produit moto sont présentées figure 1.1 [4].

Le produit se décompose en quatre formules de souscription, chacune étant composée de garanties de base et de garanties optionnelles. Les conditions d'acceptation du risque reposent en particulier sur des critères d'âge et d'adéquation entre le profil du pilote et son véhicule.

Par ailleurs, grâce à la combinaison de politiques de souscription et de la stratégie marketing, les populations assurées du portefeuille, tous sinistres confondus, ont une faible sinistralité (nettement inférieur à 10% [5]).

1.2.2 Qu'est-ce qu'un véhiculier ?

Un véhiculier est un regroupement de modèles ou versions de véhicules en groupes homogènes de risque dont la forme peut varier.

La modélisation du produit étudié est basée sur les Modèles Linéaires Généralisés (GLM) qui, par croisement des variables explicatives utilisées, forment des cases tarifaires. Le nombre de variables sélectionnées et le nombre de modalités de chacune impactent directement le nombre de cases tarifaires. **Plus le nombre de cases tarifaires est important, plus le tarif sera segmentant et plus la population représentative de chaque case tarifaire sera faible. Compte tenu de l'enjeu évoqué précédemment, il convient d'arbitrer entre le nombre de modalités et l'incertitude de modélisation en sélectionnant avec soin les variables explicatives et en regroupant les modalités d'une variable si nécessaire.**

Dans le cadre de la tarification auto ou moto, le véhicule assuré est une variable explicative dont la modalité correspond au modèle ou à la version du véhicule. Le modèle du véhicule assuré possède un nombre de modalités colossal (de l'ordre de milliers à des dizaines de milliers) qui ne permet pas de l'inclure en tant que telle dans un modèle tarifaire. Les modèles de véhicules doivent alors être regroupés afin de former des groupes homogènes. Le choix du regroupement des modèles/versions de véhicules s'appelle un **véhiculier**. Usuellement, les variables tarifaires utilisées pour l'élaboration de la prime pure se décomposent selon la figure 1.2.

4. Il n'existe pas d'unique classement des types de véhicules, il s'agit ici de la typologie de véhicule selon C14
5. Pour des raisons de confidentialité et à la demande d'Axa, la valeur moyenne de fréquence de sinistres n'est pas indiquée dans le corps de ce mémoire.

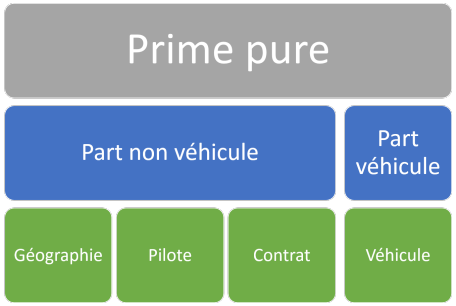


FIGURE 1.2 – Décomposition de la prime pure

Il est à noter que, dans le cas présent, l'assureur peut décider d'appliquer ou non la clause relative au bonus-malus conformément à l'article R. 311-1 du Code de la Route. Selon les experts C14, le bonus-malus ne donne pas une information fiable sur la conduite du pilote. Ainsi, cette variable a été exclue de l'étude.

D'un point de vue tarifaire, le produit étudié a également la particularité de disposer d'une double expertise, celle de l'équipe de C14 en termes de moto et celle du service actuariel d'Axa chargée de la tarification. Si la combinaison de ces deux expertises est une réelle force pour la tarification du produit, elle peut également se révéler être à double tranchant dans le cas où les expertises ne convergent pas.

1.2.3 Le véhiculier en place du produit moto étudié

Pour poursuivre la présentation d'un véhiculier, intéressons-nous à présent aux spécificités du véhiculier en place.

A - Son principe

C14 effectue sa propre classification des véhicules selon deux catégories : le groupe (V_{Groupe}) et la classe (V_{Classe}). Ces deux catégories sont utilisées dans la tarification. Le V_{Groupe} est utilisé comme un prédicteur de la fréquence de sinistres des véhicules alors que la V_{Classe} est utilisée comme un indicateur de la sévérité des sinistres. Un exemple fictif de véhiculier est présenté figure 1.3.

	Véhiculier	
	Classe	Groupe
Yamaha MT-07	M	36
Kawasaki Z800	J	32
Yamaha MT-09	K	38

FIGURE 1.3 – Exemple fictif de véhiculier

Ainsi, lors de comités de classification, un collège d'experts définit pour chaque modèle de véhicules un groupe (V_{Groupe}) et une classe (V_{Classe}) en fonction de ses caractéristiques techniques.

Ce classement des véhicules est revu périodiquement en fonction des résultats observés sur les véhicules. Suite à cette révision, certains véhicules peuvent changer de classement.

B - Ses limites

Pour autant, ce véhiculier présente de nombreuses limites :

- Il a été établi il y a de nombreuses années et n'a pas été mis à jour depuis, même si des révisions ont lieu régulièrement comme explicité précédemment.
- Une étude a été menée en 2012 sur le portefeuille et a pu montrer qu'au niveau global la répartition des véhicules selon les variables de classement ne sont pas homogènes (cf figure 1.4).
- Cette même étude menée pour les scooters met en évidence tant pour le V_{Groupe} que la V_{Classe} des incohérences (amplitude abberrante des caractéristiques véhicules - prix du véhicule, cylindrée - au sein du même groupe). Ce qui a été étudié pour les scooters se généralise au reste des véhicules du produit moto.

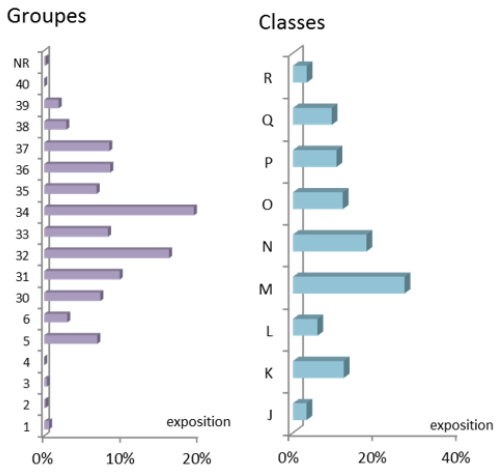


FIGURE 1.4 – Répartition des V_{Groupe} et V_{Classe} du portefeuille 2012

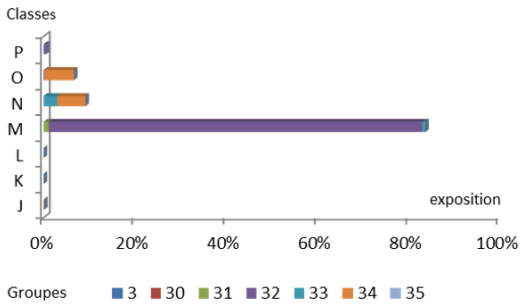


FIGURE 1.5 – Répartition des scooters du portefeuille 2012

La figure 1.5 met en évidence que les scooters sont, tout particulièrement, insuffisamment segmentés.

Au vu de l'ensemble de ces élements, il apparait nécessaire de revoir le véhiculier en place. Pour se faire, plutôt que de s'appuyer sur l'avis d'experts, l'objectif est de challenger la construction d'un véhiculier par le biais au plus exclusif des modèles et de la « machine ». Dans la poursuite de cet objectif, la partie qui suit vise notamment à détailler les contraintes imposées.

1.2.4 Le véhiculier à construire : objectif et contraintes imposées

A - Les enjeux et contraintes imposées pour la construction du véhiculier

L'objectif est donc de construire un nouveau véhiculier à dire de « machines », faisant intervenir un minimum l'avis d'experts, pour notamment évaluer la pertinence des modèles mathématiques dans l'explication de la multitude des données disponibles.

Ce nouveau véhiculier sera non seulement comparé au véhiculier C14 en place, élaboré, comme évoqué précédemment, sur avis d'experts, mais également confronté au véhiculier SRA élaboré à l'appui de modèles mathématiques mais aussi par concertation des différents acteurs ou experts de la profession, à partir de données de place. Il est à noter que pour ce dernier véhiculier, il n'a pas été possible d'accéder aux méthodes confidentielles utilisées par SRA mais nous avons tout de même pu disposer des règles de classement des véhicules, qui suffiront à l'étude.

Tant le véhiculier C14 que SRA font l'objet d'une classification par groupe et classe que l'on nommera respectivement (V_{Groupe} et V_{Classe}) et ($V_{Groupe-SRA}$ et $V_{Classe-SRA}$). **Une première contrainte imposée est donc de reprendre le format des véhiculiers existants : être constitué d'exactement deux variables de classement construites indépendamment de la garantie, soit un classement des véhicules selon la fréquence de survenance des sinistres (ci-après $V_{Groupe-new}$) et un classement des véhicules selon le coût des sinistres (ci-après $V_{Classe-new}$).**

Au delà du fait que la comparaison en sera facilitée, considérer une telle structure est nécessaire d'un point de vue économique : l'introduction d'une nouvelle variable dans le système informatique aurait un coût conséquent.

Il est à noter que le nombre de modalités par variable n'est par ailleurs pas restreint dans cette étude.

Comme dans toute classification, la question des véhicules « absents » dans les données initiales se pose. En effet, le portefeuille étudié, bien que pouvant être considéré comme représentatif des données de place, omet une partie de l'ensemble des modèles de véhicules en circulation. Sans compter que de nouveaux véhicules arrivent chaque année sur le marché, ce qui peut rendre complexe leur intégration dans un véhiculier. **Ainsi, la seconde contrainte imposée est que le nouveau véhiculier puisse proposer des façons d'intégrer ces véhicules « absents » ou futurs.**

En définitive, la qualité du nouveau véhiculier construit sous contraintes sera comparée à celle des véhiculiers C14 et SRA à l'appui de résultats statistiques selon deux critères : la réduction de l'erreur de prédiction et le caractère segmentant vis-à-vis des véhicules composant le portefeuille.

B - Les difficultés de l'étude *a priori*

Les contraintes imposées dans cette étude posent *a priori* certaines difficultés.

Tout d'abord, la modélisation de la variable d'intérêt (fréquence ou coût), toutes garanties confondues, est nécessairement associée à une certaine hétérogénéité dans les observations qui pourrait compromettre l'apprentissage de Machine Learning. Nous verrons dans la troisième partie de ce mémoire comment les théories sur la crédibilité peuvent être utilisées pour améliorer cet apprentissage.

Ensuite, les experts de par leurs connaissances du marché deux roues et des comportements des assurés sont capables de classer les véhicules en tenant compte d'effet de mode ou de cible marketing d'un modèle de véhicule donné. Ils peuvent dès lors choisir de classer un véhicule A plus à risque qu'un véhicule B qui aurait pourtant les même caractéristiques techniques (cylindrée, poids, puissance, catégorie, marque etc) car ils auront estimé que les conducteurs de véhicule A ont un comportement plus dangereux sur le route. Les Machine Learning quant à elle, auront probablement plus de mal à capter ce genre d'information non explicite dans une base de données. Ainsi, en classant les véhicules selon les caractéristiques véhicule techniques observables, les Machine Learning risquent de classer les véhicules A et B ensemble à défaut d'avoir l'information nécessaire pour les distinguer.

Dans ces conditions, les données apparaissent ainsi comme un sujet majeur au bon aboutissement de l'objectif fixé.

1.3 Les méthodes de classification des véhicules

L'effet véhicule est la part de variable d'interêt (fréquence ou coût) liée aux véhicules tel que présenté figure 1.2. Le couple "conducteur/véhicule" est en théorie indissociable. En effet, le comportement du conducteur impacte la façon dont le véhicule est conduit, et inversement, certaines caractéristiques du véhicule ont un impact sur le comportement du conducteur au "volant". De plus, d'après l'expérience de C14 en moto, certains modèles, toutes choses égales par ailleurs, attirent des profils plus risqués. C'est le cas de la Yamaha T MAX 125, les assurés conduisant ce modèle ont une probabilité de sinistres plus élevée que d'autres conduisant un modèle équivalent. Dans la pratique, la modélisation GLM peut permettre de dissocier "l'effet conducteur" de "l'effet véhicule". Elle sera d'ailleurs utilisée à cet effet dans la seconde partie de cette étude.

1.3.1 Les caractéristiques explicatives véhicule *a priori*

Une première approche de la significativité des caractéristiques véhicules aujourd'hui reconnue sur le marché vis-à-vis de la sinistralité est ici présentée.

Si aucune étude actuarielle concernant les facteurs de risque des motos n'a pu être identifiée, plusieurs publications sur le classement des automobiles ont été recensées. Malgré des périmètres d'applications différents, certains phénomènes observés dans le cas des automobiles sont tranposables dans une certaine mesure au produit moto :

- La marque du véhicule, indicateur du coût de réparation des véhicules, est un prédicteur partiel du coût des sinistres [6] ;
- La capacité du moteur (en litres) est une variable importante dans la modélisation du coût [7]. Cette variable s'est révélée être un bon candidat pour estimer la vitesse maximale et la puissance du véhicule. Les véhicules ayant une capacité moteur élevée seraient associés à une fréquence et un coût élevés ;
- Le couple puissance/poids est un bon prédicteur pour le coût des sinistres, que les variables soient utilisées en ratio ou bien séparément [8] ;
- Le type de véhicule (utilitaires, berline, etc.) est fortement corrélé à la sinistralité [9] ;
- Le nombre de portes, le type de transmissions ou encore le carburant semblent expliquer les performances maximales du véhicule.

Les différents travaux montrent qu'il n'existe pas de modèle holistique ou de combinaison de variables qui modéliserait parfaitement l'effet véhicule.

6. Sources OHLSSON (2008) et WENZEL et ROSS (2004)
7. Sources KIM et al.(2006) , OHLSSON (2008) et WANG et al.(2010)
8. Sources OHLSSON (2008) et KIM et al.(2006)
9. Sources LAWRENCE (2001), WENZEL et ROSS (2004) et KIM et al.(2006)

Côté moto, l'expertise de Axa et Run Services fait état :

- du "kilomètre départ arrêté" des motos dans la modélisation de fréquence de sinistres ;
- de l'utilisation conjointe des variables puissance et poids à privilégier par rapport au ratio puissance/poids.

Les règles de classification du véhiculier SRA donnent des indications sur les attentes a priori des variables véhicule. Ainsi, il est attendu :

- une corrélation positive respectivement entre la fréquence de sinistres et le ratio puissance/poids , et le coût des sinistres et la valeur du véhicule.

1.3.2 Quelques rappels théoriques

Maintenant que les notions de significativité de variables sont présentées, l'idée est de se diriger vers l'explication des variables véhicule pour les modèles mathématiques. Mais avant de présenter la méthodologie choisie et les travaux de référence sur laquelle elle s'appuie, quelques notions théoriques sont rappelées pour mieux appréhender le sujet.

A - GLM

Les modèles GLM, très utilisés par les compagnies d'assurances non-vie pour modéliser le risque, sont une généralisation de plusieurs modèles : la régression linéaire, l'analyse de variance et covariance. Ces méthodes, moins limitées que les méthodes de régression classiques, ont l'avantage de modéliser des variables aléatoires non nécessairement gaussiennes à l'aide de variables explicatives qualitatives ou quantitatives. De plus, les modèles linéaires simples ne permettent pas de construire un modèle où les coefficients tiennent compte des caractéristiques propres à chaque assuré dit modèle à coefficient correcteur. Si le modèle GLM permet l'utilisation de tests statistiques pour juger de la qualité du modèle, cela se fait au coût d'hypothèses fortes notamment sur la loi de la variable à expliquer. De plus, les modèles GLM peuvent théoriquement modéliser des comportements non linéaires, cependant ces interactions doivent être spécifiées *a priori*. En effet, une des limites du modèle GLM est la détection et la modélisation d'interactions entre les variables. Pour un modèle à 9 variables explicatives de 10 modalités chacune, il existe 1 milliard de possibilités. Ainsi, les méthodes d'apprentissages paramétriques permettent de prédire des modèles de façon précise, mais la structure de risque qu'ils modélisent ne correspond pas nécessairement à la réalité.

Les méthodes de statistiques d'apprentissage sont des alternatives à la statistique classique. Ces méthodes ne font pas d'hypothèses fortes sur la distribution des données à expliquer. L'unique hypothèse des méthodes sur les données à expliquer sont identiquement et aléatoirement générées par un processus à partir des variables explicatives. Par ailleurs, elles ont l'avantage de détecter les interactions entre les variables sans avoir à les spécifier au préalable. Les modèles les plus connus de Machine Learning sont les arbres de décisions, les réseaux de neurones, les Random Forest.

Par la suite, ces deux types de modélisation seront dévelopées. Le modèle GLM sera utilisé pour extraire l'effet véhicule ce qui constitutera le coeur du second chapitre. L'effet véhicule sera expliqué dans une troisième partie par des méthodes de Machine Learning évitant ainsi de faire des hypothèses fortes sur le comportement des variables véhicule.

B - Arbres CART

L'arbre de décision est devenu une méthode très prisée au vu de la rapidité de ses temps de calcul, de sa capacité à gérer tous types de variables et à selectionner les plus pertinentes, ainsi que la lisibilité et la facilité d'interprétation des résultats.

Il convient de distinguer deux types d'arbres de décision : les arbres de régression où la variable à expliquer est quantitative et les arbres de classement où la variable à expliquer est qualitative. La technique de l'arbre de régression est utilisée pour répartir les individus d'une population en k sous-populations et prédire la valeur cible pour chaque population. Il existe de nombreuses variantes d'algorithmes : CART (Classification And Regression Trees), CHAID (Chi-squared Automatic Interaction Detector), les algorithmes de classification supervisée (C5, C4.5, ID3). Nous nous intéressons dans le cadre de cette étude à l'arbre de régression CART.

L'arbre de régression CART, que nous utiliserons ici, construit des estimateurs constants par morceaux sur des partitions créées, à partir des données, par un découpage binaire récursif de l'ensemble des variables explicatives. L'algorithme commence par choisir la variable explicative, qui par ses modalités, découpe le mieux la population en deux groupes (nommés **noeuds**) en maximisant la variance inter-groupe. L'opération est réitérée jusqu'à ce qu'il n'y ait plus qu'un individu par groupe ou bien selon un critère d'arrêt à définir, obtenant les noeuds finaux appelés **feuilles**. Le coût prédit pour chaque feuille est la moyenne des valeurs cibles de chaque individu de la feuille.

La seconde étape consiste à minimiser une fonction prennant en compte l'erreur quadratique moyenne et le nombre de feuilles. Cette fonction permet d'optimiser le niveau de complexité de l'arbre de manière à prévenir le sur-apprentissage.

Lors de la troisième étape, l'arbre optimal est obtenu par élaguation selon le paramètre de complexité optimisé [10].

L'inconvénient principal des arbres de décision est que la classification dépend fortement de l'ordre des variables choisies ce qui peut nuire au pouvoir prédictif du modèle. Cette limite peut être rectifiée par des techniques de boosting ou bagging. Par la suite, les arbres vont être comparées à une méthode de bagging : le **Random Forest**. Nous verrons alors dans quelle mesure la modélisation Random Forest a un meilleur pouvoir prédictif que celle des arbres.

10. Pour de plus amples informations sur les régressions CART, se référer à la thèse de S. GEY (2002) Bornes de risque, détection de ruptures, boosting : trois thèmes statistiques autour de CART en régression

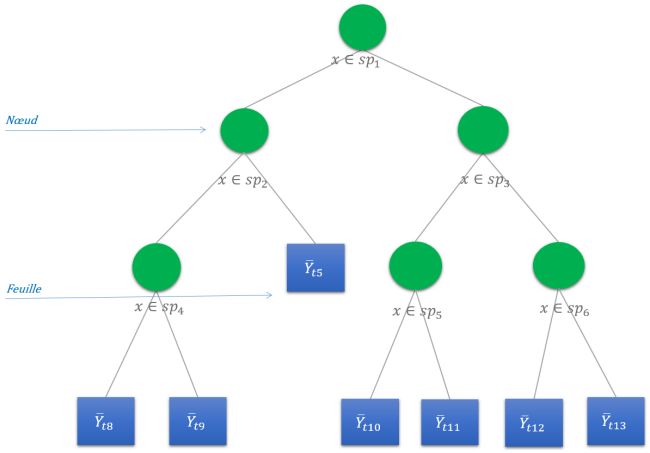


FIGURE 1.6 – Représentation d'un arbre de régression CART

C - Random Forest

Les forêts aléatoires ou Random Forest sont des cas particuliers de bagging pour les arbres de décision. L'algorithme consiste à construire une famille d'arbres de décision sur m échantillons boostrap, suivi de l'agrégation des prédictions des modèles. Le principe de l'algorithme est de chercher, pour chaque scission, non plus la meilleure scission parmi toutes variables explicatives (n), mais la meilleure scission pour p variables explicatives tirées aléatoirement parmi n. Cette double randomisation a été introduite par L. BREIMAN. [11]

Plusieurs paramètres sont à optimiser afin d'obtenir les meilleurs résultats : le nombre de variables sélectionnées par scissions, le nombre de feuilles de chaque arbre et le nombre d'arbres dans la forêt. Concernant le premier point, la littérature [12] préconise, dans le cas de problèmes de régressions, que le nombre de variables sélectionnées aléatoirement soit égal à la partie entière du tiers du nombre de variables explicatives et ne soit pas inférieur à 5. Vient ensuite la question du critère d'arrêt des arbres. Contrairement au bagging simple, il a été démontré que le Random Forest peut être appliqué avec succès à des arbres limités à deux feuilles [13]. Pour le nombre de modèles agrégés, la convergence de performance est atteinte avec un nombre de modèles agrégés à p parmi n. Ceci sous-entend que le nombre d'arbres dans la forêt croît avec le nombre de variables.

Les principes de base de la Triangulation de Delaunay, qui consiste à définir un réseau interconnecté de points dans lequels aucune nouvelle connexion en ligne entre deux ne peut être créée sans perturber une ligne existante, sont présentés en annexe.

11. BREIMAN L. (2001) Random Forest, Machine Learning 45, 5-32
12. HASTIE T., TISHIRANI R., FRIEDMAN J.H (2009) The Elements of Statistical Learning, Springer Verlag, 2nd edition, section 15.3
13. BUHLMANN YU (2002)

1.3.3 Trois approches sophistiquées de classification automobile

Le véhiculier ou de manière plus large l'utilisation des véhicules pour capter de l'information est aujoud'hui au coeur des enjeux de segmentation. Ces dernières années, la multiplication de données externes avec la commercialisation de la base de données des immatriculations (SIV) attise l'intêret. Certains percoivent, par exemple, le véhiculier comme une alternative pour contourner l'interdiction d'une tarification directe selon le sexe de l'assuré. Comme nous allons le voir, les modèles existants s'appuyant sur la théorie présentée précédemment font souvent appellent à un degré d'intervention de l'expert plus ou moins prononcé.

A - E. OHLSSON

E. OHLSSON a développé en 2007 un modèle combinant les GLM et la théorie de la crédibilité. Nommé GLMC, ce modèle permet d'intégrer, dans un modèle GLM, une variable avec de nombreuses modalités, et ainsi d'inclure le modèle du véhicule comme facteur tarifaire. Une variante de ce modèle introduit d'autres caractéristiques véhicule dans le modèle et ainsi isole la part du risque qui s'expliquerait par des attributs communs avec d'autres véhicules (par exemple, la puissance ou le poids). Ce dernier point est extrêmement intéressant, car il permet de capter des informations sur le duo conducteur/véhicule non captées jusqu'à présent.

Cette méthodologie ne permet pas directement de construire un véhiculier. De plus, cette modélisation nécessite une quantité de données considérable pour obtenir des résultats intéressants. Plus contraignant encore, elle contraint à une forte implication de l'expert dans la définition des hypothèses et dans le choix *a priori* des variables véhicule. Pour ces raisons, le modèle GLMC et ses variantes ne sont pas retenus dans ce mémoire [14]. Cependant, la théorie de la crédibilité sera, par la suite, introduite dans notre étude et combinée à des méthodes de Machine Learning pour améliorer la performance du véhiculier.

14. Pour plus d'informations sur ce modèle, se référer aux publications d'OHLSSON - OHLSSON E (2008) Combining generalized linear models and credibility models in practice.Scandinavian Actuarial Journal, 2008, 4, 301-314

B - R. SIPULSKYTE

Les travaux "Development of a Motor Vehicle Classification Scheme for a New Zealand Based Insurance Company" réalisés en 2012 par R. SIPULSKYTE ont fait l'objet de la première publication impliquant les Machine Learning dans la construction d'un véhiculier. La démarche de l'étude est présentée figure 1.7.

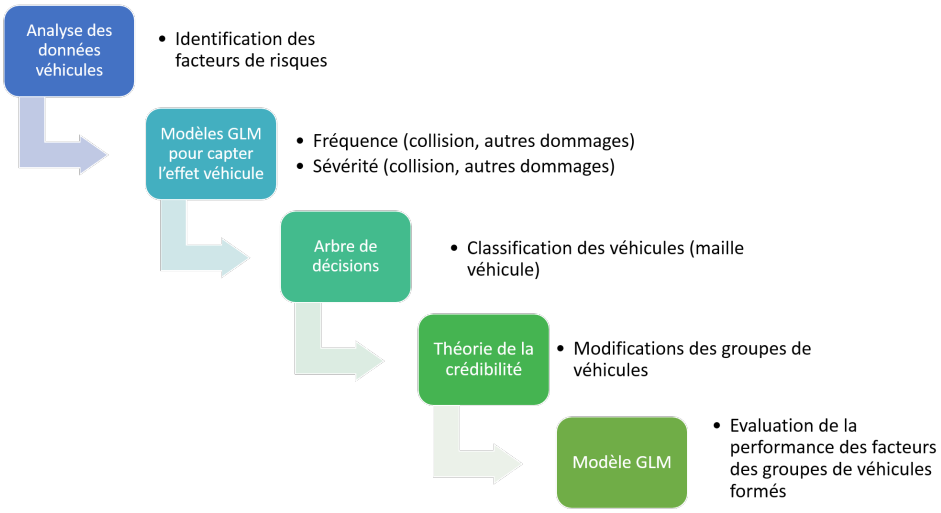


FIGURE 1.7 – Démarche du mémoire réalisé par R. SIPULSKYTE

Dans un premier temps, une analyse qualitative a été menée afin de détecter les facteurs les plus importants concernant la classification technique des véhicules et les caractéristiques du conducteur. Le deuxième temps consiste à isoler l'effet véhicule. Pour cela, les modèles GLM sont utilisés afin d'estimer la relativité de chaque facteur de risque (variables explicatives). L'effet véhicule est alors récupéré à la maille véhicule, puis il est utilisé en input d'un arbre de régression CART formant des groupes de véhicules. De par le choix de la maille, les véhicules ont été traités de la même manière indépendamment de leurs poids dans le portefeuille. La dernière étape consiste alors à fiabiliser les groupes formés par l'arbre. Pour cela, chaque groupe est subdivisé en sous-groupes auxquels est appliquée la théorie de la crédibilité. Selon leur coefficient de crédibilité, certains sous-groupes sont déplacés vers un autre groupe. L'opération est répétée. Et après deux itérations, les nouveaux groupes sont stables et considérés comme finaux.

Cette étude, par rapport à celle de E. OHLSSON, fait peu d'hypothèses en amont sur les variables véhicule et nécessite donc une moindre intervention de l'expert. Néanmoins, la manière dont la crédibilité est appliquée, requiert un arbitrage de l'expert dans le déplacement des groupes et leurs justifications.

Un détail de cette méthode et de ses limites est présenté en annexe

En dépit de ses limites, cette approche demeure intéressante par son utilisation d'un algorithme de Machine Learning et l'emploi de la crédibilité, c'est pourquoi dans le cadre de ce mémoire une méthode ajustée à l'appui de ces principes sera dévelopée au dernier chapitre.

C - AXA Global P&C

AXA Global P&C (IARD) a développé en 2012, une méthode permettant de capter une partie de variance résiduelle par la création d'une variable de classement des véhicules. Si la part de la variable d'intêret (ici la fréquence) liée au véhicule est captée par un modèle GLM, l'originalité de l'étude réside dans la création d'une carte des véhicules. L'objectif de cette carte est double ; capter la part de signal dans la variance résiduelle et fiabiliser l'information des véhicules par lissage spatial. La méthodologie développée est résumée figure 1.8.

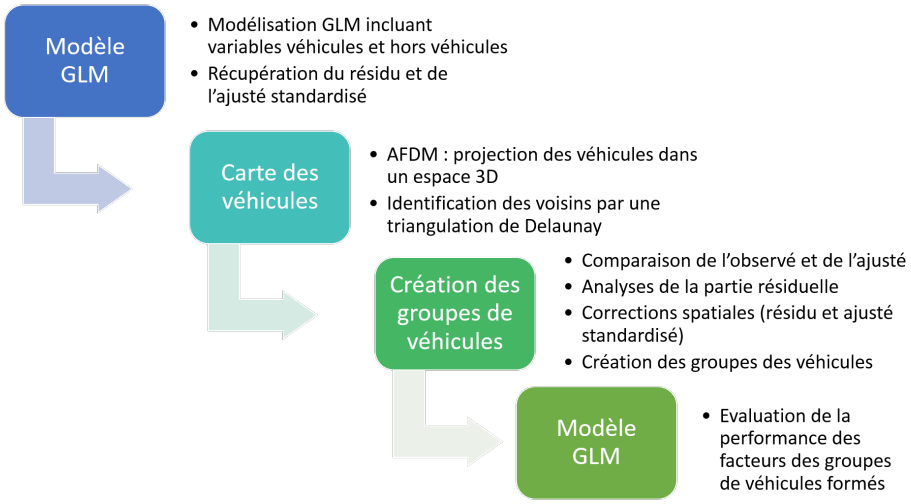


FIGURE 1.8 – Démarche de la méthodologie développée par Global P&C

La première étape consiste à construire en parallèle un modèle GLM et une carte des véhicules. D'un côté, le modèle GLM est construit de manière classique c'est-à-dire en incluant les variables véhicule et non véhicule. Le résidu du GLM est ensuite isolé et analysé afin de détecter la présence potentielle de variance non expliquée (signal). Si une présence de variance inexpliquée est détectée, la carte des véhicules sera utilisée pour capter ce signal. D'un autre côté, la carte des véhicules est construite en effectuant une analyse de réduction de dimension sur la base des données des caractéristiques véhicule. Les modèles de véhicules sont projetés selon leurs caractéristiques véhicule uniquement (en opposition aux caractéristiques techniques de sinistralité) dans un espace de dimension réduite. La carte est ensuite matérialisée par la création de connexions entre les véhicules grâce à une triangulation de Delaunay. Deux véhicules sont considérés comme voisins "s'ils sont reliés" directement l'un à l'autre. La triangulation de Delaunay ne considérant que la position géographique, forme un réseau fermé, excluant de fait la possibilité de véhicule isolé. Pour cette raison, les voisinages sont analysés et les liens aberrants supprimés.

La deuxième étape est l'analyse de la classification elle-même. Cette dernière suit un processus selon trois étapes : un lissage spatial du résidu, une correction spatiale de l'ajusté standardisé

et le partitionnement des véhicules. Le lissage spatial du résidu sur la carte des voisins cherche à isoler le signal du bruit. Dans le cas où le résidu s'éloigne de 1, le résidu n'est pas aléatoire et contient une part de signal. Il est alors lissé et noté résidu lissé. Ensuite, l'ajusté standardisé est corrigé spatialement selon le voisinage et l'exposition : noté ajusté standardisé lissé. Ainsi, l'ajusté standardisé des modèles de véhicules bien représentés (c'est-à-dire une exposition au risque élevé) est utilisé pour estimer les modèles de véhicules sous représentés (exposition au risque faible). Dans un troisième temps, pour chaque modèle de véhicules le résidu lissé et l'ajusté standardisé lissé sont additionnés formant l'effet véhicule lissé. Les modèles de véhicules sont ensuite partitionnés selon l'effet véhicule lissé créant ainsi la variable de classement des véhicules.

Un détail de cette méthode et de ses limites est également présenté en annexe

Tout comme la méthode mise au point par E. OHLSSON, la méthode d'Axa Global nécessite une forte intervention de l'expert dans la modélisation des variables véhicule par un modèle GLM. Par la suite, nous allons exploiter la richesse de l'approche pour une perspective proche du 100% machine.

1.4 Détail des deux approches exploitées

Sur la base des travaux présentés précédemment, deux approches ont été créées et ajustées pour éviter au maximum l'intervention de l'expert dans l'élaboration du véhiculier.

1.4.1 Principe général : deux approches en trois étapes

A - Une méthodologie en trois étapes

Ces deux approches ont un déroulement similaire en trois étapes résumé à la figure 1.9.

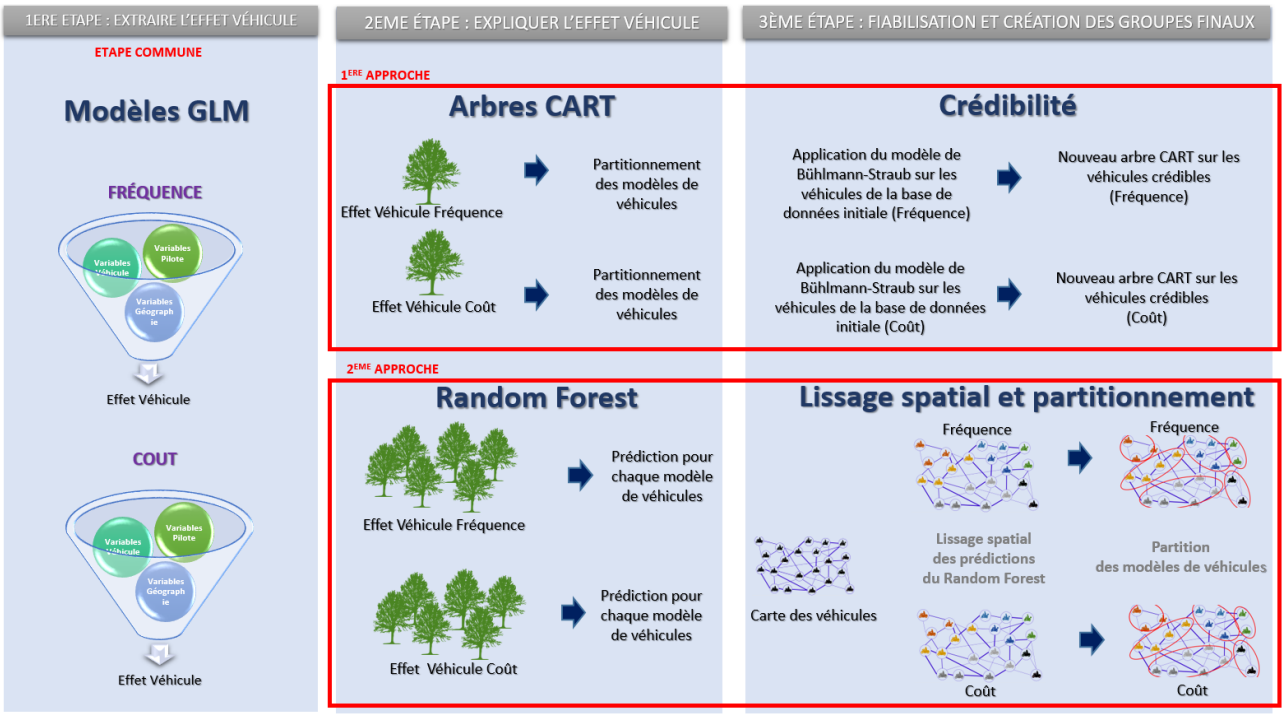


FIGURE 1.9 – Schéma récapitulatif de la méthodologie générale choisie pour l'étude

La première étape, commune aux deux approches, consiste à isoler par un modèle GLM l'effet véhicule dans le coût (respectivement la fréquence) observé(e). Puis, cet effet véhicule est expliqué dans une seconde étape par des méthodes de Machine Learning à l'aide des variables véhicule. La dernière étape cherche à fiabiliser les prédictions des modèles de l'étape précédente.

B - Une première étape commune

La première étape est déterminante pour la bonne extraction de l'effet véhicule. Nous avons pu évoquer le fait de l'utilisation de modèle GLM pour extraire l'effet véhicule. Mais la question qui se pose est : quelles variables choisir pour élaborer le GLM ?

Le modèle le plus intuitif consisterait à construire un GLM en excluant les variables véhicule. Or un tel modèle est potentiellement instable, dans la mesure où des variables explicatives importantes pourraient être oubliées. L'autre limite importante est que les coefficients estimés dans le modèle ne prennent pas en compte l'interaction/la corrélation entre les variables véhicule et non véhicule. Pour ces raisons, il a été choisi d'inclure certaines variables véhicule dans la construction du modèle GLM.

La seconde question qui se pose naturellement ensuite est la suivante : comment extraire l'effet véhicule compte tenu des interactions existantes entre les différentes variables ?

Une première méthode consiste à recalculer les coefficients du modèle GLM en excluant les variables non véhicule. Les variables véhicule ne feront ainsi plus partie de la modèlisation mais elles auront servis pour sélectionner les variables du modèle. Cette méthode sera nommée par la suite "méthode résiduelle".

R. SIPULSKYTE propose une alternative qui consiste à diviser chaque observation par le produit des coefficients GLM non véhicule. La valeur obtenue correspond au coût (respectivement la fréquence) à profil identique. Dans le cas du GLM, le profil choisi est le profil le plus courant dans le portefeuille. De cette manière, les effets non véhicule sont supprimés. Cette seconde méthode qui a l'avantage de ne pas nécessiter le recalcul d'un intercept, sera nommée par la suite "méthode relative".

Afin d'obtenir un critère objectif de choix entre ces deux méthodes d'extraction, elles vont être mises en place et comparées dans le chapitre suivant.

1.4.2 Spécificités de la première approche : arbre CART et crédibilité

Cette première approche, basée sur l'étude de R. SIPULSKYTE, suit un déroulement très proche de l'étude d'origine. Pour se faire, nous allons repartir des limites identifiées de cette méthode pour proposer des alternatives permettant de construire un véhiculier avec un minimum d'intervention d'avis d'experts.

La **première limite** de l'étude de R. SIPULSKYTE est le sur-apprentissage de classification construite. L'over-fitting peut principalement s'expliquer par deux points : le choix de la maille véhicule et la non-optimisation du critère d'arrêt de l'arbre CART.

Quelle maille est la plus pertinente pour expliquer l'effet véhicule ?

La maille la plus naturelle pour expliquer l'effet véhicule est la maille contrat. Cependant, il est possible qu'étant donné la forte volatilité des observations, notamment pour le modèle de fréquence où de nombreux contrats sont sans sinistre, la maille véhicule soit plus adaptée. C'est l'hypothèse faite par R. SIPULSKYTE dans son étude. Ces deux mailles seront testées et comparées.

Quel critère d'arrêt doit être privilégié pour limiter l'overfitting ?

Pour limiter le risque de sur-apprentissage, le niveau d'élaguage privilégié de chaque arbre sera celui obtenu par validation croisée.

La **seconde limite** du modèle développé par R. SIPULSKYTE, est l'absence de méthode pour classer de nouveaux véhicules. En effet, la manière dont la crédibilité est utilisée rend impossible l'utilisation du pouvoir prédictif de l'arbre CART qui perd alors son avantage premier. Ce mémoire propose alors une utilisation différente de la théorie de la crédibilité.

Comment crédibiliser l'information sans déplacement manuel des véhicules ?

Plutôt que de mettre en place une méthode de crédibilité sur des sous-groupes de chaque feuille de l'arbre CART, il est proposé dans ce mémoire de mettre en place la crédibilité en amont de l'arbre. L'idée développée est la suivante : le modèle de Bühlmann-Straub, rappelé en annexe, est appliqué aux véhicules puis les données en input de l'arbre CART sont sélectionnées selon le coefficient de crédibilité du véhicule assuré. Ainsi, le concept d'utilisation du facteur de crédibilité proposé par R.SIPULSKYTE est détourné afin de définir dans l'effet véhicule les données crédibles.

1.4.3 Spécificités de la seconde approche : Random Forest et lissage spatial

Cette seconde approche fait une utilisation détournée de la méthode d'Axa Global P&C.

Toujours dans le but de construire un véhiculier à dire de machines, nous n'avons pas envisagé la modélisation GLM telle que proposée par l'auteur de la méthode. L'approche sélectionnée ici est l'extraction de l'effet véhicule à l'identique de la première approche. L'effet véhicule est ensuite expliquée par une méthode de Machine Learning : le **Random Forest**. L'intêret de ce choix sera de challenger les prédictions de Random Forest avec celles de l'arbre CART.

L'**utilisation de la carte** est **différente** par rapport à la méthode initialement proposée. Si dans cette dernière, le lissage est utilisé en partie sur les résidus, dans la présente d'étude, l'effet véhicule, prédit par le Random Forest, va être lissé grâce à la carte des véhicules afin de crédibiliser l'information.

Quelle méthode de lissage utiliser ?

Il existe dans la littérature diverses méthodes de lissages. Nous avons souhaité utiliser un lissage tenant compte de l'exposition de chaque modèle de véhicules (c'est-à-dire le nombre d'années polices pour le modèle de fréquence et le nombre de sinistre pour le modèle de coût). Ainsi, pour les véhicules sous-représentés, l'effet véhicule prédit du Random Forest sera fiabilisé par celui des véhicules suffisamment représentés.

Le modèle de lissage spatial dans le cadre de cette étude est inspiré d'une méthode implémentée dans le prologiciel Classifier dite « Distance (Credibilité Weighted) ».

Définition 1 (*Formule de lissage spatial*)

Soient

- R_i *l'effet véhicule lissé du véhicule* i
- r_i *l'effet véhicule prédit par le Random Forest pour le véhicule* i
- V_i *le facteur de crédibilité défini de la manière suivante :* $V_i = \left(\frac{w_i}{w_i+w_0}\right)$
- wk *le nombre d'années d'exposition du véhicule* k *pour le modèle de fréquence, le nombre de sinistres pour le modèle de coût*
- d_{ik} *la distance euclidienne sur la carte des véhicules entre le véhicule* i *et son véhicule voisin* k
- P *paramètre de lissage*
- $\overline{r_i}$ *la moyenne des valeurs non lissées du voisinage du véhicule* i *pondérée par la distance euclidienne du véhicule* i *à voisins* j *selon la formule suivante :* $\overline{r_i} = \frac{\sum_{k\neq i} r_j d_{ik}^{-P} w_k}{\sum_{k\neq i} d_{ik}^{-P} w_k}$

$$R_i = V_i r_i + (1 - V_i)\overline{r_i}$$

L'une des principales limites du modèle initialement développé par Axa Global P&C est que le partitionnement des véhicules favorise l'homogénéité des classes vis-à-vis du nombre de véhicules par groupe au dépend de l'homogénéité des variables véhicule au sein de chaque groupe.

Quelle méthode de classification utiliser ?

Les techniques de classification sont nombreuses, parmi celles-ci les algorithmes de classification ascendante hiérarchique reposent sur la recherche itérative de rapprochement entre deux classes. Le cœur de cette famille réside dans la définition des distances entre deux classes. La méthode Ward définie la distance entre deux classes par la diminution d'inertie associée à leur fusion. Elle revient à créer un regroupement de classes de manière à maximiser la variance intra-classe. Cette dernière a été retenue pour la classification des véhicules. Il convient de noter que la définition de la distance tend à construire des classes sphériques et de mêmes effectifs.

Chapitre 2

De la constitution de la base de données à l'isolement de l'effet véhicule

2.1 Elaboration de la base de données de l'étude

La qualité de la base de données est un prérequis essentiel pour atteindre l'objectif. En effet, pour envisager challenger le véhiculier à dire d'experts, les machines nécessitent des données de qualité pour leur apprentissage. Avant de faire une analyse quant à la qualité des données utilisées pour cette étude, et notamment les retraitements réalisés pour gagner tant en exhaustivité qu'en pertinence, commençons par regarder la manière avec laquelle nous avons construit la base brute initiale.

2.1.1 Les différentes bases de données brutes

La sélection du périmètre de l'étude est une étape décisive. Le périmètre doit à la fois être défini de sorte que l'étude soit représentative du portefeuille du produit et doit limiter la présence de cas pathologiques qui pourraient atteindre l'intégrité des résultats de l'étude.

Le **périmètre final** défini pour l'étude est le suivant :

- Les années 2012 à 2014 [1] ;
- Le produit moto, à savoir les véhicules présentés figure 1.1, hors usage tout terrain ;
- Les personnes physiques, hors société ;
- Les contrats non temporaires, avec effet (exposition non nulle) et à tacite reconduction ;
- Le nouveau produit de tarification [2] ;
- Sinistres hors graves (seuil des graves : 150 000 €) [3].

1. Amplitude d'exposition imposée

2. Exclusion des anciens produits où un certain nombre de variables tarifaires sont manquantes. En effet le risque de biais non négligeable lors de la construction du modèle GLM a mené à l'exclusion de ces contrats.

3. Sinistres graves exclus sont des sinistres atypiques et non représentatifs, risquant de biaiser l'étude GLM du coût moyen.

Les bases de données brutes utilisées pour cette étude sont résumées figure 2.1. Un détail de ces bases est, par ailleurs, présenté en annexe.

Les bases internes brutes ont été extraites à l'aide du logiciel SAS. Cette étape a nécessité le développement de programmes adéquats et de contrôles de cohérences associées. Afin de récupérer l'ensemble des données assurés, deux bases conséquentes ont du être extraites : une base avec des modalités prédéfinies, une seconde renseignée manuellement et retraitée par la suite. Les bases externes brutes sont, quant à elles, des bases commercialisées. La base SRA est fournie par l'association alors que la base SIV est directement intégrée dans le système de données Axa nécessitant également le recours à SAS.

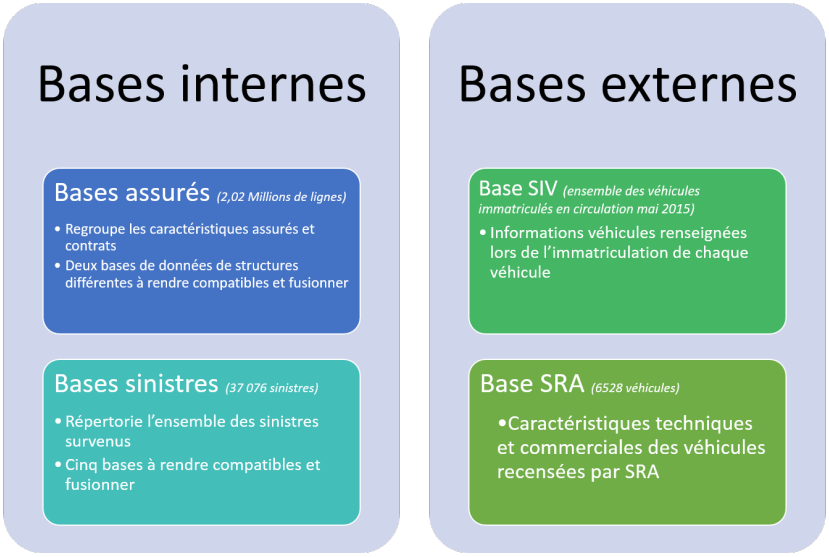


FIGURE 2.1 – Récapitulatif des bases de données brutes

Parce qu'il est important d'avoir le même recul et donc pour cela de définir un vieillissement, les sinistres sont vieillis au mois M+6 de l'année suivant la survenance, c'est-à-dire vus à cette date.

2.1.2 Le rapprochement des bases

Le rapprochement entre les différentes bases de données a nécessité de nombreuses étapes de fiabilisation qui ont été résumées dans le schéma figure 2.2. L'ordonnancement du rapprochement est le suivant : après avoir rapproché les bases internes entre elles, le lien avec la base externe SIV a été fait avant de rapprocher la base SRA. Les bases internes et SIV ont été rapprochées grâce à une variable commune aux deux bases : le numéro d'immatriculation. Ce rapprochement a été rendu possible grâce à la fiabilisation au préalable des informations SIV. En effet, comme nous le verrons par la suite, la base SIV a une qualité de données qui peut laisser à désirer.

L'étape la plus délicate a été le rapprochement entre les bases internes/SIV et la base SRA. Les principales difficultés rencontrées sont les suivantes :

- Une écriture différente des modèles de véhicules ;
- La multiplicité des véhicules SRA pour un même véhicule C14 ;
- Des catégorisations différentes de véhicules.

Plusieurs moyens ont alors été mis en place pour permettre ce rapprochement. Dans un premier temps, les erreurs d'écriture les plus courantes dans le nom des véhicules ont été analysées et corrigées. Puis, un processus itératif à trois étapes intégrant un ensemble de tests de cohérence a été mis en place. Lors de ce dernière, 40 clés ont été formées pour identifier le véhicule SRA correspondant. Lorsque cela a été nécessaire des hypothèses ont été faites. Des tests de concordance entre les valeurs rapprochées ont été réalisés tout au long de ce rapprochement [4].

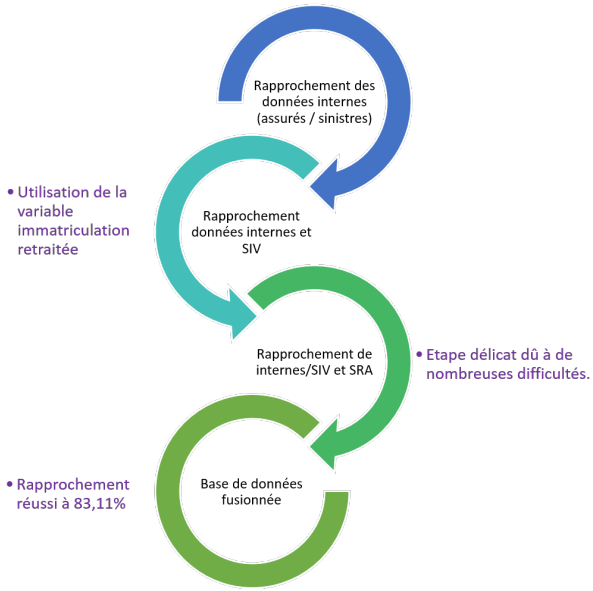


FIGURE 2.2 – Récapitulatif de l'étape de rapprochement des bases de données brutes

Le rapprochement entre les variables exogènes et la base assurés a été réussi à $83,11\%$, donnant une base finale de $1,69$ millions de lignes dont 34274 sinistres, suffisamment exhaustive pour mener à bien l'étude. Les $16,89\%$ de la base initiale pour lesquels l'information n'a pas pu être récupérée constitutent des véhicules non renseignés dans SRA ou bien ceux pour lesquels les informations dont nous disposons ne nous permettent pas de définir un identifiant SRA sans introduire un biais considéré trop important pour l'étude. Puisque ces dernières ne semblent pas, par ailleurs, représenter de profil particulier, nous pouvons faire l'hypothèse pour la suite que la base obtenue est représentative du portefeuille moto étudié.

4. Un détail des principales difficultés, des moyens mis en oeuvre et des hypothèses faites est présenté en annexe.

2.1.3 Fiabilisation et traitement de la base après rapprochement

Une fois les bases rapprochées, un ensemble de retraitements a dû être effectué (corrections, regroupements, exclusions) afin de fiabiliser les données de l'étude.

A - Traitement des variables internes

La base de données a été explorée afin de **remplacer ou supprimer les données incorrectes**. Les variables ayant trop de valeurs manquantes ou ayant une qualité très faible ont été retirées de l'étude et celles ayant des individus aberrants ont été remplacées ou exclues de la base de données finale.

Ensuite, afin d'assurer la robustesse des méthodes statistiques utilisées par la suite, notamment pour les GLM, les données multifactorielles ont été traitées afin de rendre chaque modalité représentative. Lorsque cela a été possible, les modalités représentant moins de 5% d'exposition ont été **regroupées** avec des modalités ayant une fréquence et un coût moyen similaire et sur l'avis d'experts.

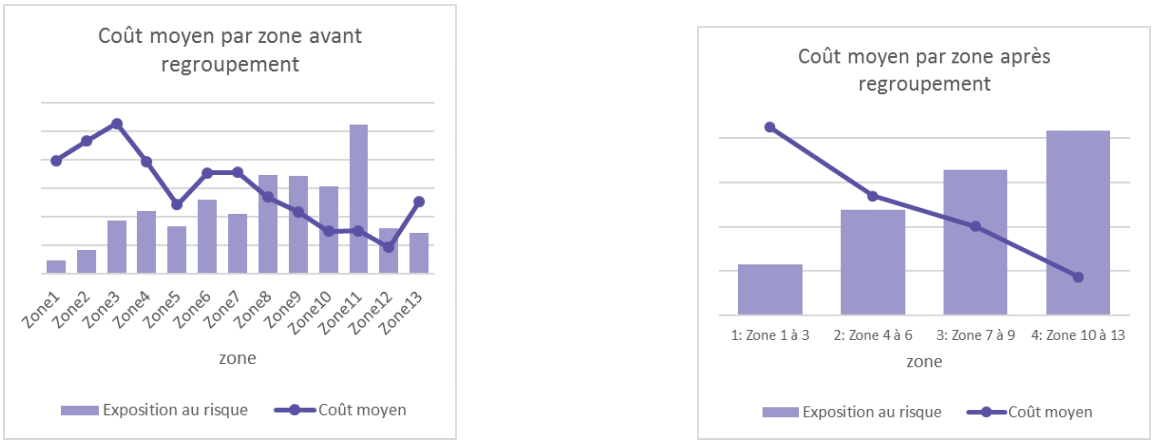


FIGURE 2.3 – Exemple de regroupement des modalités d'une variable

B - Traitement des variables externes

Concernant les **données SIV**, de nombreuses informations étaient **manquantes ou erronées**, si bien que la majorité des variables de cette **base ont été exclues** de l'étude par manque de fiabilité. Les variables sur l'identité du véhicule (notamment immatriculation, marque, modèle, CNIT, etc.) étaient suffisamment complètes pour les comparer avec les informations des autres bases. En cas d'ambiguïté sur la modalité d'une variable SIV pour un identifiant SRA donné, la modalité la plus courante pour ce même identifiant a été choisie. Au vu de la faible qualité des données SIV et de leur redondance avec des informations dans la base SRA, les seules

variables SIV conservées pour l'étude sont les variables vitesse du moteur et niveau sonore. Les variables concernant une maille plus fine que le modèle du véhicule, telles que sa couleur, ont été exclues de l'étude.

Ensuite, les **variables SRA**, mieux renseignées que celles SIV, ont des informations **redondantes ou très corrélées**. Par exemple, la *puissance réelle CEE* (kWatt) et la *puissance réelle DIN* (cH DIN) représentent la puissance réelle du véhicule selon une unité de puissance différente. Dans le cadre de cette étude, nous avons **choisi de garder** l'unité de cH DIN. Cependant, deux variables correspondent à la puissance du véhicule DIN : la *puissance réelle DIN* et la *puissance réelle DIN libre*. La différence provient du bridage des véhicules par les constructeurs afin de respecter la législation européenne. La puissance DIN libre, correspondant à la puissance non bridée, ne sera pas prise en compte dans notre étude puisqu'elle ne correspond pas à la puissance des véhicules en pratique.

La variable *ratio puissance-masse* de la base SIV étant rendue inutilisable faute de valeurs suffisamment renseignées, **une variable équivalente a été construite** en utilisant la *puissance* et le *poids* de la base SRA.

2.1.4 Base de données finale

Les figures 2.5 et 2.4 présentent les variables de la base de données finale. Il est à noter que seules les variables utilisables dans la modélisation sont présentées ici. Les variables utilisées pour le rapprochement, notamment concernant l'identification des variables (modèle, version, numéro CNIT, plaque d'immatriculation, etc.), ne sont pas détaillées.

Par rapport à la base initiale, le coût moyen de la base finale a diminué de moins de $0,96\%$ et la fréquence moyenne a augmenté de près de 10%. L'augmentation de la fréquence moyenne est, dans notre étude, un point positif puisque nous avons en moyenne plus de sinistres, ce qui est une information précieuse. Par la suite, nous parlerons de variables véhicule pour évoquer les variables externes et les variables internes de la section véhicule (figures 2.5 et 2.4), toutes les autres variables seront appelées variables non véhicule.

La base finale obtenue va être divisée en deux parties, les années 2013 et 2014 formeront notre base d'apprentissage sur laquelle l'étude sera faite et l'année 2012 servira de base test pour valider les résultats.

	Variables externes	
	Variables	Modalités (après traitement des données)
Source SRA	Puissance réelle	numérique (DIN)
	Identifiant SRA	3271 modèles
	Cylindrée	numérique
	Ratio puissance/poids	numérique
	Capacité réservoir essence	numérique
	Groupe SRA ($V_{Groupe\text{-}SRA}$)	4 à 19
	Classe SRA ($V_{Classe\text{-}SRA}$)	Z A B C D E F G H J *
	Freinage couple	Oui - Non
	ABS	Oui - Non
	Marque	BMW - HARLEY DAVIDSON - HONDA - SUZUKI - YAMAHA - Autres italiennes - Autres asiastiques - Autres
	Catégorie	Buggy (BUG) - Cross (CRO) - Custom (CUS) - Divers (DIV) - Enduro (END) - Grande Touriste (GRT) - Quad (QAD) - Roadster(ROA) - Routière Basique (RBA) - Routière sportive (RTS) - SSV (SSV) - Sportive (SPO) - Scooter 2 roues (SCO) - Scooter 3 roues (SC3) -Trail (TRA) - Trial (TRI) - Trike (TRK)
	Nombre de roues	2 roues - 3 roues - 4 roues
	Assistance freinage d'urgence	Oui - Non
	Nombre de cylindres	numérique
	Puissance fiscale	numérique
	Aide à la conduite	Oui - Non
	Poids à sec	numérique (en kg)
	Dernier prix connu	numérique (en €)
Source SIV	Vistesse moteur	numérique
	Niveau sonore	numérique

FIGURE 2.4 – Tableau récapitulatif des variables externes après traitement des données

	Variables internes	
	Variables	Modalités (après traitement des données)
Pilote	Âge du pilote	<25 ans [30 ; 35 ans[≥50 ans [25-30 ans[[40- 50 ans[
	Ancienneté du permis	< 3 ans [10 ; 15 ans[≥30 ans [3 ; 5 ans[[15 ; 20 ans[[5 ; 10 ans[[20 ; 30 ans[
	Conditions Sociaux Professionnelles (CSP)	Autres Fonctionnaire Salarié
	Situation matrimoniale	Autres Célibataire Marié\Concubin
	Coefficient Technique réel(CT)	< 0.7]0,8 ; 0.9[[1; 1,10[[0.7 ; 0.8[ ]0,9 ; 1[≥ 1,10
	Antécédent-Expérience	Avec-Avec Sans
	Groupe d'écart	0 à 1 groupe d'écart 2 à 4 groupes d'écart 5 groupes d'écart
	Usage du véhicule	Privé Trajet-travail Professionnel\Tournée
	Ancienneté du véhicule	<1 an [3; 4 ans[[6; 8 ans[≥15 ans [1; 2 ans[[4; 5 ans[[8; 10 ans[[2; 3 ans[[5;6 ans[[10 ; 15 ans[
	Durée de détention du véhicule	<1 an [2; 3 ans[[4; 5 ans[[6; 8 ans[[1; 2 ans[[3; 4 ans[[5; 6 ans[≥8 ans
Géographie	Zone	Zones 1 à 3 Zones 7 à 9 Zones 4 à 6 Zones 10 à 13
	Région du risque	Autres Sud-Est Ile-De-France Sud-Ouest Nord-Est
	Réseau de distribution du produit	Agent Courtier CCAS/Salarié
Contrat	Formule	F1: Responsabilité Civile (RC) F2: RC+Incendie Vol (IV) F3: RC + IV + Dommages Collisions (DC) F4: RC + IV + Tous dommages (TC)
	Fractionnement	Annuel Mensuel/Trimestriel/Semestriel
	Option Sécurité du conducteur étendue (sdc étendue)	Oui Non
	Option garantie Accessoires	Oui Non
	Coefficient Technique Personnalisé (CTP)	[65;70[[75;80[[70;75[≥80
Véhicule	Cylindrée du véhicule	≤125 cm3]500;750 cm3[ ]125;500 cm3] ≥750 ans
	Marque du véhicule	Autres Honda Autres asiatiques BMW Suzuki Autres italiennes Harley Davidson Yamaha
	Groupe C14 (V_{Groupe})	Groupes 30 , 31 et 05 Groupes 36 et 37 Groupes 32, 33 et 06 Groupes 38 et 39 Groupes 34 et 35
	Classe C14 (V_{Classe})	J/K/L O M P N Q/R
	Genre C14	CUS ROA SPO/HYP GTO ROU/BAS TRA/SMO QUAD/END/EXO/SID SCO/SC3

FIGURE 2.5 – Tableau récapitulatif des variables internes après traitement des données

2.1.5 Contrôle de la qualité des données

Cohérence des données

Une cohérence est assurée par une approche unique sur l'ensemble des données exploitées. La démarche utilisée pour la constitution de la base de données finale, précédemment détaillée, a en effet été appliquée de manière équivalente pour chaque période étudiée (2012 à 2014).

Exactitude et exhaustivité des données

La réalisation de l'objectif de cette étude est intimement liée à l'exactitude et l'exhaustivité des données, particulièrement les données liées aux véhicules.

Nous avons peu d'information sur la traçabilité et l'alimentation des données externes (SIV et SRA) qui sont pourtant utilisées sur la place. C'est grâce à des retraitements, des exclusions et des tests de cohérence que la base de données finale a pu être considérée comme fiable. En effet, les bases de données brutes utilisées n'avaient pas toute le même degré de fiabilité. Par exemple, comme évoqué précédemment, une partie de la base de données assurés avait un taux d'erreur de saisie manuelle relativement important et la base SIV avait une faible qualité de données.

Suite à des informations en doublon ou un manquant de fiabilité dans les données, seules 16 variables externes sur une cinquantaine ont été sélectionnées. Ce nombre de variables semble faible au regard des objectifs. Nous déplorons notamment l'absence d'information sur le *kilomètre départ arrêté* qui s'était révélé significatif lors d'études antérieures. Nous essayerons néanmoins de l'approximer à l'aide des différentes informations dont nous disposons sur la capacité de puissance du véhicule (vitesse du moteur, nombre de cylindres, cylindrée).

Enfin, l'historique de données imposé pour cette étude paraît lui aussi limité compte tenu de la faible sinistralité et de variabilité des modèles de véhicules. Il se justifie cependant par le souhait de créer un véhiculier en adéquation avec le portefeuille récent, et également par la nécessité d'utiliser la base SIV disponible uniquement pour le mois de mai 2015.

Maintenant que la base de données finale est constituée, nous allons nous intéresser à ce que les modèles de Machine Learning sont capables d'expliquer à partir de ces données et notamment de l'effet véhicule sous-jacent. Pour isoler ce dernier par un modèle GLM, une étape d'analyse de données doit au préalable être effectuée.

2.2 Etude et pré-sélection des variables candidates pour le modèle GLM par une analyse univariée et multivariée

Nous présentions au premier chapitre la décision d'utiliser à la fois **les variables véhicule et non véhicule** pour le **modèle GLM** dans l'objectif de capter la corrélation entre conducteur et véhicule. Dès lors que les données ont été introduites, nous pouvons préciser quelles variables sont concernées.

L'emploi de toutes les variables véhicule nécessiterait de faire des hypothèses fortes *a priori*. De plus, l'utilisation des variables externes comme variables candidates exigerait l'intervention d'avis d'experts pour les regroupements. Nous avons, par conséquent, choisi d'utiliser uniquement les **variables internes** (véhicule et non véhicule) dans le modèle GLM.

Une analyse univariée et multivariée de ces dernières va permettre de nous forger un avis sur les variables les plus pertinentes dans nos modèles GLM. En préambule de ces analyses, l'indépendance entre la fréquence et le coût, qui est un pré-requis pour la modélisation GLM telle que définie, va être justifiée.

2.2.1 Justification de l'indépendance fréquence et coût

Les contraintes de l'étude imposent implicitement une modélisation en séparant la fréquence et le coût moyen. Avant de mettre en place une telle modélisation, il convient de vérifier l'indépendance entre ces deux variables.

A - Coefficient de corrélation

Afin de quantifier le degré de dépendance entre la charge et la fréquence des sinistres, trois mesures de corrélations sont utilisées :

τ de Pearson	ρ de Spearman	τ de Kendall
0,04	0,12	0,08

FIGURE 2.6 – Tableau de corrélation entre la fréquence et le coût moyen

La mesure de Pearson est un indicateur de la corrélation linéaire entre deux variables. Une faible corrélation linéaire entre la fréquence et la charge est constatée. La mesure de Spearman donne une indication sur la corrélation entre les rangs des différentes valeurs de nos variables. Le ρ de Spearman calculé reflète une faible corrélation monotone entre les données. Pour finir, la mesure de Kendall compare la probabilité de concordance ou discordance de deux couples indépendants. Ce τ est estimé à 0,08 avec une p-value de $2,22e-16 < 1\%$. **Ces différentes mesures indiquent que la fréquence et le coût moyen ne sont pas corrélés**.

B - Analyse graphique

Afin de conforter les premiers résultats d'indépendance, une analyse graphique du diagramme de dispersion et des graphes de rangs est réalisée.

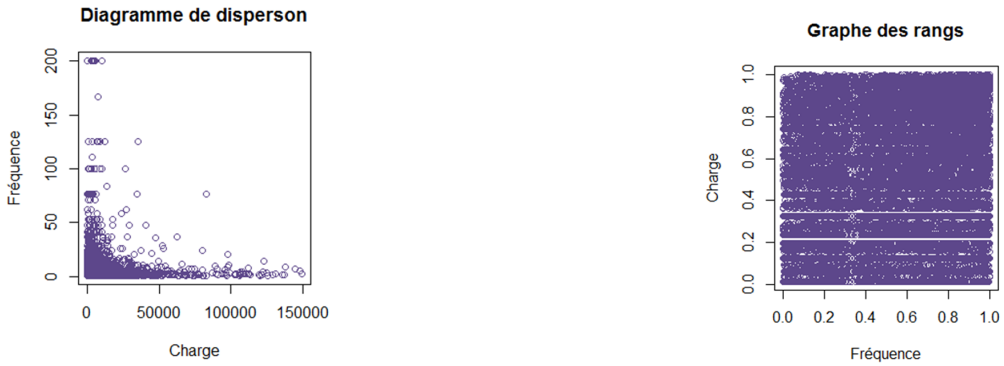


FIGURE 2.7 – Analyse graphique de corrélation entre fréquence et coût moyen

D'une part, le graphique 2.7 de gauche représente la fréquence observée en fonction du coût moyen observé. Nous n'observons pas de couple charge élevée et fréquence élevée. Par ailleurs, le graphe des rangs à droite indique que pour un rang donné de fréquence, le rang de charge semble prendre une valeur aléatoire et vice versa. **L'analyse graphique conclut donc à l'indépendance entre la fréquence et le coût.**

C - Copule empirique

Pour confirmer l'hypothèse d'indépendance, nous allons comparer la copule empirique et la copule d'indépendance.

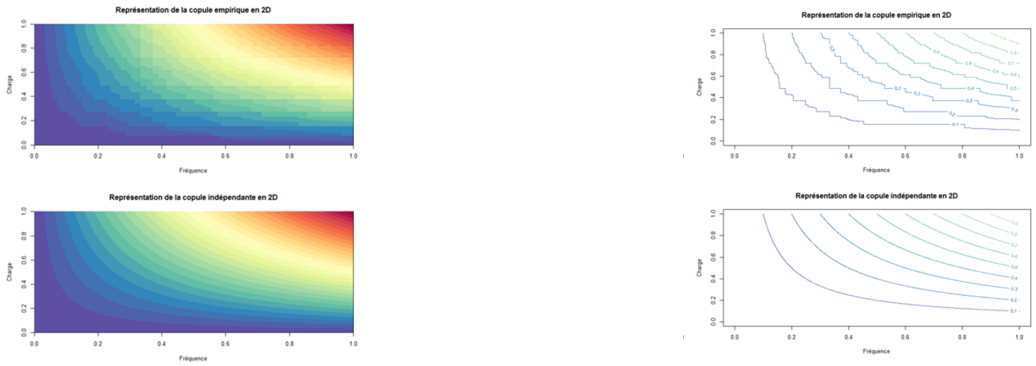


FIGURE 2.8 – Comparaison de la copule empirique et de la copule indépendante

La copule empirique a bien la même forme que la copule indépendante (cf figure 2.8). **En définitive, pour la suite de l'étude, nous acceptons l'hypothèse d'indépendance entre la fréquence et le coût.**

2.2.2 Statistiques descriptives

Avant d'examiner le comportement des variables internes vis-à-vis des variables d'intêret, quelques informations sur le profil de risque du portefeuille étudié sont présentées ci-après.

Les assurés de produit moto étudié ont une moyenne d'âge assez élevée. Plus de 30% ont entre quarante et cinquante ans, 34% ont plus de cinquante ans, et plus de 75% vivent en couple. Le véhicule est utilisé par plus de la moitié des assurés comme un moyen de transport pour aller travailler. Le parc des véhicules assurés est relativement "âgé", plus d'un tiers des véhicules ont plus de 10 ans. Nous constatons cependant que seul un sixième des véhicules est détenu depuis plus de huit ans. Le type de véhicule le plus représenté est le scooter deux roues de $125cm^3$ de la marque Yamaha. Les tendances des tris à plat sont résumées dans le tableau 2.9.

	Tendances tris à plats		V de cramer		Analyse de la variance (ANOVA)		Remarque
Variables	**Fréquence**	**Coût**	**Fréquence**	**Coût**	**Fréquence**	**Coût moyen**	
Âge du pilote	-	+/-	3,09%	16,20%	10	rejetée	Le coût est décroissant jusqu'à 40 ans et croissant au-delà
Ancienneté du permis	-	+/-	3,02%	16,30%	3	12	
Ancienneté du véhicule	-	+/-	3,99%	17,10%	2	4	
Antécédent\Expérience	Sans	Sans	6,24%	18,20%	9	11	Les assurés avec antécédent d'assurance et de l'expérience en moto ont moins d'accident et des accidents de graves que ceux ne respectant pas l'une de ces conditions.
Classe C14 (V_{Classe})	+	+/-	2,88%	19,80%	11	3	
CSP	Salarié et Autres	Fonctionnaire	2,38%	16,70%	20	1% < p-value < 5%	
CT	+/-	+/-	2,71%	16,70%	17	10	
CTP	+	-	3,09%	15,90%	14	rejetée	
Cylindrée du véhicule	Ø	+	2,47%	21,10%	19	7	
Durée de détention du véhicule	-	Ø	3,39%	16,20%	5	rejetée	
Formule	+	+	3,79%	22,10%	6	1% < p-value < 5%	
Fractionnement	Mensuel/Trimestriel/Semestriel	Mensuel/Trimestriel/Semestriel	3,37%	16,00%	15	rejetée	
Genre C14	scooters et GTO	custom et sportives	3,31%	19,30%	4	9	
Groupe C14 (V_{Groupe})	32 et 35	+	3,00%	20,60%	12	8	allure non croissante pour la fréqence
Marque du véhicule	autres italiennes et BMW	Harley Davidson	2,73%	17,00%	13	1	
Option SDC étendue	Ø	Oui	2,35%	16,60%	rejetée	2	
Option garantie accessoires	Ø	Oui					
Région du risque	Ile de France	Nord-Est	3,19%	16,80%	8	rejetée	
Réseau de distribution	Courtier	Agent	3,29%	17,20%	16	rejetée	
Situation matrimoniale	Célibataire	marié\concubin	2,85%	17,60%	18	1% < p-value < 5%	
Usage du véhicule	Professionnel tournee	Privée	4,51%	16,40%	7	5	
Zone	+	-	5,47%	17,40%	1	6	La fréquence de sinistres est plus élevée dans les zones urbanisées et inversement pour le coût des sinistres
Nombre de groupe d'écart	2 à 4 groupes d'écart	2 à 4 groupes d'écart					

	Tendances tris à plats	V de cramer	Analyse de la variance
Légende	+: tendance croissante - : tendance décroissante +/- : tendance non monotone nom: modalité ayant la fréquence (coût moyen) le plus elevé Ø : pas de tendance visible	V de cramer en %	1 à X : p-values inférieures à 5% par ordre décroissant 1% < p-value < 5% rejetée : p-value supérieure à 5%

FIGURE 2.9 – Tableau récapitulatif de l'analyse de données

Nous notons notamment que le coût moyen est décroissant avec l'âge pour les assurés de moins de 40 ans et croît à partir de 40 ans. De plus, les assurés avec antécédent d'assurance et de l'expérience en termes de conduite de moto ont moins d'accidents et en particulier moins d'accidents graves que ceux ne respectant pas l'une de ces conditions. Autre point notable, la fréquence de sinistres est plus élevée dans les zones urbanisées et inversement pour le coût des sinistres.

Ces statistiques descriptives nous donne une première idée du comportement des variables internes vis-à-vis des variables d'intérêt. Cependant, ils ne permettent pas de conclure à l'effet direct d'une variable sur la fréquence ou le coût moyen. De nombreuses interactions et effets cachés influencent l'évolution. Nous continuons alors notre analyse par une étude des corrélations.

2.2.3 Corrélations

L'étude des corrélations a ici un double objectif : identifier les variables explicatives corrélées entre elles pour éviter les problèmes de multicolinéarité dans les GLM, mais aussi pour évaluer la corrélation entre variables explicatives et à expliquer, à savoir fréquence et coût.

Le V de Cramer est préféré au test du χ^2 couramment utilisé pour tester l'indépendance entre les variables. En effet, ce dernier, sensible au nombre de modalités ne permet pas de comparer les mesures d'associations entre elles. Plus les variables sont correlées entre elles plus le V de Cramer sera proche de 100%.

Dans le cadre d'un modèle GLM, si deux variables sont trop fortement corrélées, nous pouvons choisir d'en exclure une ou bien de croiser les variables fortement corrélées entre elles. Un résumé des variables fortement corrélées est proposé figure 2.10.

Variable 1	Variable 2	V de cramer
Groupe écart	Antécédent/Expérience	98,9%
Formule	Accessoire	70,4%
Groupe	Classe	61,7%
Genre	Groupe	58,3%
CTP	Réseau	57,3%
Cylindrée	Classe	57,1%
Réseau	Région	56,5%
Cylindrée	Genre	56,5%
Cylindrée	Groupe	56,1%

FIGURE 2.10 – Tableau récapitulatif des variables explicatives fortement corrélées

Le groupe d'écart est très fortement corrélé avec l'antécédent et l'expérience de l'assuré. La variable antécédent/expérience, étant construite par rapport au nombre de groupes d'écart, nous choisissons d'**exclure la variable groupe d'écart**. Le **niveau d'accessoires garanti** est, quant à lui, fortement lié à la formule choisie. Le niveau de garanties des accessoires choisi étant majoritairement celui inclu de base dans la garantie, cette variable **n'est pas retenu pour le GLM**. Pour les autres variables, nous choisirons par la suite celles à éliminer en fonction de leur caractère discriminant.

La corrélation entre la variable à expliquer (fréquence respectivement coût moyen) et les variables explicatives (les variables internes), est par ailleurs résumée dans la figure 2.9. Il ressort que les variables **Antécédent-Expérience**, la **zone géographique**, l'**usage**, l'**ancienneté du véhicule** et la **formule** semblent être les variables les plus corrélées à la variable d'intêret **pour le modèle de fréquence**. D'autre part, la **formule**, **les variables liées au véhicule** (cylindrée, classe, genre et groupe), et l'**ancienneté du véhicule** semblent être les variables les plus corrélées à la variable d'intérêt pour le **modèle de coût**.

2.2.4 ACM et test ANOVA

Pour mettre en avant les facteurs discriminants, il est indispensable de passer par une analyse simultanée des variables. Pour cela, nous avons souhaité réaliser des ACM et des test ANOVA.

A - ACM

L'ACM avait initialement été mise en place pour identifier les variables les plus significatives pour nos deux modèles. Cependant, nous n'avons pu tirer aucune conclusion probante à cause de la forte proximité à des variables d'intérêt avec l'axe.

Cette analyse nous a cependant permis de mettre en évidence la corrélation entre conducteur et véhicule et dès lors de justifier le choix d'inclusion des variables véhicule dans le modèle GLM.

À l'aide du critère de JP. Benzécri [5], deux axes sont conservés [6]. Les variables ayant au moins deux modalités dont l'inertie par axe est supérieure à l'inertie moyenne, sont les variables significatives sélectionnées pour l'interprétation des résultats.

Variables	Axe 1	Axe 2
Cylindrée	+++	++
Classe	+++	++
Genre	+++	++
Groupe	+++	
Marque	+++	++
Zone	++	
Ancienneté permis	++	+++
Région	+	
Réseau	+	
CTP	+	
Ancienneté du véhicule		+++
Formule		+
Age	+	+++
Durée de détention du véhicule		+
Antécédent Expérience	+	++
Groupe d'écart	+	
Fractionnement		+
CSP		+

FIGURE 2.11 – Signification des axes pour l'ACM de la fréquence

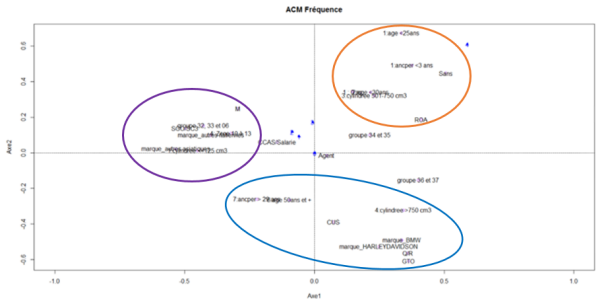


FIGURE 2.12 – ACM de la fréquence

Pour le modèle fréquence, le premier axe oppose les véhicules puissants de type sportif avec un groupe et une classe élevée aux véhicules de type scooter à deux ou trois roues de moins de $125cm^3$ de marques italiennes utilisés principalement en zone urbaine. Le second axe sépare les assurés jeunes avec peu d'expérience aux assurés de plus de 30 ans avec de l'expérience. Cet axe permet de constater que les moins de 30 ans roulent avec des motos de cylindrée inférieure à $750cm^3$ de type routière ou basique alors que les plus de 30 ans utilisent plutôt des engins de type customs dont la cylindrée dépasse les $750cm^3$ et met donc en avant la corrélation entre le conducteur et son véhicule.

5. Ce critère consiste à recalculer l'inertie des axes ayant une inertie supérieure à la moyenne.
6. 2 axes correspondent à pourcentage cumulé de l'inertie des axes de plus de 60% pour le modèle de coût et de 72% pour le modèle de fréquence.

B - Test ANOVA

Une analyse de la variance a été ensuite effectuée afin de conforter les observations de l'analyse uni-variée. Pour cela, des tests de significativité d'ANOVA ont été réalisés sous SAS avec la fonction PROC ANOVA. Les résultats de ces tests sont résumés dans le tableau 2.9 précédent.

Ce test permet de confirmer que les variables identifiées précedemment sont statistiquement discriminantes aussi bien pour le modèle de fréquence que de coût.

Pour le modèle de fréquence, en plus des variables déjà distinguées à savoir : l'Antécédent-Expérience, la zone géographique, l'usage, l'ancienneté du véhicule et la formule, nous remarquons que l'ancienneté de permis, le genre et l'usage apparaissent également comme discriminants.

Pour le modèle de coût, en plus de la formule, des variables véhicule et de l'ancienneté du véhicule, nous remarquons que l'option SDC [7] semble fortement discriminante.

En définitive, cette analyse de données, nous conduit à émettre un avis critique sur la sélection automatique des variables qui sera développée par la suite tant sur le plan de pertinence des variables que sur leurs corrélations. Cette analyse nous permettra ainsi de sélectionner au plus juste les variables à conserver dans nos modèles GLM.

7. SDC : Sécurité du conducteur

2.3 Isolement de l'effet véhicule : le GLM, première étape des approches envisagées

Une fois les études préliminaires réalisées, nous allons procéder au paramètrage puis à la mise en oeuvre des modèles GLM pour en extraire l'effet véhicule.

2.3.1 Variables explicatives : valeurs atypiques

Dans un premier temps, les variables d'intérêt sont analysées afin de détecter d'éventuelles anomalies dans leur distribution et plus particulièrement des valeurs extrêmes ou atypiques.

Le coût moyen

D'une part, nous constatons que certains sinistres ont une charge négative. Ce phénomène s'explique par la convention IDA/IRSA [8] mise en place entre les assureurs. Ainsi, par la suite, seules les charges de sinistres positives sont considérées. D'autre part, les variables extrêmes ont été traitées en amont en éliminant du périmètre d'étude les sinistres graves selon le seuil d'écrêtement défini par AXA. Dans cette étude, nous n'avons pas cherché à réévaluer ce seuil.

La fréquence

La fréquence peut, par définition, être élevée lorsque les sinistres sont survenus sur des images de faible durée. Par exemple, un sinistre survenu le jour suivant la création de l'image aura une fréquence de 365. L'analyse de la répartition de la fréquence selon l'exposition journalière montre la sur-représentation des images de faible exposition. Ce phénomène peut s'expliquer dans la pratique par des rectifications ou des mises à jour des informations assurés quelques jours après une nouvelle souscription. Après avoir vérifié qu'elles ne correspondaient pas à un profil de véhicule particulier, les images dont l'exposition est inférieure à 8 jours ont été exclues de l'étude du modèle de fréquence. Les images exclues représentent moins de 0,8% des sinistres.

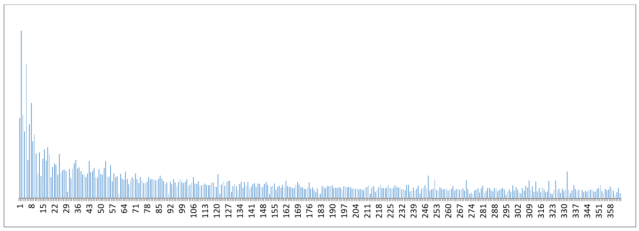


FIGURE 2.13 – Répartition de la fréquence de sinistres selon l'exposition en jour

8. Afin de simplifier et traiter plus rapidement les sinistres, différents assureurs automobiles/motos ont mis en place les conventions IRSA (Indemnisation Règlement des Sinistres Automobiles) – IDA (Indemnisation Directe de l'Assuré). Chaque assureur indemise alors son assuré puis l'assureur du non responsable exerce un droit de recours.

2.3.2 Elaboration des modèles GLM

A - Choix de la loi de probabilité et choix de la fonction lien

Loi de probabilité pour le coût

Les lois de probabilité testées sont les lois gamma et log-normale utilisées en assurance pour modéliser le coût.

L'application d'un modèle GLM repose sur l'appartenance de la variable à expliquer à la famille exponentielle ; or la loi log-normale ne fait pas partie de la famille exponentielle. Dans la pratique, cette contrainte est contournée en modélisant le logarithme du coût qui suit alors une loi normale.

Cependant, le recours au modèle log-normal oblige à une double transformation logarithme exponentielle qui transforme les écarts entre les valeurs observées et entraîne *in fine* une sous-estimation du coût. En effet, il est possible de montrer que l'espérance du coût estimée par un modèle log-normal est inférieure à l'espérance du coût.

Aussi, une solution envisageable serait de ne pas revenir au coût et donc d'extraire un effet véhicule correspondant au log du coût pour éviter cette double transformation. Cette modélisation du logarithme du coût entrainerait toutefois une distorsion de la structure des écarts ; or la distance entre nos points est déterminante pour les méthodes utilisées par la suite (CART comme Random Forest). De ce point de vue, la loi Gamma semble celle à privilégier.

Nous avons donc testé les deux lois sur notre jeu de données. Comme l'indique la figure 2.14 la comparaison de la déviance standardisée et du χ^2 de Pearson standardisé laisse à penser une meilleure adéquation du modèle Gamma.

	Déviance standardisée/ddl	X² de Pearson/ddl
Modèle log normal	2,09	2,09
Modèle gamma	1,08	1,64

FIGURE 2.14 – Comparaison des modèles Gamma et Log-normal

Ainsi bien que les graphes QQ-plots ont mis en évidence que la loi gamma possède une queue de répartition moins épaisse que nos données [9], **la loi Gamma est celle retenue pour notre étude**.

9. Des graphiques et remarques complémentaires sont présentés en annexe.

Loi de probabilité pour le nombre de sinistres

Lorsque la variable à expliquer dépend linéairement d'une autre variable, cette dernière est mise en offset. Dans le modèle de fréquence, la variable modélisée est le nombre de sinistres. L'exposition (ou nombre d'années polices) est alors mise en variable offset.

Les lois de probabilité couramment utilisées pour modéliser le nombre de sinistres sont la loi de Poisson et la loi Binomiale négative. Dans la mesure où le paramètre de la loi de Poisson est identique pour tous les assurés, une modélisation par cette loi suppose que la population est homogène face au risque c'est-à-dire que la fréquence individuelle des assurés est déterministe. Cette hypothèse ne paraît pas fondée *a priori*. Dès lors, une généralisation de la loi de Poisson envisageable est le modèle Binomial négatif (ou Poisson-Gamma). Une nouvelle hypothèse moins restrictive a alors été considérée : la fréquence individuelle est une variable aléatoire de loi Gamma.

Le test du χ^2 a été utilisé pour tester l'adéquation de la fréquence à ces deux lois (cf figure 2.15).

Au vu des résultats de ce test, **la loi Binomiale négative est celle retenue dans la suite de l'étude**.

Loi	Conclusion
Loi de Poisson	Rejet au seuil 5%
Loi Binomiale négative	Acceptation au seuil 1%

FIGURE 2.15 – Comparaison des lois Poisson et Binomiale négative

Loi de la fonction lien

Chaque loi de la famille exponentielle est associée à une fonction de lien canonique. L'utilisation de la fonction de lien canonique dans le GLM fournit des propriétés statistiques intéressantes, dont des simplifications dans l'estimation des paramètres du modèle. Cependant, il est tout à fait possible de choisir une autre fonction pour la fonction lien.

Dans notre étude, la fonction de lien a été choisie selon la structure du modèle souhaité. Les modèles additifs ou multiplicatifs permettent de récupérer simplement le résidu. Le modèle additif modélise le risque par une somme d'effets explicatifs alors que le modèle multiplicatif modèlise le risque par un produit d'effets explicatifs. Dans le cadre de cette étude, le modèle multiplicatif a été privilégié car, usuellement utilisé en IARD, il a l'avantage de garantir des prédictions positives. L'utilisation de la fonction log permet par inversion d'obtenir des effets multiplicatifs par propriété de la fonction exponentielle. Ainsi, **la fonction logarithme népérien est choisie comme fonction lien**.

B - Choix des variables explicatives

Nous avons étudié et défini à la section 2.2 les variables candidates aux modèles GLM. A présent, des méthodes de sélection automatique vont être mises en place et comparées aux résultats de l'analyse uni-varié et multi-varié afin de choisir les variables explicatives à retenir. La méthode *FORWARD*[10] va être utilisée selon deux critères : le critère AIC et le critère BIC[11].

Sélection des variables les plus pertinentes pour le coût moyen

La méthode Forward avec le critère BIC introduit les variables dans l'ordre suivant : la cylindrée, l'option SDC étendue, l'usage et le genre. Nous constatons que l'option SDC étendue ressort parmi les variables ayant le plus grand pouvoir explicatif. L'option SDC étendue est fortement corrélée aux sinistres corporels qui ont souvent des coûts élevés. Afin de distinguer les sinistres corporels des autres sinistres, une indicatrice de sinistres corporels a été introduite dans le modèle.

Lors d'un accident touchant la garantie RC, le montant du dommage est lié au véhicule de l'assuré, mais également à l'état de santé dans le cas d'une RC corporelle et au véhicule de l'autre conducteur dans le cas d'une RC dommage. Pour cette raison, une variable indicatrice de RC a été créée.

Après l'introduction de ces deux variables supplémentaires, une sélection automatique avec le critère BIC a été réalisée et comparée aux résultats de la section 2.2. Nous avons alors pu choisir les variables à conserver parmi celles fortement corrélées entre elles. Les choix suivants ont été faits :

- Conservation de la variable réseau et exclusion de la région et du CTP ;
- Conservation de la V_{Classe} et du genre et exclusion du V_{Groupe} et de la cylindrée.

Enfin, après avoir éliminé les variables corrélées, de nouvelles sélections automatiques respectivement avec les critères AIC et BIC, ont été mises en place. Celles-ci sont résumées figure 2.16.

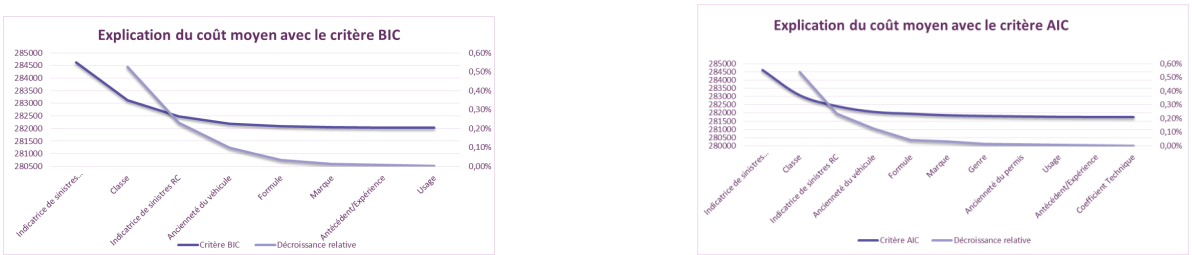


FIGURE 2.16 – Résultat du *FORWARD* selon les critères BIC (à gauche) et AIC (à droite) pour le modèle de coût

10. La sélection pas-à-pas ascendante (« FORWARD ») est une sélection ne contenant initialement aucune variable explicative. Elle consiste à intégrer une à une les variables contribuant le plus au modèle compte tenu des variables sélectionnées. L'algorithme s'arrête lorsque l'ajout d'une variable supplémentaire n'améliore plus le modèle selon un critère préfini.

11. Le critère d'Akaïke (AIC) et le critère de Schwartz souvent noté BIC (Bayesian information criterion).

Il n'existe pas de règles pour déterminer à l'avance le nombre optimal de variables. Il convient d'arbitrer entre un modèle avec un nombre important de variables qui s'ajuste parfaitement au modèle mais qui se généralisera mal et un modèle trop simple qui sera insatisfaisant du point de vue prédictif.

A partir des résultats des sélections automatiques résumées aux figures 2.16 et des résultats de l'analyse des données section 2.2, les variables suivantes ont été sélectionnées pour le modèle de coût :

- Indicatrice de sinistres corporels ;
- Classe ;
- Indicatrice de sinistres RC ;
- Ancienneté du véhicule ;
- Formule ;
- Marque.

Sélection des variables les plus pertinentes pour la fréquence

Selon le même principe que le modèle de coût, des sélections automatiques de variables ont été mises en place (cf figure).

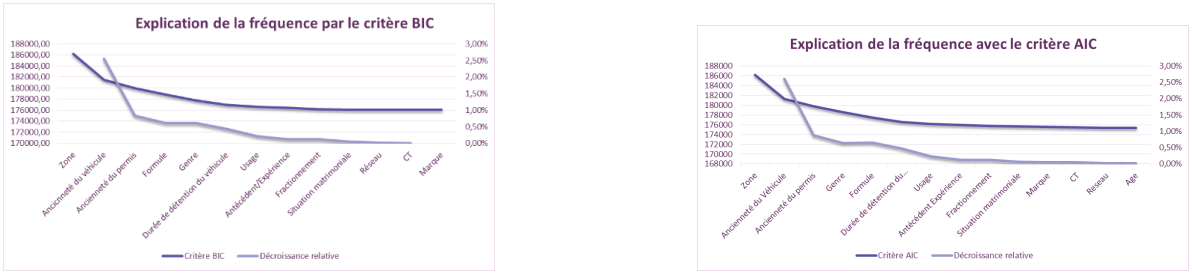


FIGURE 2.17 – Résultat du *FORWARD* selon les critères BIC (à gauche) et AIC (à droite) pour le modèle de fréquence

Les variables du modèle de fréquence ont été sélectionnées à partir des résultats des analyses de données préliminaires et des sélections des variables automatiques. Les variables sélectionnées pour le modèle de fréquence sont les suivantes :

- Zone ;
- Ancienneté du véhicule ;
- Ancienneté du permis ;
- Formule ;
- Genre ;
- Durée de détention du véhicule ;
- Usage ;
- Antécédent/Expérience.

C - Les modèles GLM

Avant de valider les modèles choisis, nous allons résumer les modèles GLM construits.

Le modèle de coût peut, de manière simplifiée, être écrit de la façon suivante :

Soit le Y la charges de sinistres

$$\begin{aligned} log(E(Y|x_1,..,x_{33})) =& \beta_0 + \beta_1 1_{\left\{1_{\text{Corporels}=1}\right\}} + \ldots + \beta_3 1_{\{V_{Classe}=\text{'J/K/L'}\}} + \ldots \\ &+ \beta_9 1_{\{1_{\text{RC}}=1\}} + \ldots + \beta_{11} 1_{\{\text{Formule}=F1\}} + \ldots \\ &+ \beta_{15} 1_{\{\text{Marque}=Honda\}} + \ldots + \beta_{23} 1_{\{\text{Ancienneté du véhicule='}< 1 \text{ an'}\}} + \ldots \\ &+ \beta_{33} 1_{\{\text{Ancienneté du véhicule='} \leq 15 \text{ ans'}\}} \end{aligned}$$

où Y suit une loi Gamma

Le modèle de fréquence peut, de manière simplifiée, être écrit de la façon suivante :

Soit le Z le nombre de sinistres.

$$\begin{aligned} log(E(Z|x_1,..,x_{37})) =& \beta_0 + \beta_1 1_{\{1_{\text{Zone}=\text{"}Zone13\text{"}}\}} + \ldots + \beta_4 1_{\{1_{\text{Ancienneté du véhicule}=\text{"}<1ans\text{"}}\}} + \ldots \\ &+ \beta_{13} 1_{\left\{1_{\text{Ancienneté du permis='}< 3 \text{ ans'}}\right\}} + \ldots + \beta_{19} 1_{\{1_{\text{Formule='F1'}}\}} + \ldots \\ &+ \beta_{22} 1_{\{1_{\text{Genre='CUS'}}\}} + \ldots + \beta_{29} 1_{\{1_{\text{Durée de détention du véhicule='0 an'}}\}} + \ldots \\ &+ \beta_{35} 1_{\left\{1_{\text{Usage='Privé'}}\right\}} + \ldots + \beta_{37} 1_{\left\{1_{\text{Antécédent/Expérience='Sans'}}\right\}} \\ &+ \text{offset}\,(\ \log(\text{nombre d'années polices})) \end{aligned}$$

où Z suit une binomiale négative

2.3.3 Validation des modèles et extraction de l'effet véhicule

Avant de pouvoir extraire l'effet véhicule de validation, il convient de valider les modèles construits.

A - Validation des modèles

Afin de valider le modèle, nous avons testé pour commencer la significativité des coefficients. Ensuite, une analyse des résidus a été réalisée. Pour finir, les valeurs prédites et observées sont comparées.

Test de significativité des coefficients

La significativité des paramètres estimés est ici analysée. Cette significativité s'observe par rapport à une modalité de référence. Pour chaque variable, la modalité de référence est la modalité la plus représentée dans la base de données. L'individu caractérisé par toutes les modalités de référence est considéré comme l'individu moyen. Dans le cas d'un modèle avec une fonction lien logarithmique, cet individu a pour valeur prédite l'exponentielle de l'intercept. Comme l'indique la figure 2.18, les coefficients utilisés sont significatifs. Le constat est le même pour le coût moyen. Dans le cas où certaines modalités ne seraient pas significatives, il convient de faire des regroupements.

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     [illegible]  0.03525 -98.017  < 2e-16 ***
ancper1:ancper <3 ans           [illegible]  0.02637  23.776  < 2e-16 ***
ancper2:ancper 3-4 ans          [illegible]  0.03012  14.683  < 2e-16 ***
ancper3:ancper 5-9ans           [illegible]  0.02320  14.696  < 2e-16 ***
ancper4:ancper 10-14ans         [illegible]  0.02341   8.272  < 2e-16 ***
ancper5:ancper 15-19ans         [illegible]  0.02309   3.353 0.000799 ***
ancper7:ancper > 29 ans         [illegible]  0.02131  -6.200 5.66e-10 ***
usage1 Privé                    [illegible]  0.01510 -17.739  < 2e-16 ***
usage3 Professionel/Tourn       [illegible]  0.02356   5.678 1.36e-08 ***
anccir1:anccir <1 an            [illegible]  0.02986  22.090  < 2e-16 ***
anccir10:anccir 15 ans +        [illegible]  0.03310  -6.161 7.22e-10 ***
anccir2:anccir 1 an             [illegible]  0.03036  16.532  < 2e-16 ***
anccir3:anccir 2 ans            [illegible]  0.03108  13.387  < 2e-16 ***
anccir4:anccir 3 ans            [illegible]  0.03112  11.047  < 2e-16 ***
anccir5:anccir 4 ans            [illegible]  0.03094   9.664  < 2e-16 ***
anccir6:anccir 5 ans            [illegible]  0.03099   8.087 6.11e-16 ***
anccir7:anccir 6-7 ans          [illegible]  0.02765   7.963 1.68e-15 ***
anccir8:anccir 8-9 ans          [illegible]  0.03044   4.551 5.33e-06 ***
ddv1 : 0 an                     [illegible]  0.02465  10.965  < 2e-16 ***
ddv2 : 1 an                     [illegible]  0.02286   7.473 7.81e-14 ***
ddv4 : 3 ans                    [illegible]  0.02498  -2.608 0.009096 **
ddv5 : 4 ans                    [illegible]  0.02708  -4.506 6.61e-06 ***
ddv6 : 5 ans                    [illegible]  0.03039  -4.959 7.09e-07 ***
ddv7 : 6-7 ans                  [illegible]  0.03073  -7.539 4.74e-14 ***
ddvplus de 8 ans                [illegible]  0.03018  -8.144 3.83e-16 ***
ant_expSans                     [illegible]  0.02275  15.141  < 2e-16 ***
zone1: Zone 1 à 3               [illegible]  0.02498  -8.664  < 2e-16 ***
zone3: Zone 7 à 9               [illegible]  0.01912  30.458  < 2e-16 ***
zone4: Zone 10 à 13             [illegible]  0.01958  51.740  < 2e-16 ***
formuleF1: RC                   [illegible]  0.02061 -33.369  < 2e-16 ***
formuleF2: RC+IV                [illegible]  0.02645 -12.939  < 2e-16 ***
formuleF3: RC+IV+DC             [illegible]  0.02582  -8.558  < 2e-16 ***
genreCUS                        [illegible]  0.02777 -15.783  < 2e-16 ***
genreGTO                        [illegible]  0.03025   2.988 0.002810 **
genreQUAD/END/EXO/SID           [illegible]  0.07007 -13.596  < 2e-16 ***
genreROA                        [illegible]  0.02246 -15.340  < 2e-16 ***
genreROU/BAS                    [illegible]  0.02530 -16.836  < 2e-16 ***
genreSPO/HYP                    [illegible]  0.03278   1.959 0.050073 .
genreTRA/SMO                    [illegible]  0.02549 -16.232  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 2.18 – Significativité des coefficients du modèle de fréquence

Analyse des résidus

L'analyse des résidus permet de mettre en avant le pouvoir explicatif du modèle.

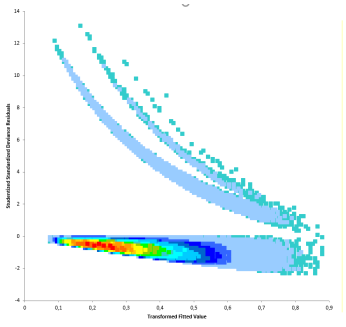


FIGURE 2.19 – Résidus du modèle fréquence

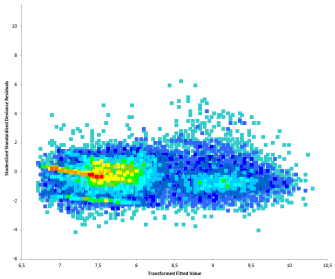


FIGURE 2.20 – Résidus du modèle coût

Les résidus pour le modèle de fréquence ont une allure atypique. La plupart des résidus sont centrés en 0. Les résidus supérieurs à 0 se justifient par la présence de sinistres avec peu d'années d'exposition. Cette allure des résidus est classique dans un modèle de fréquence. Le nuage de point ne semble pas avoir d'allure particulière, nous considérons que l'hypothèse de constance de la variance est vérifiée.

Pour le modèle de coût, nous notons que les points les plus éloignés du nuage sont ceux proches du seuil d'écrêtement. L'analyse graphique des résidus conclut que le modèle de coût semble relativement bien respecter les hypothèses des GLM.

Comparaison des variables prédites et observées

Pour que le modèle ait un bon pouvoir explicatif, l'écart entre la courbe observée et prédite pour chaque variable doit être le plus petit possible. Afin de juger du caractérère explicatif du modèle, nous avons comparé les écarts entre les moyennes observées et modélisées. Les deux modèles semblent corrects.

Pour exemple, pour le V_{Groupe}, nous observons que le coût moyen modélisé s'ajuste relativement bien à nos données alors que cette variable ne fait pas partie des variables du modèle. Le modèle semble correct même sans inclure cette variable dans le modèle.

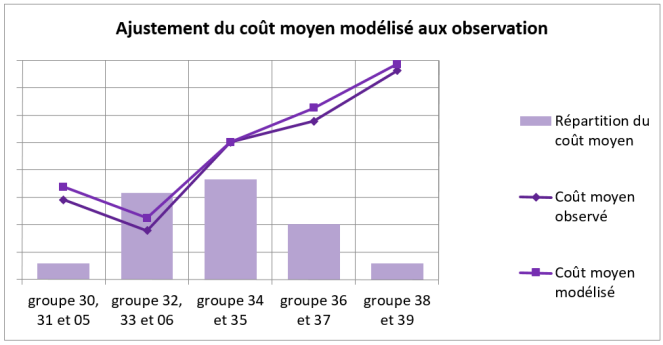


FIGURE 2.21 – Comparaison des variables prédites et observées sur le modèle de coût

B - Extraction de l'effet véhicule

Les modèles GLM construits et validés, nous allons pouvoir extraire l'effet véhicule. Pour cela, comme explicité dans le chapitre précédent, deux démarches vont être utilisées : la "méthode relative" et la "méthode résiduelle" ; chacune étant appliquée à la fois à la maille contrat et la maille véhicule.

La "méthode résiduelle" a été mise en place en recalculant les coefficients du modèle GLM sans variable véhicule. L'effet véhicule dans cette méthode n'est autre que le résidu du modèle recalculé.

La «méthode relative» présentée à la section 1.9 a été mise en place et utilisée pour isoler l'effet véhicule en divisant le résidu du modèle GLM (avec variables véhicule et non véhicule) par le coefficient des variables véhicule. Les premiers résultats de cette méthode ont mis en évidence une très forte sensibilité des résultats selon le profil moyen choisi (c'est-à-dire caractérisé par toutes les modalités de référence) de par l'intégration des variables véhicule dans le profil de référence. L'effet véhicule extrait selon la "méthode relative" donnant des arbres non stable selon le profil choisi, il a été décidé de ne considérer que la "méthode résiduelle" qui donne des résultats robustes.

Quartiles de l'effet véhiule	Maille	Q1	Médiane	Moyenne	Q3
Modèle Fréquence	Maille contrat	-0,02335	-0,01112	0,03563	-0,00509
	Maille véhicule	0,008701	0,02431	0,03563	0,04993
Modèle Coût	Maille contrat	-1889	-712,2	-113,3	637,4
	Maille véhicule	-1337	-540,4	-113,3	749,6

FIGURE 2.22 – Quartiles de l'effet véhicule selon le modèle et la maille

Comme l'indique la figure 2.22 présentant les quartiles de l'effet véhicule pour chaque modèle et chaque maille, l'effet véhicule résultant de "la méthode résiduelle" peut prendre des valeurs négatives. Il s'interprète non plus comme un coût ou une fréquence, mais comme une partie résiduelle que nous allons chercher à expliquer dans le prochain chapitre par des Machine Learning.

À la maille véhicule, la base de données pour le coût est composée de 1880 véhicules, celle pour la fréquence est composée de 3221 véhicules alors que, à la maille contrat, la base de données pour le coût est composée de 1177800 lignes et la modèle de fréquence de 19529 lignes.

Chapitre 3

Elaboration et étude de la pertinence des véhiculiers à dire de machines

L'effet véhicule identifié et extrait lors du chapitre précédent, va être ici exploité par des Machine Learning pour construire des véhiculiers. En comparant ces véhiculiers à dire de machines avec ceux à dire d'experts, nous allons être en mesure d'évaluer la capacité ou non des machines à constuire un véhiculier pertinent et ainsi d'apprécier le degré d'intervention adapté de l'avis d'experts.

3.1 Première approche : arbre CART et crédibilité

Comme présenté au premier chapitre, un premier modèle combinant régression CART et crédibilité va être mis en place. Afin d'avoir un critère objectif pour choisir la maille la plus appropriée entre la maille contrat et celle véhicule, il a été choisi préalablement de mettre en place et comparer ces deux mailles.

3.1.1 Etape 2 - Arbre CART à la maille contrat

La fonction *rpart* du logiciel R a été utilisée pour construire les arbres. Nous avons laissé chaque arbre "s'étendre" sans critère d'arrêt. Puis, afin de limiter le sur-apprentissage ils ont ensuite été respectivement "élagués" en définissant le paramètre de la fonction de complexité *cp* qui minimise l'erreur relative par validation croisée.

La validation croisée par défaut dans la fonction *rpart* consiste à subdiviser la base de données en dix échantillons, puis à successivement utiliser neuf échantillons pour construire un arbre et le dixième comme base de test.

Les arbres ont été dans un premier temps construits à partir d'un effet véhicule à la maille contrat.

Remarques générales : Les graphiques 3.1 et 3.2 présentent l'évolution de l'erreur relative par rapport au *cp* pour le modèle de fréquence et celui de coût. Nous notons tout d'abord que **le pouvoir explicatif des arbres est relativement faible**, l'erreur relative reste en effet proche de 1 quelle que soit la valeur *cp* choisie. La subdivision des données explique ainsi une faible part de la variance. Par ailleurs, les lignes verticales, correspondant à l'écart type de l'erreur, ont une amplitude importante, suggérant des écarts relativement conséquents dans les erreurs. **Ce dernier point semble indiquer une importante variabilité de la variable d'intérêt**.

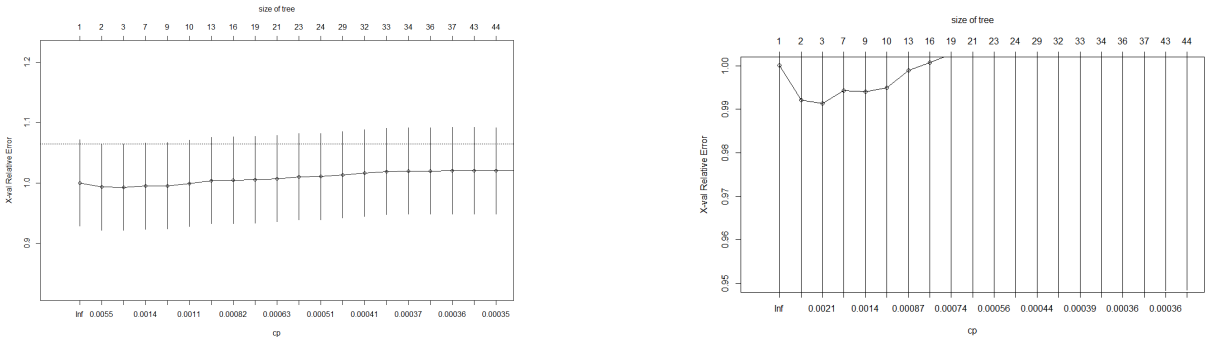


FIGURE 3.1 – Maille contrat - Coût - Erreur relative par rapport au paramètre de complexité (le graphique de droite est un zoom de celui de gauche sur les premières valeurs *cp*)

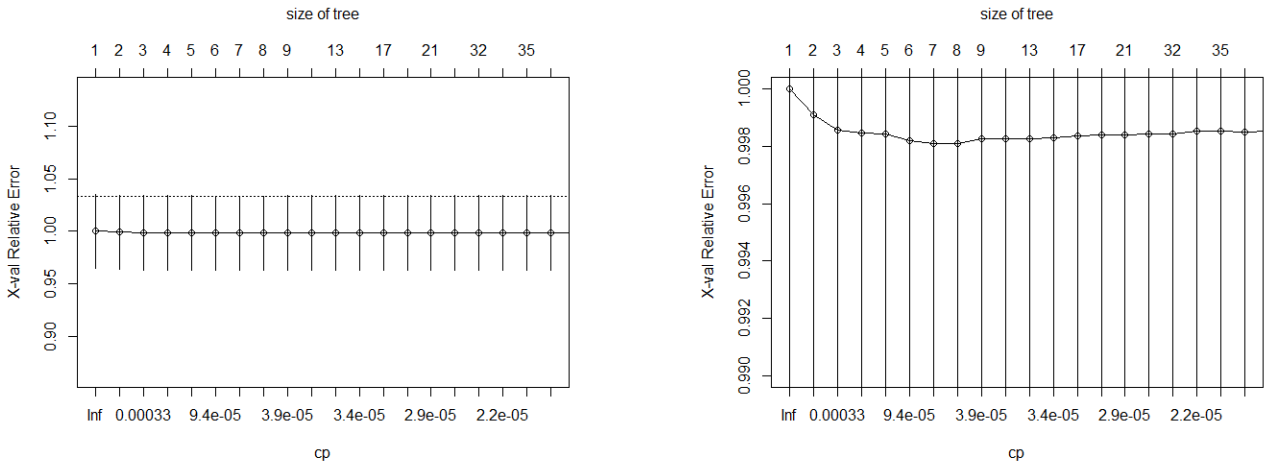


FIGURE 3.2 – Maille contrat - Fréquence - Erreur relative par rapport au paramètre de complexité (le graphique de droite est un zoom de celui de gauche sur les premières valeurs *cp*)

Modèle de coût : le *cp* optimal donne un arbre à 3 nœuds soit 4 feuilles, cet arbre sera nommé par la suite *arbre-cm-mc-c*0 [1]. Le nombre de feuilles de l'arbre *arbre-cm-mc-c*0 semble insuffisant. C'est pourquoi une seconde élagation va être mise en place. Pour ce second arbre, une valeur de *cp* de $0,0014$ a été choisie permettant d'obtenir un arbre de 9 feuilles sans augmenter fortement l'erreur relative. Dans ce nouvel arbre nommé ci-après *arbre-cm-mc-c*0-*bis* (cf figure 3.3), nous avons identifié 3 feuilles potentiellement non fiables : la cinquième, la sixième et la dernière feuille en partant de la gauche figure 3.3. En effet, la cinquième feuille est composée de 39 sinistres correspondant à 12 véhicules. Et, parmi ces véhicules, un unique véhicule représente plus de 5 sinistres. Dès lors, sur ces 39 sinistres, seuls plus d'un $1/3$ sont liés à des sinistres sous-représentés. Le constat est encore plus frappant sur le sixième nœud composé de seulement 16 sinistres correspondant à 11 véhicules qui sont en moyenne associés à 1,6 sinistres seulement. De plus, parmi ces 11 véhicules, le véhicule associé au plus grand nombre de sinistres n'a que 4 sinistres au compteur.

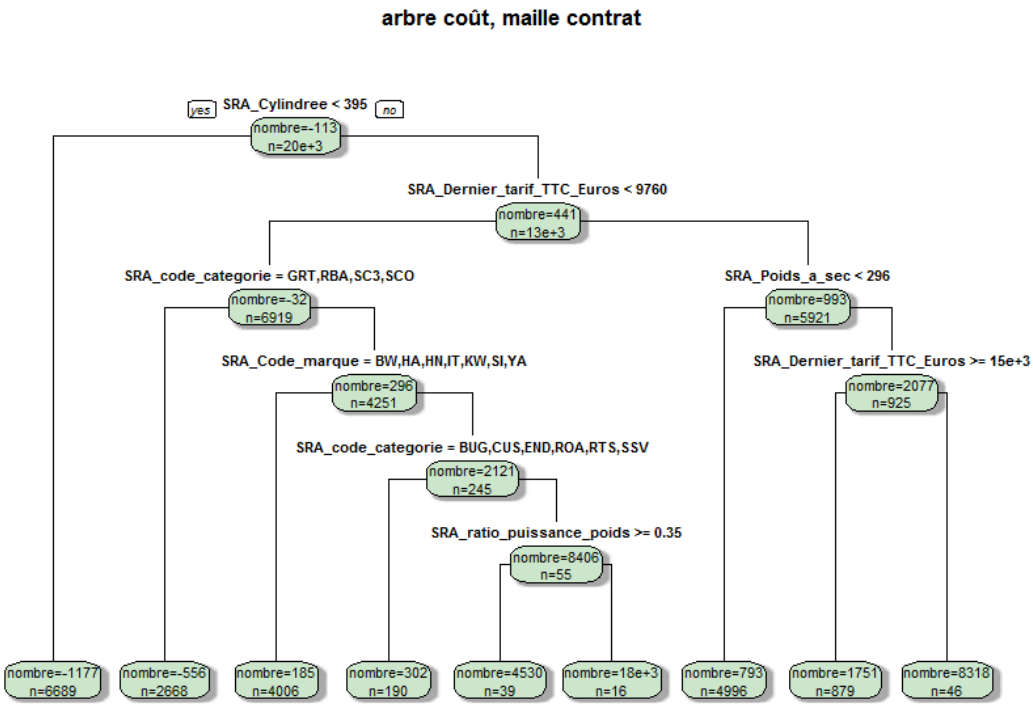


FIGURE 3.3 – Arbre *arbre-cm-mc-c*0-*bis*

1. La structure de la notation des arbres est la suivante : arbre-modèle-maille-crédibilité-version. Les modèles possibles sont le modèle de coût (cm) ou le modèle de fréquence (freq). Les mailles possibles sont la maille contrat (mc) et la maille véhicule (mv). c0 correspond au niveau de crédibilité inexistant à ce stade.

Modèle de fréquence : le *cp* optimal donne l'arbre nommé *arbre-freq-mc-c*0 à 9 feuilles présenté figure 3.4. L'analyse détaillée des feuilles ne met pas en avant de feuilles potentiellement non fiables. De tout évidence, chaque feuille est constituée de véhicules associés de à nombreuses images.

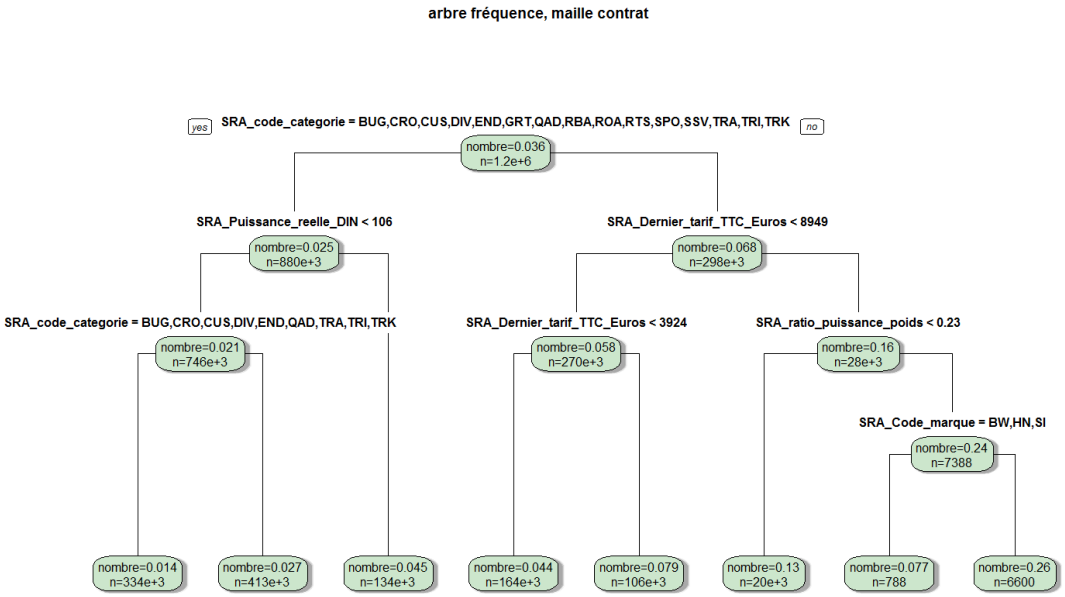


FIGURE 3.4 – Arbre *arbre-freq-mc-c*0

3.1.2 Etape 2 bis - Arbre CART à la maille véhicule

Dans un second temps, d'autres arbres CART ont été construits selon la même méthode à partir de l'effet véhicule à la maille véhicule.

Modèle de coût : Nous observons à la figure 3.5, que l'erreur relative est croissante par rapport au *cp*. L'arbre optimal est donc réduit à une racine.

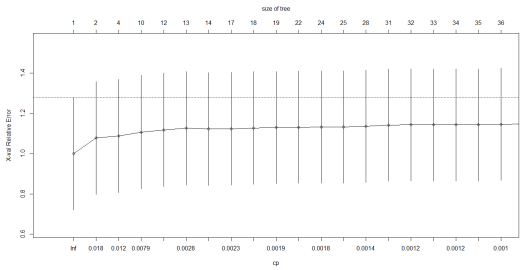


FIGURE 3.5 – Maille véhicule - Coût - Erreur relative par rapport au paramètre de complexité

D'une part, ceci pourrait s'expliquer par la sensibilité de la validation croisée aux bases de données de petite taille. Cela pourrait d'autre part être dû à la forte variabilité de la variable d'intérêt. Afin de comparer les résultats entre les mailles, un *cp* de $0,0079$ est choisi afin d'avoir un arbre nommé *arbre-cm-mv-c*0 de taille similaire à *arbre-cm-mc-c*0 (voir figure 3.6).

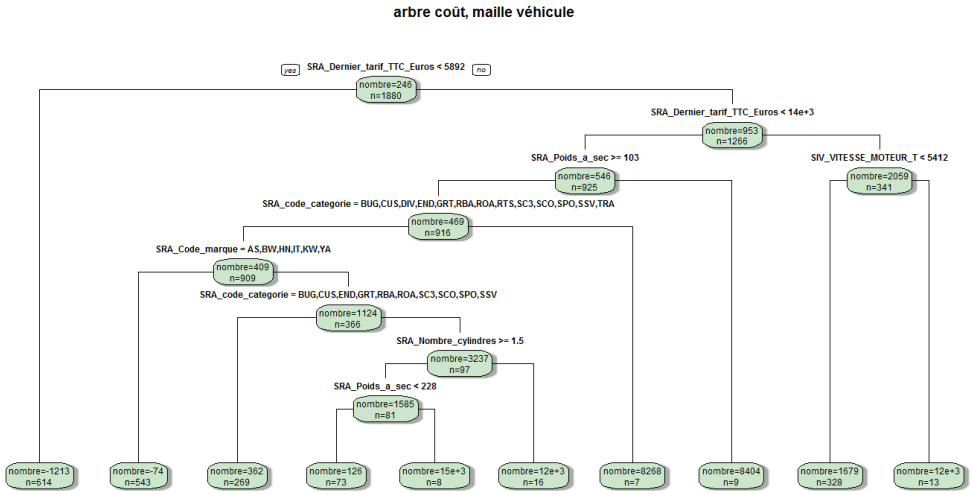


FIGURE 3.6 – Arbre *arbre-cm-mv-c*0

L'existence de nœuds potentiellement non fiables est assez intuitive à la maille véhicule dans la mesure où les véhicules sont considérés indépendamment du nombre de sinistres qu'ils représentent. Ainsi, le cinquième nœud en partant de la gauche figure 3.6 est composé de 8 véhicules correspondant au total à 28 images. Parmi ceux-ci, seul un véhicule représente plus de 10 sinistres, les autres véhicules représentant au plus respectivement 3 sinistres.

Par ailleurs, comme l'indique la figure 3.6, le passage à la maille véhicule a permis de réduire le coût maximal prédit par feuille.

En comparant les deux mailles, la valeur du véhicule, la marque, et la catégorie apparaissent comme les variables les plus utilisées par les arbres.

Modèle de fréquence : Nous constatons que, comme pour le modèle de coût, l'erreur relative est croissante avec le *cp*. Selon la même logique, l'arbre est élagué selon un *cp* de $0,006$ donnant un arbre *arbre-freq-mv-c*0 de 9 feuilles présenté à la figure 3.7.

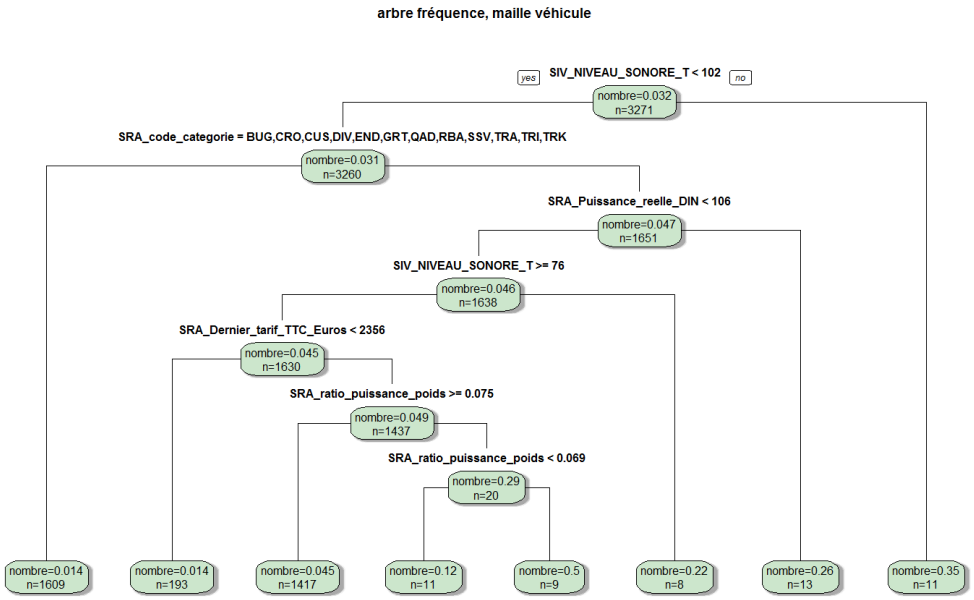


FIGURE 3.7 – Arbre *arbre-freq-mv-c*0

Les règles de décision de l'arbre *arbre-freq-mv-c*0 par rapport à la maille contrat sont significativement différentes. Une comparaison des classements par tableau croisé a mis en évidence un écart important de classement des véhicules d'une maille à l'autre.

Les variables les plus discriminantes pour le modèle de fréquence, toute maille confondue, sont la catégorie du véhicule ainsi que sa puissance réelle, sa valeur, et son ratio puissance/poids. Ce sont en effet ces variables qui sont le plus utilisées pour la scission des noeuds. Nous constatons, par ailleurs, que la variable niveau sonore est segmentante à la maille contrat mais pas à la maille véhicule.

La mise en place de régression CART a permis d'identifier les variables véhicule les plus discriminantes pour l'effet véhicule de coût comme de fréquence. Si, à ce stade, aucune conclusion ne peut être tirée sur la pertinence des mailles ou la qualité de regroupements construits pour chaque maille, l'allure de l'erreur relative par rapport au *cp* et la forte dissimilarité des arbres entre les deux mailles semblent indiquer un mauvais apprentissage des arbres. Dans la prochaine section, nous allons essayer d'améliorer leur apprentissage à l'aide du modèle de Bühlmann-Straub.

3.1.3 Etape 3 - Mise en place de la crédibilité et conclusions

Comme évoqué précédemment, la première approche se propose d'intégrer la théorie de la crédibilité afin d'améliorer l'apprentissage des arbres.

A - Mise en place

Pour ce faire, le modèle de Bühlmann-Straub tel que défini au premier chapitre est appliqué aux véhicules à l'aide de la fonction *cm* du package "actuar" sous R. A titre de rappel, la crédibilité est utilisée dans cette étude non pas pour crédibiliser des estimations, mais pour estimer les véhicules dont l'information est suffisante pour être jugés comme crédibles. Pour chaque modèle, un niveau de crédibilité va être choisi et, par la suite, seuls les véhicules ayant un facteur de crédibilité supérieur ou égal à ce seuil seront utilisés en input de l'arbre CART.

Les hypothèses prérequises à l'application du modèle de Bühlmann-Straub sont supposées vérifiées.

B - Choix du paramètre

Modèle de coût : Le choix du facteur de crédibilité tient compte de trois éléments résumés figure 3.8 : le nombre d'observations de la variable d'intérêt, le nombre de véhicules et le représentativité de la variable d'intérêt par véhicule. Par exemple, pour le modèle de coût, en choisissant un seuil de $0,40$, $76,24\%$ des sinistres seraient conservés, mais uniquement 23% des véhicules de la base totale des sinistres seraient représentés. Le choix de ce seuil enlèverait les véhicules ayant moins de 11 sinistres. En arbitrant sur ces trois composantes, le niveau de crédibilité choisi pour le modèle de coût est de $0,25$. Par ce choix, $87,75\%$ des sinistres sont conservés et seulement 40% des véhicules. Ainsi, seuls les véhicules ayant au moins 5 images sont conservés.

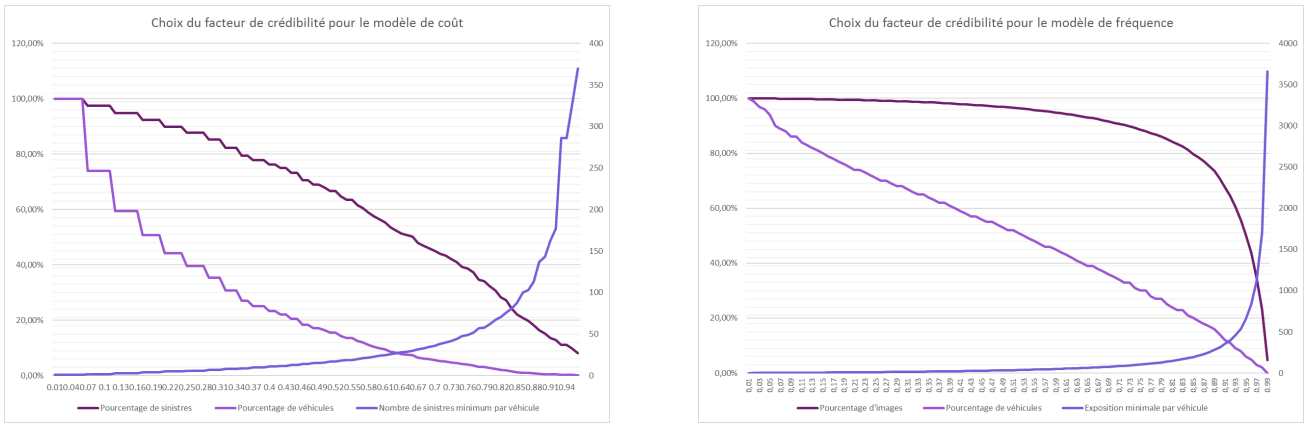


FIGURE 3.8 – Choix des seuils du facteur de crédibilité

Modèle de fréquence : Pour le modèle de fréquence, les trois critères équivalents sont résumés figure 3.8. Pour ce modèle, trois seuils ont été choisis :

- $Z = 0,53$ permet de conserver $96,21\%$ des images et 50% des véhicules de la base initiale. Les véhicules ayant moins de 38 images ne sont pas exclus.
- $Z = 0,60$ permet de conserver $94,62\%$ des images et 44% des véhicules de la base initiale. Les véhicules ayant moins de 50 images sont ainsi exclus.
- $Z = 0,70$ permet de conserver $91,14\%$ des images et 35% des véhicules de la base initiale. Les véhicules ayant moins de 80 images sont ainsi exclus.

C - Nouveaux CART

Maintenant que les seuils de crédibilité ont été définis, les données sont sélectionnées et de nouveaux arbres sont alors bâtis.

Maille contrat :

Modèle de coût : La sélection des images par crédibilité à la maille contrat ne modifie pas le nombre de feuilles optimal. Deux *cp* sont alors définis afin d'obtenir des arbres plus segmentants. Le premier *arbre-cm-mc-c*25 est composé de 5 feuilles. Le second arbre *arbre-cm-mc-c*25-*bis* de 18 feuilles est présenté figure 3.9.

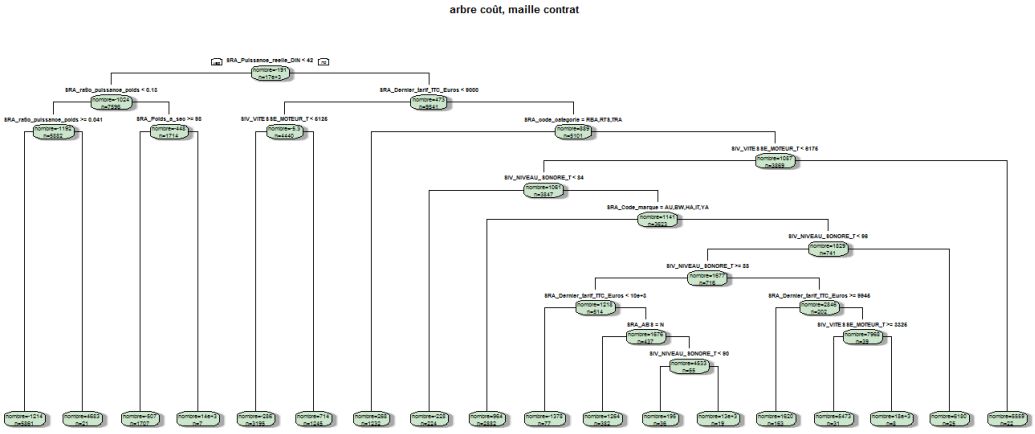


FIGURE 3.9 – Arbre *arbre-cm-mc-c*25-*bis*

Nous constatons que **l'exclusion des véhicules les moins représentés modifie significativement les règles de l'arbre**.

Modèle de fréquence : La sélection des images par crédibilité à la maille contrat améliore considérablement le nombre optimal de feuilles qui passe de 4 feuilles à 23 feuilles pour un facteur de crédibilité de $0,53$ (respectivement 14 et 22 pour les niveaux $0,60$ et $0,70$).

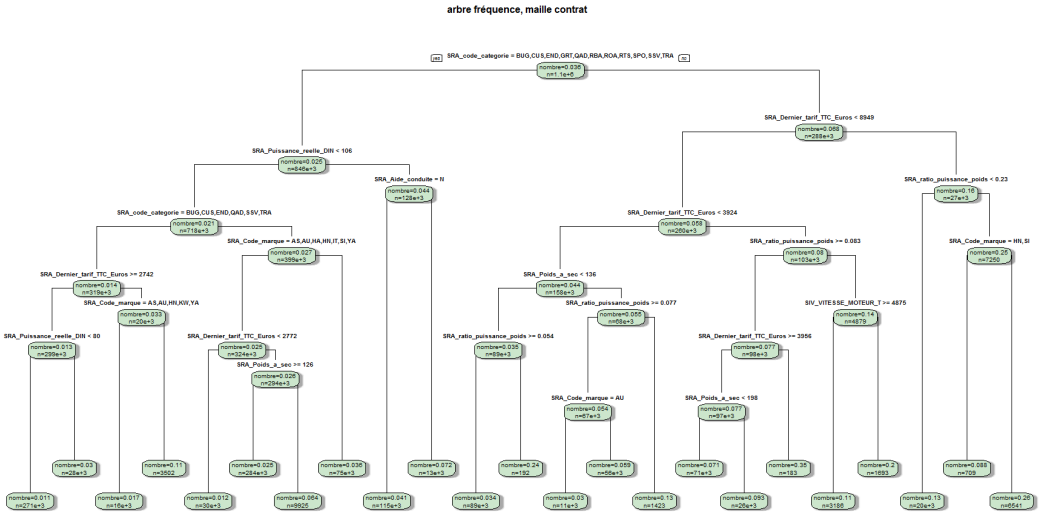


FIGURE 3.10 – Arbre *arbre-freq-mc-c*53

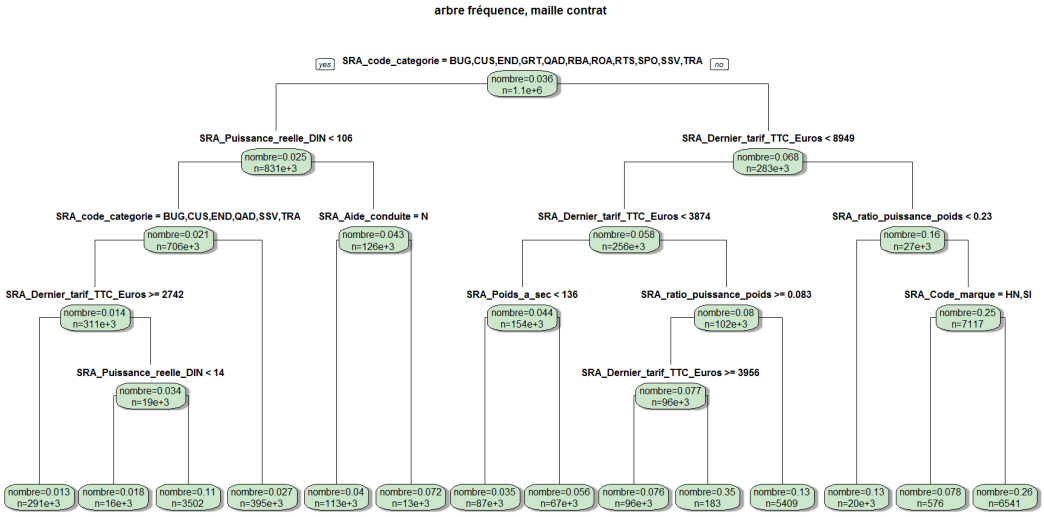


FIGURE 3.11 – Arbre *arbre-freq-mc-c*60

Nous constatons que les règles de décisions des arbres après crédibilité sont relativement proches de celles de l'arbre *arbre-cm-mc-c*0.

Maille véhicule :

Modèle de coût : L'étape de crédibilité à la maille véhicule permet de résoudre le problème de croissance de l'erreur relative par rapport au *cp*, ainsi l'arbre optimal a maintenant 3 feuilles au lieu d'être réduit à une racine. En observant la figure 3.12 de plus près, nous constatons que la courbe présente une légère constante aux minimums.

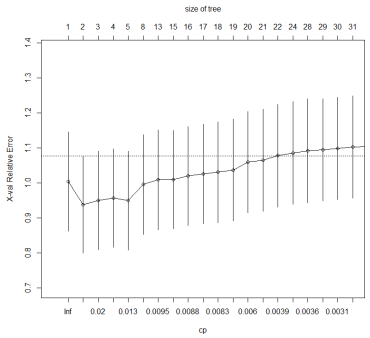


FIGURE 3.12 – Maille véhicule - Coût - Crédibilité $Z = 0,25$ - Erreur relative par rapport au paramètre de complexité

Nous choisissons d'élaguer l'arbre selon un *cp* de $0,013$. L'arbre obtenu nommé *arbre-cm-mv-c*25 est composé de 5 feuilles. Un deuxième *cp* est choisi afin de pouvoir challenger l'*arbre-cm-mv-c*25 à un arbre plus grand. L'arbre *arbre-cm-mc-c*25*-bis* de 18 feuilles alors obtenu est présenté figure 3.13.

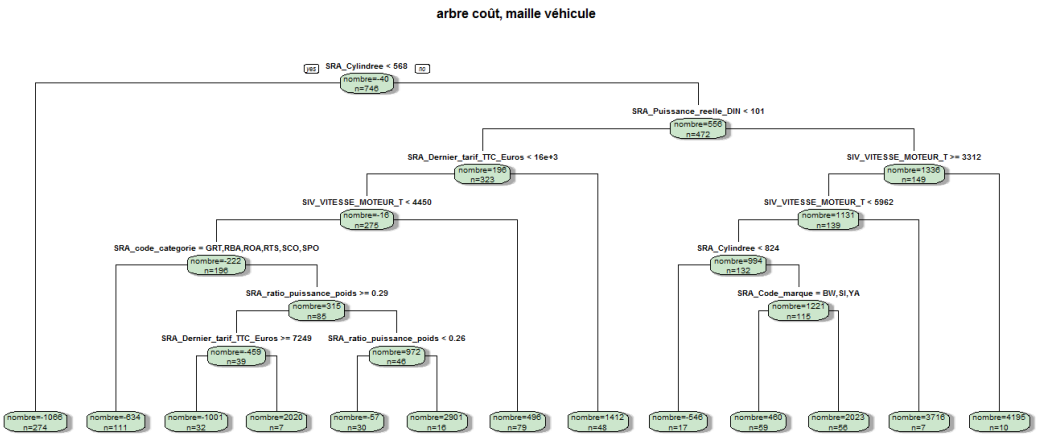


FIGURE 3.13 – Arbre *arbre-cm-mv-c*25*-bis*

Nous constatons que **l'exclusion des véhicules les moins représentés modifie significativement les règles de l'arbre**.

Modèle de fréquence : L'étape de crédibilité à la maille véhicule permet d'obtenir une décroissance de l'erreur relative sur les premiers *cp*. Pour le niveau de crédibilité 0,60 (respectivement 0,53 et 0,70), la taille optimale de l'arbre est de 14 feuilles (respectivement 6 et 5). Pour le niveau de crédibilité, la légère constante de la courbe *cp* aux valeurs minimales nous permet de considérer un deuxième *cp* donnant lieu à un arbre nommé *arbre-freq-mv-c*53-*bis* de 18 feuilles.

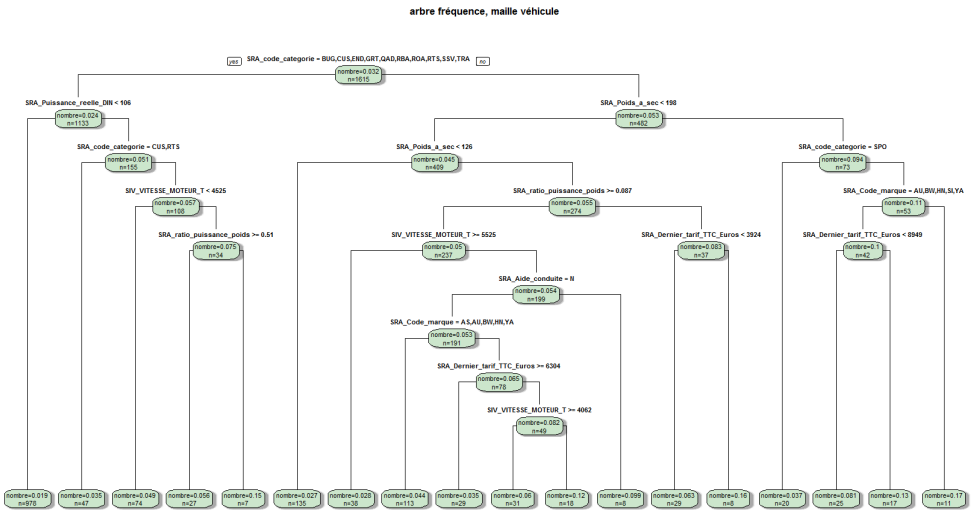


FIGURE 3.14 – Arbre *arbre-freq-mv-c*53-*bis*

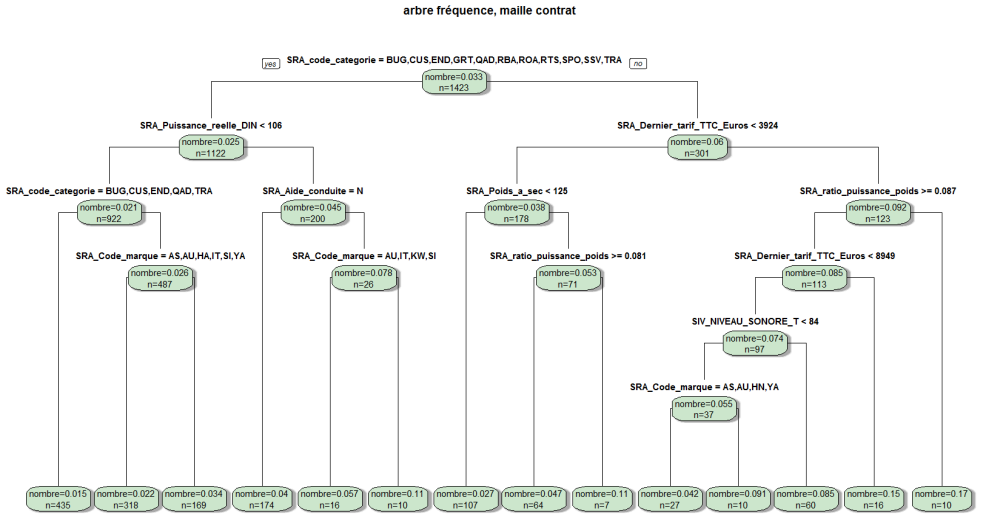


FIGURE 3.15 – Arbre *arbre-freq-mv-c*60

Les règles de l'arbre *arbre-freq-mv-c*0 sont très différentes des règles après l'étape de crédibilité. Cependant, nous observons que les différents arbres construits après l'étape de crédibilité ont des règles très proches. **Plus intéressant encore, l'arbre de niveau de crédibilité** $0,60$ **est très sensiblement le même que les arbres de la maille contrat après crédibilité**.

Au final, l'étape de crédibilité semble globalement améliorer l'apprentissage de l'arbre. En effet, hormis le modèle de coût à la maille contrat, les autres arbres ont tous une taille optimale supérieure après crédibilité.

Par ailleurs, dans son ensemble les arbres du modèle de coût sont sensibles aussi bien à la maille qu'à la sélection des données par crédibilité donnant ainsi des arbres aux règles significativements différentes.

Le modèle de fréquence est moins sensible. Si les règles de l'arbre se sont révélées différentes selon maille sans crédibilité, l'étape de crédibilité a non seulement créée des arbres relativement proches de ceux avant crédibilité pour chaque maille respective, mais a également formé après crédibilité des arbres similaires entre les mailles. Ce dernier point laisse à penser, qu'après crédibilité, la classification des véhicules dans le modèle de fréquence est sensiblement la même quelque soit la maille utilisée.

Avant de pouvoir analyser la pertinence du véhiculier ainsi construit et de confirmer nos intuitions par des critères objectifs en comparant la pertinence de la crédibilité sur la base d'apprentissage et la base test en termes de diminution de l'erreur, nous allons construire un véhiculier selon une deuxième approche.

3.2 Seconde approche : Random Forest et lissage spatial

Un second véhiculier va être développé avec l'effet véhicule extrait lors du deuxième chapitre en combinant cette fois-ci Random Forest et lissage spatial.

3.2.1 Etape 2 - Ramdom Forest

Sur le même principe que la première approche, le Random Forest a été mis en place à la maille contrat et de la maille véhicule afin de définir la maille la plus adaptée. Pour ce faire, la fonction *randomForest* de R a été utilisée. Les paramètres par défaut de la fonction ont été employés à l'exception du Random Forest du modèle de fréquence à la maille contrat dont nous avons réduit le nombre d'arbres dans la forêt à 50 pour des contraintes informatiques. Par la suite, la forêt construite pour le modèle de fréquence (respectivement pour le modèle de coût) à la maille contrat sera nommée *rf-freq-mc* (respectivement *rf-cm-mc*), celle à la maille véhicule sera nommée rf-freq-mv (respectivement rf-cm-mv).

A - Analyse de l'importance des variables

La fonction *randomForest* sous R permet d'examiner l'influence des variables dans le modèle construit. Nous observons dans les graphiques 3.16 et 3.17 suivants les variables participant le plus à la pureté des nœuds c'est-à-dire leur qualité de scissions et de fiabilité des scissions créées.

Modèle coût : Nous remarquons, tout d'abord, que de manière cohérente aux attentes *a priori*, la variable plus importante pour le modèle de coût quelle que soit la maille est la valeur du véhicule (le prix du véhicule).

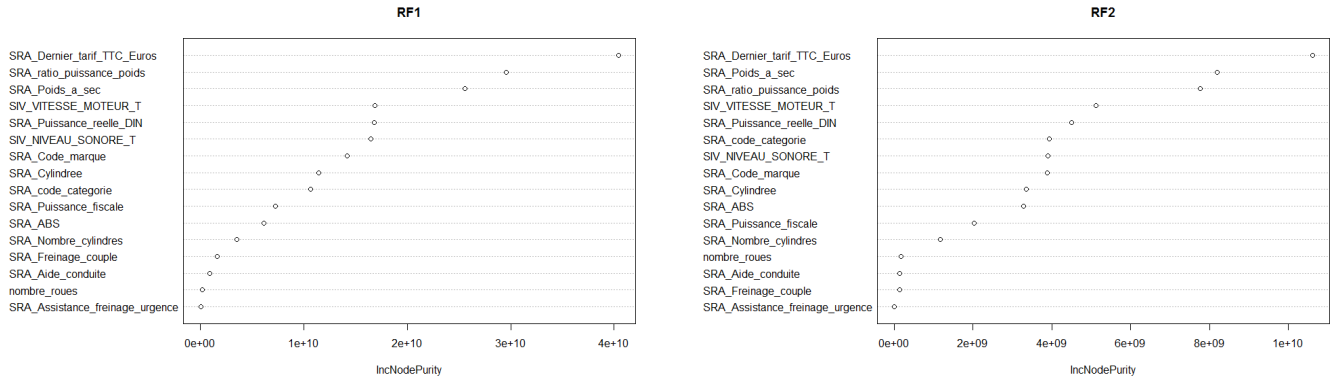


FIGURE 3.16 – Modèle de coût - Importance des variables véhicule selon la maille utilisée (maille contrat à gauche et maille véhicule à droite)

A l'appui des figures 3.16, le ratio puissance/poids et le poids apparaissent également comme des variables prédictives. Nous constatons, par ailleurs que, quelle que soit la maille, le poids et

la vitesse du moteur ont un pouvoir de prédiction plus grand que la puissance réelle. Ensuite, la marque des véhicules a une moindre importance que ce qui était attendu au préalable. Concernant les écarts entre les mailles véhicules, les classements des variables sont assez similaires. Nous constatons enfin que la marque du véhicule est un meilleur prédicteur que la catégorie à la maille contrat alors que le pouvoir explicatif de ces deux variables est sensiblement le même à la maille véhicule.

Modèle de fréquence : Nous relevons que le prix du véhicule est, pour le modèle de fréquence également, la variable la plus discriminante, suivie ensuite des variables de capacité du véhicule (ratio puissance/poids, poids, puissance).

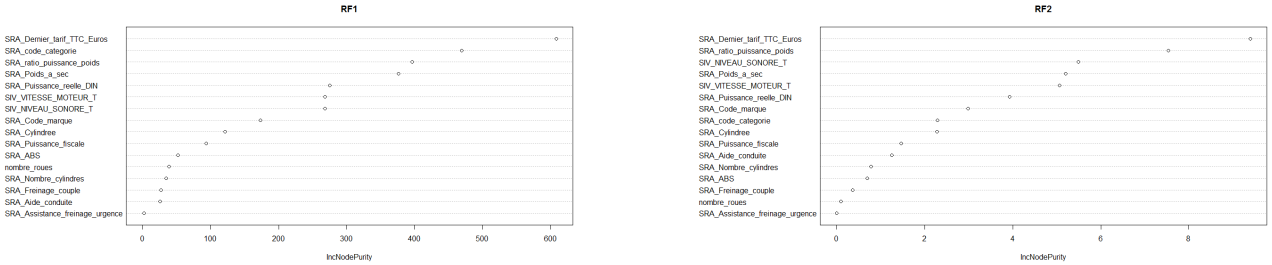


FIGURE 3.17 – Modèle de fréquence - Importance des variables véhicule selon la maille utilisée (maille contrat à gauche et maille véhicule à droite)

À la maille contrat, la catégorie est un très bon prédicteur pour la fréquence de sinistres. Ceci est cohérent avec les résultats d'études automobiles présentées section 1.3.1. Ce dernier constat n'est cependant pas vérifié à la maille véhicule.

A la maille véhicule toujours, le ratio puissance/poids est plus segmentant que le poids, lui-même plus segmentant que la puissance.

La catégorie, bien que n'ayant pas la même importance selon la maille, est dans tous les cas plus performante que la marque du véhicule.

B - Analyse de la prédiction des arbres

Nous avons ensuite comparé pour chaque véhicule, la prédiction des forêts avec les observations moyennes. Comme l'indique les graphes 3.18 et 3.19, les écarts entre les prédictions et observations moyennes pour chaque véhicule sont moindres pour les Random Forest construits à la maille contrat que ceux construits à la maille véhicule.

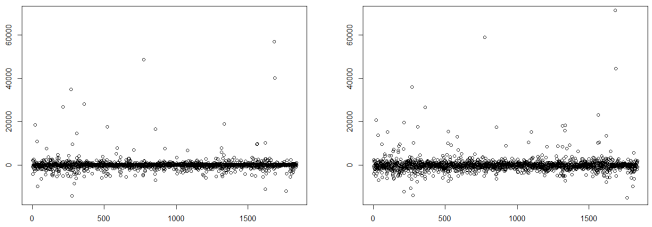


FIGURE 3.18 – Modèle de coût - Comparaison selon la maille des prédictions des Random Forest à la moyenne des observations par véhicule (à gauche la maille contrat et à droite la maille véhicule)

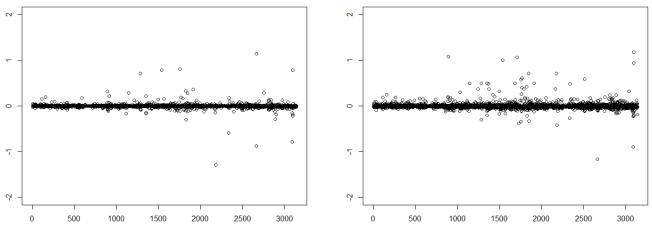


FIGURE 3.19 – Modèle de fréquence - Comparaison selon la maille des prédictions des Random Forest à la moyenne des observations par véhicule (à gauche la maille contrat et à droite la maille véhicule)

Ainsi malgré la forte hétérogénéité des données, particulièrement pour le modèle de coût, le Random Forest apparaît avoir un meilleur pouvoir prédictif à la maille contrat qu'à la maille véhicule.

3.2.2 Etape 3 préliminaire - Carte des véhicules

La carte des véhicules est construite selon la méthodologie développée par AXA Global P&C présentée à la section 1.3.3. Celle-ci est composée des trois phases clés détaillées ci-après : la projection en 3D, la triangulation et la création du voisinage et les corrections associées.

A - Les données utilisées

L'objectif de la carte des véhicules est double, elle permet non seulement de fiabiliser l'information des véhicules sous-représentés, mais également de récupèrer de l'information sur les véhicules absents de l'étude. La base de données SRA répond totalement à ces deux objectifs. Ceci n'est pas le cas de la base de données SIV qui nécessite le numéro d'immatriculation dont nous ne disposons que pour les véhicules assurés. Ainsi, les variables véhicule SRA sans valeur manquante utilisées pour construire les cartes des véhicules sont : l'ABS, l'aide à la conduite, l'assistance freinage d'urgence, la catégorie, la puissance fiscale, la marque, la cylindrée, le poids à sec, la puissance réelle, le dernier prix connu, le ratio puissance/poids, le freinage couple et le nombre de roues.

Selon ces critères, la base de données utilisée contient 5732 modèles de véhicules différents.

B - Les projections AFDM

La projection AFDM constitue la première étape de la constitution de la carte des véhicules. L'objectif ici est de projeter les véhicules dans un espace de dimension réduit pour répondre aux questions suivantes : Quelles sont les ressemblances / différences entre les individus ? Quelles sont les relations entre les variables ? D'un point de vue pratique, cette méthode a été mise à place à l'aide de la fonction *dudi.mix* du package *ade4* du logiciel R.

Choix du nombre d'axes : Des critères vont être utilisés afin de choisir le nombre d'axes :

- Le critère de Kaiser : tous les axes dont l'inertie est supérieure à 1 devraient être conservés ;
- Le critère de Karlis Sporta & Spinakis (KSS) : tous les axes dont l'inertie est supérieure ou égale à une valeur seuil devraient être gardés.

Cette valeur seuil est calculée de la manière suivante :

$$\text{Valeur du seuil} = 1 + 1,65\sqrt{\frac{p-1}{n-1}}$$

où $p = \frac{\text{variation totale}}{\text{inertie}} = 36$ et n =nombre d'observations= 5732

Le critère de Kaiser nous indique de conserver 16 axes pour une analyse approfondie. Les axes à conserver selon ce critère sont ceux en bleu sur le graphique 3.20. Le critère de KSS (en vert) nous indique de conserver les axes dont l'inertie est supérieure ou égale à $1,13$ soit 9 axes.

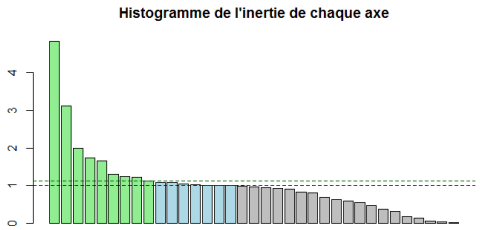


FIGURE 3.20 – Histogramme de l'inertie de chaque axe

Ici, l'axe 1 explique 2,21% de la variation totale, c'est l'axe qui a le plus grand pouvoir explicatif. Les 3 premiers axes expliquent cumulativement 27,64% de la variation totale.

Pour la suite de l'étude, les 3 premiers axes ont été sélectionnés pour les raisons suivantes :

- Au-delà de 3 dimensions, il est difficile de présenter et d'interpréter les résultats. Avec un plus grand nombre d'axes, le nombre maximum de voisins d'un véhicule (la notion de voisinage est expliquée par la suite) augmente considérablement (de 65 en 3D à 229 en 4D, 667 en 5D) ;
- Dans cette étude, l'AFDM n'est effectuée que pour extraire des informations latentes dans les données.

Ainsi, **seuls les 3 premiers axes seront utilisés par la suite**, cela signifie que chaque véhicule SRA sera représenté dans un espace en trois dimensions.

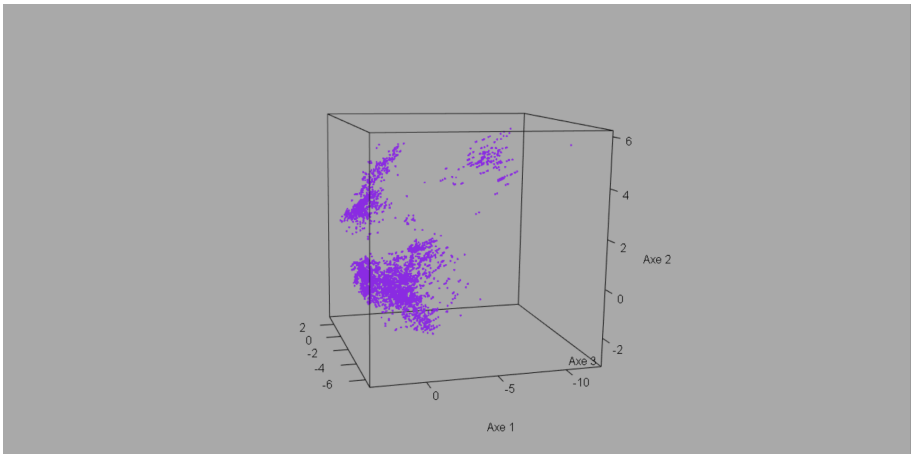


FIGURE 3.21 – Projection des véhicules en trois dimensions

C - Analyse de la projection

Une analyse de la composition des axes est indispensable afin de comprendre la manière dont les individus sont positionnés dans les axes principaux et *in fine* la manière dont notre carte de véhicule est construite.

Le tableau ci-dessous résume les variables et leurs contributions aux axes selon deux critères. La colonne « %Ligne » indique la proportion de chaque variable expliquée par les axes respectifs et la colonne « %Colonne » indique la proportion de chaque variable dans la composition des axes respectifs.

Variables	Type	Niveau	RS1	%Colonne	%Ligne	RS2	%Colonne	%Ligne	RS3	%Colonne	%Ligne
SRA_ABS	f	2	18%	4%	9%	8%	3%	4%	0%	0%	0%
SRA_Aide_conduite	f	2	3%	1%	1%	7%	2%	3%	4%	2%	2%
SRA_Assistance_freinage_urgence	f	2	5%	1%	3%	2%	1%	1%	11%	5%	5%
SRA_code_categorie	f	17	74%	15%	4%	92%	29%	5%	84%	42%	5%
SRA_Puissance_fiscale	q	1	87%	18%	87%	0%	0%	0%	3%	2%	3%
SRA_Code_marque	f	9	27%	6%	3%	33%	11%	4%	33%	16%	4%
SRA_Cylindree	q	1	87%	18%	87%	0%	0%	0%	2%	1%	2%
SRA_Poids_a_sec	q	1	41%	9%	41%	43%	14%	43%	1%	0%	1%
SRA_Puissance_reelle_DIN	q	1	7%	1%	7%	1%	0%	1%	11%	6%	11%
SRA_Dernier_tarif_TTC_Euros	q	1	72%	15%	72%	4%	1%	4%	1%	0%	1%
SRA_ratio_puissance_poids	q	1	36%	8%	36%	34%	11%	34%	2%	1%	2%
SRA_Freinage_couple	f	2	2%	0%	1%	5%	2%	2%	0%	0%	0%
nombre_roues	f	3	25%	5%	8%	84%	27%	28%	48%	24%	16%

FIGURE 3.22 – Tableau récapitulatif de la contribution des variables dans chaque axe

Le « % Ligne » doit être supérieur à $\frac{1}{p}$ où p est l'inertie totale de l'ensemble de données (ici, $p = 36$). Le « % Colonne » doit être supérieur à $\frac{1}{q}$ où q est le nombre total de variables à l'étude (ici, $q = 13$).

Les principales observations sont les suivantes :

- La catégorie, la marque et le nombre de roues sont expliqués par les trois axes ;
- La puissance des véhicules (puissance fiscale, cylindrée) et la valeur du véhicule sont expliquées dans une large mesure par le premier axe. La puissance réelle est en partie expliquée par le troisième axe ;
- Le poids et le ratio puissance/poids sont expliqués par les deux premiers axes ;
- L'ABS est partiellement expliqué pour les deux premiers axes, cependant les variables aide à la conduite, assistance de freinage d'urgence et freinage couple sont mals représentées dans l'espace.

Des études graphiques ont été réalisées afin d'appréhender le positionnement des véhicules dans l'espace 3D construit. Comme l'indique les graphiques 3.23 et 3.24, les variables *nombre de roues* et *catégorie* sont bien représentées dans l'espace.

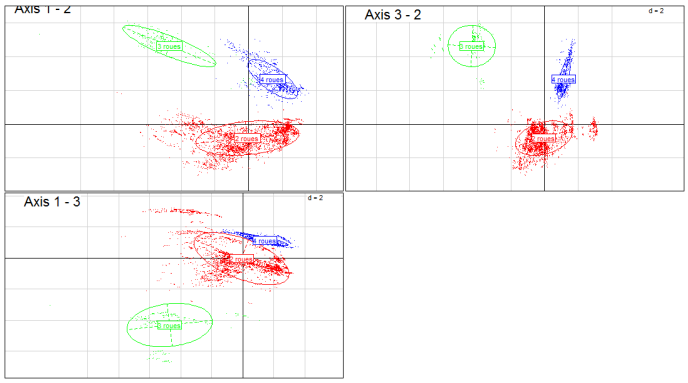


FIGURE 3.23 – Analyse du nombre de roues dans l'espace

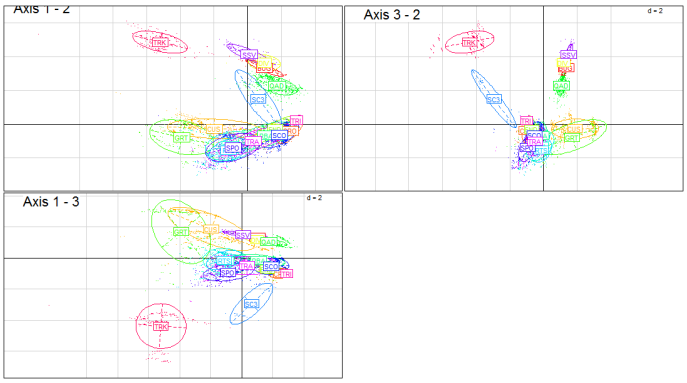


FIGURE 3.24 – Analyse de la catégorie dans l'espace

Le groupement de variables dans l'espace est nettement moins visible pour certaines variables [2]. La figure 3.25 montre que le prix du véhicule est bien modélisé dans le premier axe. Il est cependant plus difficile de désigner une zone des véhicules puissants par rapport aux autres véhicules (cf figure 3.26).

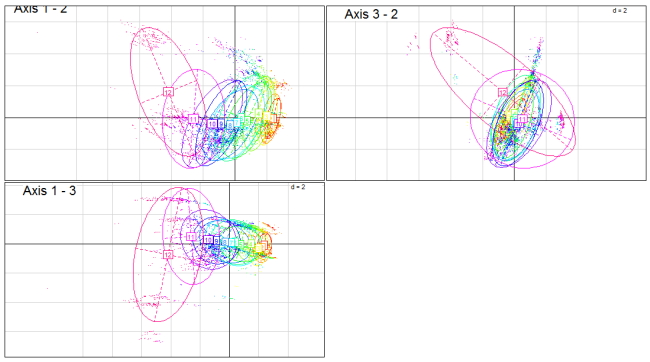


FIGURE 3.25 – Analyse de la variable dernier prix connu dans l'espace

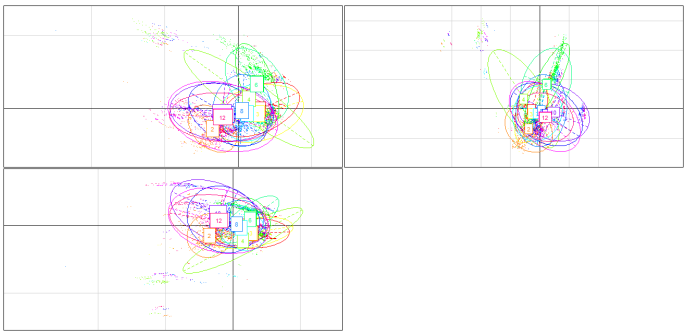


FIGURE 3.26 – Analyse de la variable puissance dans l'espace

2. Pour l'analyse graphique des figure 3.25 et 3.26, les variables quantitatives ont été regroupées en 12 groupes en utilisant les quantiles.

Les véhicules de la figure 3.27 sont colorés selon les quantiles auxquels ils sont associés. D'une part, les groupes de puissances différentes se repèrent visuellement. La projection en trois dimensions a ainsi bien conservé, dans une certaine mesure, l'information de la puissance. D'autre part, contrairement aux variables *nombre de roues* et *catégorie* qui sont nettement séparées dans l'espace, il est observé dans certaines zones de l'espace un mélange de couleur qui traduit une certaine hétérogénéité des véhicules. Ce dernier point n'est pas surprenant puisque, dans l'espace, les groupes sont formés selon leur genre, et qu'au sein d'un même genre il existe une hétérogénéité des autres variables.

La distance entre les véhicules sera la base de la carte des véhicules. Il conviendra donc par la suite de tenir compte des informations insuffisamment mises en avant par les distances, telle que la puissance du véhicule.

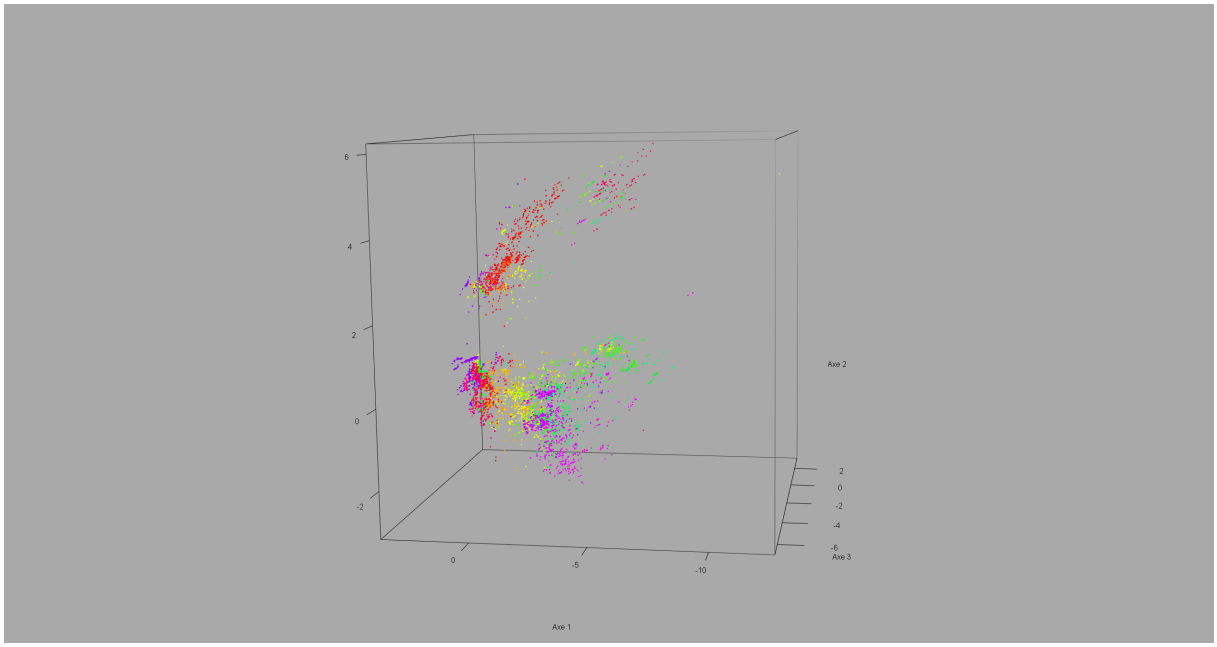


FIGURE 3.27 – Détails de la variable puissance en 3D

D - Triangulation de Delaunay

Maintenant que les véhicules sont projetés dans l'espace, la triangulation de Delaunay va être utilisée pour relier les véhicules entre eux à l'aide de la fonction *delaunayn* du package « geometry » de R. Les tétraèdres de Delaunay ont été tracés à partir du nuage de véhicules généré par l'ADFM, formant 36869 tétraèdres différents. Les véhicules reliés seront dits voisins ou adjacents. Nous connaissons à ce stade le voisinage de chaque véhicule.

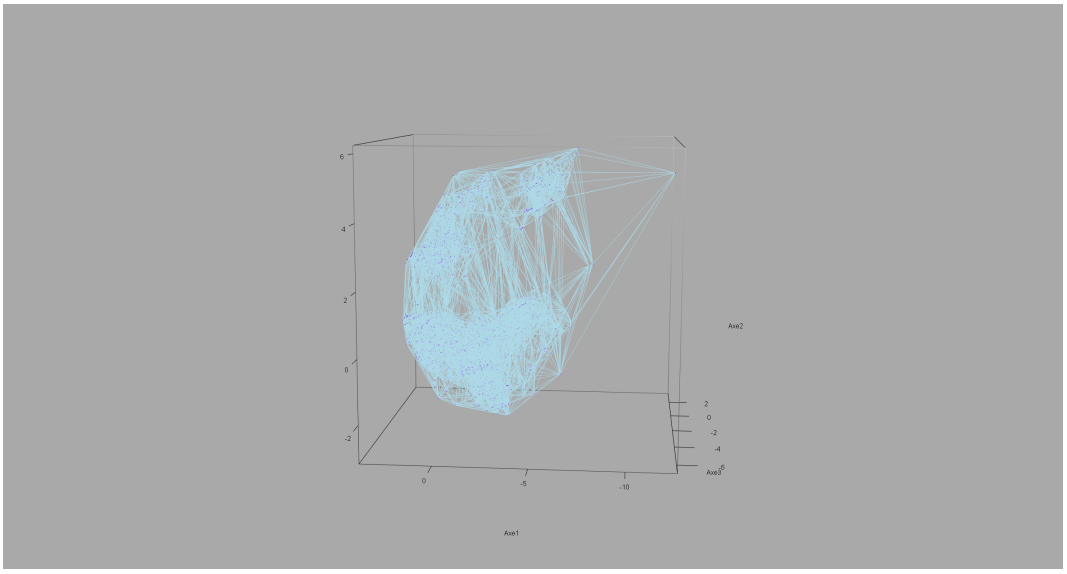


FIGURE 3.28 – Triangulation de Delaunay n 3D

Les connexions sont présentées à la figure 3.28. Les points violets sont la position des véhicules (c'est-à-dire les sommets des tétraèdres) et les lignes bleues sont les arêtes des tétraèdres formées par la triangulation de Delaunay. Comme nous pouvons le faire, l'espace créé est un réseau fermé : il n'y a pas de véhicule isolé.

L'inconvénient majeur de la triangulation de Delaunay dans notre véhiculier est de connecter les véhicules selon la possibilité géométrique uniquement ; donnant ainsi des connexions qui ne semblent pas rationnelles. En effet, comme le montre la figure 3.28, certains véhicules isolés aux extrémités sont connectés au cœur du nuage.

Dès lors une étape de contrôle visuel suivie d'une analyse plus approfondie des voisinages semblent nécessaires.

E - Table d'adjacence et corrections

A partir de la carte des véhicules construite, le voisinage des véhicules a été formalisé par un tableau d'adjacence facilitant ainsi l'analyse et la suppression des liens aberrants.

Par ailleurs, le voisinage de deux véhicules ne signifie pas nécessairement qu'ils seront dans le même groupe ou la même classe à la fin du classement. Par exemple, dans le cas où le modèle Yamaha Xmax et le modèle Honda Forza sont voisins et que ce dernier est sous-représenté, alors l'expérience en fréquence et coût moyen de la Yamaha Xmax sera utilisée pour obtenir des estimations crédibles pour la Honda Forza.

Dans la suite, les voisinages des véhicules vont être étudiés avec précision et les liens aberrants supprimés selon deux critères : un critère de distance des véhicules et un critère d'incohérence des liens.

Critère de distance : Si la distance dans l'espace 3D entre deux véhicules voisins est supérieure à une valeur seuil, le lien est alors supprimé. Les quantiles des longueurs de liens sont résumés dans le tableau 3.29.

0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
5,13255E-05	0,021611361	0,037824369	0,053149001	0,067382448	0,08212085	0,099000819	0,115808566	0,134640493	0,154952083	0,177535134
55%	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**	**100%**	
0,199533595	0,224190622	0,253618158	0,290739638	0,337646456	0,398310639	0,492457576	0,641914664	0,961111003	1,17E+01	

FIGURE 3.29 – Quantiles de la distance des véhicules

Tous les liens supérieurs à $0,96$ ont été rompus afin de supprimer 5% des liens les plus longs.

Les résultats de la correction avec le critère de distance sont les suivants :

- Le nombre maximal de voisin passe de 65 à 50 ;
- 8 véhicules SRA sont isolés, c'est à dire sans voisin. Les points en question sont tracés à la figure 3.30. Nous constatons que les points isolés sont bien éloignés de la partie dense du nuage modélisée en bleu ciel sur les graphes (cf figure 3.30).

Critère de liens aberrants : Les liens jugés injustifiés entre les véhicules vont être supprimés en fonction des caractéristiques véhicule, c'est-à-dire selon le degré de dissemblance entre deux véhicules en termes de variables non véhicule.

Afin d'évaluer le degré de cohérence entre les caractéristiques des véhicules, une analyse précise de ces dernières a été réalisée. Les tableaux croisés de caractéristiques des voisins ont aidé à la prise de décision. Nous avons arrêté notre choix sur les variables cylindrées, poids, prix et puissance réelle pour définir les cohérences des liens [3].

3. Les variables poids, prix, et puissance ont été divisées en 12 quantiles, la cylindrée en 8 quantiles.

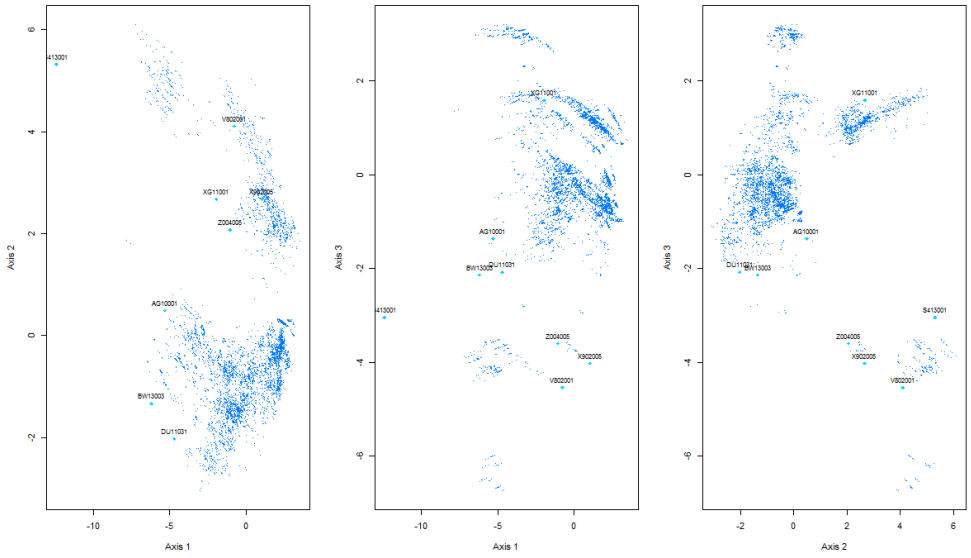


FIGURE 3.30 – Véhicules isolés suite à la correction selon le critère de distance

Les liens des véhicules ayant plus ou moins X variables d'écart avec leurs voisins ont été supprimés. Plusieurs valeurs de X ont été testées et résumées dans le tableau 3.31 ; pour chacune, les voisinages restants et supprimés ont été analysés. Après avoir testé différentes valeurs pour X et analysé pour chacune les voisinages restants et supprimés, la valeur X=2 a été retenue.

Version	Nombre de liens	% de lien non affecté	Nombre de véhicules sans voisin
Sans correction	42140	100%	0
Critère de distance	37906	90%	8
Critère de distance + Critère de liens aberrants x=3	32056	76%	17
Critère de distance + Critère de liens aberrants x=2	28052	67%	27
Critère de distance + Critère de liens aberrants x=1	20400	48%	100

FIGURE 3.31 – Tableau choix de X pour le critère de liens aberrants

Finalement, comme le montrent les figures 3.32, 3.33 et 3.34, ces deux étapes de corrections améliorent nettement la carte des voisins. Suite à des tests manuelles de cohérences, la carte des véhicules construite est considérée comme suffisamment réaliste pour être utilisée pour les techniques de lissage spatial.

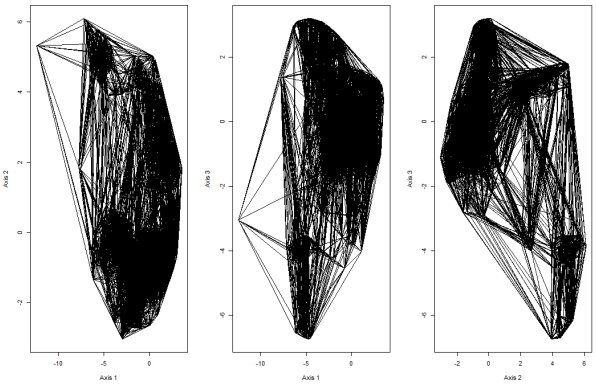


FIGURE 3.32 – Carte des véhicules avant correction

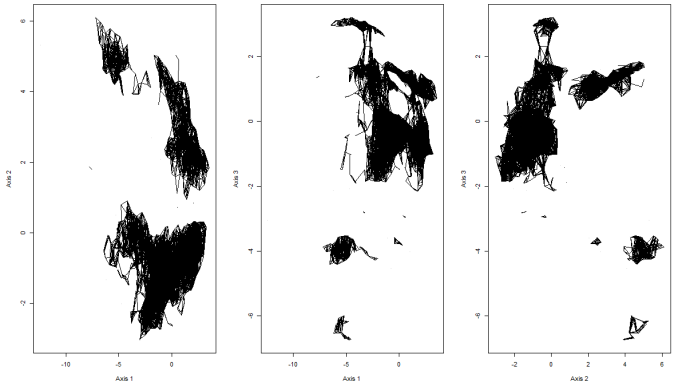


FIGURE 3.33 – Carte des véhicules après correction selon critère de distance

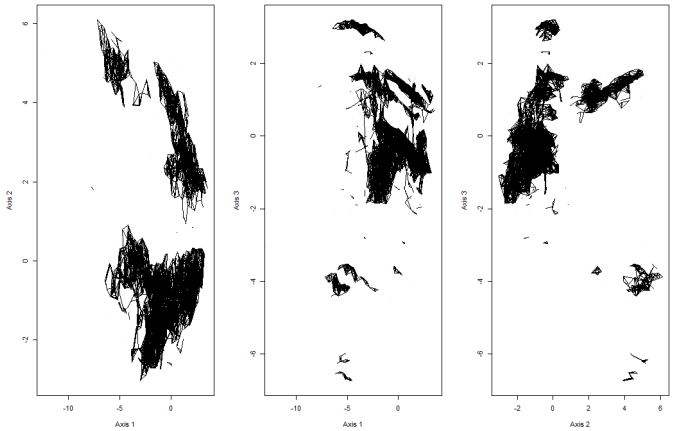


FIGURE 3.34 – Carte des véhicules après correction selon critère de distance et le critère de liens aberrants

3.2.3 Etape 3 - Lissage spatial et conclusions

La carte des véhicules achevée et le random forest détaillé, la dernière étape de cette seconde approche est l'étape de crédibilité qui prend ici la forme d'un lissage spatial.

A - Lissage spatial

La méthode de lissage spatial a été mise en place selon la méthodologie présentée section 1.4.3. Un même paramètre de lissage fixé à 0,5 a été utilisé pour les modèles Random Forest construits. Ainsi, pour chaque véhicule i, $\overline{r_i}$ est la valeur moyenne des prédictions du Random Forest des véhicules voisins calculée en tenant compte de la distance des voisins aux véhicules x et de leur représentativité en termes de nombre d'images pour le modèle de fréquence (respectivement de nombre de sinistres pour le modèle de coût).

FIGURE 3.35 – Modèle de coût - Comparaison de la variable d'intérêt avant et après lissage (à la maille contrat à gauche et à la maille véhicule à droite)

Le facteur de crédibilité de la méthode de lissage V_i (voir section 1.4.3) est sensible au nombre de voisins. Par exemple, un véhicule ayant un nombre d'images correspondant à 100 aura un V_i égale à 99% s'il a un unique voisin associé à 1 image. Dans le cas où, ce même véhicule est entouré de 9 voisins, correspondant respectivement à 100 images, son V_i sera égal à 10%. Nous avons souhaité analyser le comportement des lissages vis-à-vis de la représentativité du portefeuille. Pour cela, une analyse de l'écart entre les valeurs prédites par le Random Forest et la valeur lissée selon le facteur de crédibilité de la méthode de Bühlmann-Straub a été menée ; créant ainsi les graphiques figures 3.35 et 3.36 où l'axe vertical correspond aux facteurs de crédibilité selon la méthode de Bühlmann-Straub par ordre croissant. Nous observons tout

FIGURE 3.36 – Modèle de fréquence - Comparaison de la variable d'intérêt avant et après lissage (à la maille contrat à gauche et à la maille véhicule à droite)

d'abord que le lissage est plus fort (c'est-à-dire l'écart avant et après lissage est plus élevé) pour la maille contrat que pour celle véhicule. Ce phénomène est cohérent avec l'essence même des mailles puisque celle véhicule contient des informations moyennisées.

Ensuite, nous remarquons que pour le modèle de fréquence, le lissage est plus important pour les véhicules de faible crédibilité. Ce constat correspond tout à fait à l'objectif fixé de fiabiliser l'information des véhicules sous-représentés par celle des véhicules bien représentés. De plus, l'allure des graphiques sous-entend une certaine homogénéité de la valeur à expliquer entre les véhicules voisins fortement représentés.

Pour le modèle de coût, indépendamment de la maille, le lissage semble plus fort pour les modèles de forte crédibilité que pour ceux de faible crédibilité. Tout d'abord, un faible écart entre les variables faiblement représentées peut s'interpréter de deux manières. Une première explication pourrait être que l'étape de crédibilité n'est pas nécessaire. Cet argument semble peu convainquant puisque les prédictions du Random Forest à la maille contrat sont très proches de la moyenne des observations des véhicules et que le Random Forest à la maille véhicule ne fait aucune distinction selon le poids des véhicules. Une autre explication pourrait être que les véhicules sous-représentés n'ont pas de voisins ayant une crédibilité suffisante. Cette raison serait cohérente avec le fait que certaines typologies de véhicules sont plus sinistrées que d'autres. En effet plus de 66% des sinistres sont des véhicules dans l'une des quatre catégories suivantes : scooter, custom, roadster ou trail. Ensuite, les véhicules ayant une forte crédibilité subissent un lissage fort. Un tel phénomène n'est possible que si les véhicules crédibles sont voisins avec des véhicules fortement représentés. De plus, l'importance de ce lissage se traduit par une forte hétérogénéité de la variable d'intérêt. Cette disparité peut avoir différentes origines. Elle peut être due à une insuffisance de données qui entrainerait un mauvais apprentissage des forêts. Elle peut également provenir d'une inadéquation de la carte des véhicules qui aurait des voisins discutables. Une autre explication serait que les caractéristiques véhicules utilisées n'expliquent pas ou insuffisamment le coût et ainsi que des facteurs explicatifs soient omis. Cette dernière possibilité rejoint d'ailleurs la non-conformité de la carte.

B - La classification automatique

Il ne reste plus qu'à classer les véhicules selon leur effet véhicule prédit lissé (c'est-à-dire après l'étape de Random Forest et de lissage spatial). La méthode de Ward présentée section 1.4.3 a été mise en place à l'aide de la fonction *hclust* du logiciel R. Le nombre de classes a été défini à l'aide du package R *NbClust* qui fournit 30 indices pour déterminer le nombre de groupe optimal. Les différentes classifications créées sont résumées dans le tableau 3.37 ci-après.

	Maille	Nombres de groupe le plus plus proposé par la fonction NbClust	Nombres de classes mises en place	Nom du modèle associé
$V_{Classe\text{-}new}$	Contrat	9 ou 14	9	rf-cm-mc-al-w9
			14	rf-cm-mc-al-w14
			18	rf-cm-mc-al-w18
	Véhicule	9 ou 15	9	rf-cm-mv-al-w9
			15	rf-cm-mv-al-w15
			18	rf-cm-mv-al-w18
$V_{Groupe\text{-}new}$	Contrat	14	13	rf-freq-mc-al-w13
			14	rf-freq-mc-al-w14
			20	rf-freq-mc-al-w20
	Véhicule	12	12	rf-freq-mv-al-w12
			13	rf-freq-mv-al-w13
			20	rf-freq-mv-al-w20

FIGURE 3.37 – Tableau récapitulatif des classifications construites après lissage spatial

Nous constatons que pour un même modèle donné, la classification des véhicules est peu sensible au nombre de classes. En effet, selon le tableau croisé 3.41 comparant les modèles *rf-freq-al-w13* et *rf-freq-al-w20*, le changement de classification des véhicules est marginal à l'ajout d'une classe supplémentaire : la majorité des véhicules reste dans la même classe. À l'inverse, en comparant des modèles construits à des mailles différentes, la classification des véhicules est très sensible au nombre de classes. En effet, une comparaison de la classification d'un modèle à l'autre met en évidence une forte sensibilité du classement des véhicules comme le montre le tableau croisé 3.41.

Classification Ward	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total général
1	200				115																315
2		81					199														280
3			400																		400
4				33																	33
5					374																374
6								202		200											402
7								317													317
8											48						62				110
9												99			70						169
10													365								365
11														110		53					163
12																		27			27
13																			7	1	8
Total	200	81	400	33	115	374	199	202	317	200	48	99	365	110	70	53	62	27	7	1	2963

Classification Ward	1	2	3	4	5	6	7	8	9	10	11	12	13	Total général
1		1	3	2	5	43	8	22	81	2	89			256
2	5	4	75	2	20	63	100		12	14	4			299
3	1	124		25						2				152
4	179	128	3	1	28					82				421
5	23	3	78		117	9	27			84				341
6	3	4	165	1	105	42	76			27	1			424
7	102	13	33	1	78	3	7			148				385
8	2	2	35		18	146	78	1	24	5	10			321
9		1	6		3	84	17	3	40	1	18			173
10			1	1		9	4	47	7		32	6		107
11			1			3		29	2		5	8	1	49
12								7	3		3	12	5	30
13								1			1	1	2	5
Total	315	280	400	33	374	402	317	110	169	365	163	27	8	2963

FIGURE 3.38 – Modèle de fréquence - Comparaison des classifications à 13 et 20 groupes de la maille contrat à gauche, comparaison des classifications à 13 groupes entre la maille contrat et la maille véhicule à droite

En plus du tableau 3.37, afin de juger par la suite de la pertinence de la carte des véhicules et du lissage spatial, nous avons également construit les classifications équivalentes à partir des prédictions du Random Forest (c'est-à-dire sans l'étape de lissage spatial).

3.3 Analyses comparatives des deux approches réalisées

En combinant les deux approches développées, nous avons créé 19 classifications candidates pour la $V_{Classe-new}$ et 22 classifications candidates pour le $V_{Groupe-new}$. Cette partie se propose de comparer dans un premier temps les différentes classifications construites à dire de machines. Puis ces véhiculiers construits seront comparés à ceux à dire d'experts ce qui nous amènera à nous interroger sur la place des experts.

3.3.1 Détail des véhiculiers obtenus pour chacune des deux approches

Dans une démarche comparative des classifications construites, chaque classification a respectivement été injectée dans le modèle GLM défini à la section 2.3 en remplacement des variables véhicule. Pour comparer ces différentes classifications, il a fallu déterminer un critère : le critère retenu pour cette étude est la racine carrée de l'erreur quadratique moyenne (RMSE). Ensuite un modèle GLM à iso-variables a été construit sur la base test. Ce dernier point va nous permettre de rendre compte de la qualité du véhiculier construit et donc de son pouvoir à s'appliquer à d'autres données. Les tableaux 3.39 et 3.40 résument les résultats obtenus.

Modélisation	Maille	Nom du modèle	Nombre de groupes	RMSE sur base d'apprentissage	RMSE sur base test
Arbres sans crédibilité	Contrat	arbre-cm-mc-c0	4	7500,79	6681,42
		arbre-cm-mc-c0-bis	9	7443,65	6666,84
	Véhicule	arbre-cm-mv-c0	10	7463,83	6669,27
Arbres avec crédibilité	Contrat	arbre-cm-mc-c25	5	7496,61	6665,26
		arbre-cm-mc-c25-bis	18	7448,41	6643,22
	Véhicule	arbre-cm-mv-c25	5	7513,58	6690,42
		arbre-cm-mv-c25-bis	13	7490,56	6687,38
Random forest sans lissage	Contrat	rf-cm-mc-sl-w9	9	7347,69	8818,73
		rf-cm-mc-sl-w14	14	7707,87	8811,47
		rf-cm-mc-sl-w18	18	7746,95	8811,03
	Véhicule	rf-cm-mv-sl-w9	9	7686,04	8809,75
		rf-cm-mv-sl-w15	15	7736,72	8810,49
		rf-cm-mv-sl-18	18	7736,30	8810,71
Random forest avec lissage	Contrat	rf-cm-mc-al-w9	9	7244,50	8550,65
		rf-cm-mc-al-w14	14	7202,40	8552,78
		rf-cm-mc-al-w18	18	7194,23	8553,06
	Véhicule	rf-cm-mv-al-w9	9	7342,19	8553,03
		rf-cm-mv-al-w15	15	7336,48	8553,81
		rf-cm-mv-al-w18	18	7335,59	8554,18

FIGURE 3.39 – Tableau récapitulatif du RMSE des classifications construites candidates pour la $V_{Classe-new}$

Modélisation	Maille	Nom du modèle	Nombre de groupes	RMSE sur base d'apprentissage	RMSE sur base test
Arbres sans crédibilité	Contrat	arbre-freq-mc-c0	9	0,618117	0,710948
	Véhicule	arbre-freq-mv-c0	8	0,618205	0,711036
Arbres avec crédibilité	Contrat	arbre-freq-mc-53	23	0,618101	0,710937
		arbre-freq-mc-60	14	0,618128	0,710981
		arbre-freq-mc-70	22	0,618117	0,710953
	Véhicule	arbre-freq-mv-53	6	0,618111	0,710942
		arbre-freq-mv-53-bis	18	0,618120	0,710959
		arbre-freq-mv-60	14	0,618104	0,710953
		arbre-freq-mv-70	5	0,618131	0,710978
Random forest sans lissage	Contrat	rf-freq-mc-w13	13	0,609256	0,707243
		rf-freq-mc-w14	14	0,609255	0,707243
		rf-freq-mc-w20	20	0,609248	0,707238
	Véhicule	rf-freq-mv-w13	12	0,609301	0,707250
		rf-freq-mv-w14	14	0,609302	0,707251
		rf-freq-mv-w20	20	0,609299	0,707234
Random forest avec lissage	Contrat	rf-freq-mc-w13	13	0,609130	0,704432
		rf-freq-mc-w14	14	0,609129	0,704431
		rf-freq-mc-w20	20	0,609127	0,704430
	Véhicule	rf-freq-mv-w13	12	0,609163	0,704442
		rf-freq-mv-w14	14	0,609166	0,704444
		rf-freq-mv-w20	20	0,609162	0,704439

FIGURE 3.40 – Tableau récapitulatif du RMSE des classifications construites candidates pour la $V_{Groupe-new}$

A - Résultats de la première approche

Analysons pour commencer les classifications issues de la première approche. Tout d'abord, plusieurs remarques peuvent être formulées sur les mailles utilisées.

En absence d'étape de crédibilité, le choix de la maille à retenir semble étroitement lié à la base de données (nombre de lignes, volatilité). En effet, pour la $V_{Classe-new}$ construite à partir d'une base de données avec peu de lignes et une importante hétérogénéité, la maille la plus adéquate paraît être la maille véhicule. Au contraire, pour le $V_{Groupe-new}$ construit à partir d'une base de données aux caractéristiques opposées (nombre important de lignes et hétérogénéité des données relatives) la maille à privilégier se révèle être la maille contrat.

Après l'étape de crédibilité, la maille contrat se démarque. Pour le modèle de coût, ce constat est lisible sans ambiguïté dans le tableau correspondant. L'analyse des résultats est plus délicate quant au modèle de fréquence pour deux raisons. La première raison réside dans la sensibilité des résultats au seuil de crédibilité choisi ; la maille contrat est celle la plus appropriée selon le critère RMSE (RMSE sur base test plus faible) pour le niveau de 0,53 et inversement pour le niveau 0,60. Le second critère se situe dans l'apparenté des arbres construits entre les deux mailles mise en évidence précédemment. La maille contrat est tout de même celle à privilégier pour le modèle de fréquence, car elle utilise une base d'information plus complète et a un RMSE plus petit.

Considérons maintenant, plus en détail, l'étape de crédibilité. Dans l'ensemble, elle diminue significativement l'erreur RMSE. Ce constat est apparent pour le modèle de fréquence, la mise en place de la sélection des véhicules grâce aux modèles de Bühlmann-Straub a l'effet escompté ; à savoir améliorer la classification construite. Il est à noter que l'effet nombre de groupes peut également jouer sur le niveau de RMSE à la maille contrat par rapport à la classification du modèle arbre-freq-mc-c0. Pour le modèle de coût, la crédibilité à la maille véhicule améliore la segmentation, ce qui n'est pas le cas à la maille véhicule. Cependant ce dernier point ne remet pas nécessairement en cause la pertinence du modèle dans la mesure où le sur-apprentissage peut être dû à la baisse significative du nombre de lignes lors de l'exclusion des véhicules « non crédibles ».

B - Résultats de la seconde approche

Nous nous intéressons dans un premier temps **aux résultats du Random Forest** c'est-à-dire avant tout lissage spatial. La comparaison des RMSE révèle un résultat surprenant puisque pour le modèle de coût, l'arbre de décision a un pouvoir prédictif plus fort que celui des Random Forest. Ainsi pour ce modèle, la double randomisation de l'algorithme de BREIMAN et CUTLER est nuisible au pouvoir prédictif. Si un tel résultat est facilement compréhensible à la maille véhicule à cause de la faible quantité de données, il est à première vue plus surprenant à la maille contrat. Cette situation pourrait, cependant, s'expliquer par la forte volatilité des données et par l'inefficacité de certaines variables qui biaiseraient l'apprentissage. Il serait alors intéressant de tester si l'algorithme gradient boosting [4] diminue l'erreur RMSE de la seconde approche.

Pour le modèle de fréquence, les résultats sont plus en accord avec les attentes *a priori* : l'erreur RMSE est globalement plus petite à la maille contrat qu'à la maille véhicule. Nous notons l'exception des classifications de 20 groupes où l'erreur RMSE est inférieure à la maille véhicule par rapport à la maille contrat.

Ensuite, l'étape de lissage sur la carte des véhicules améliore manifestement la classification des véhicules. Après cette étape de lissage, la maille la plus appropriée est visiblement la maille contrat.

Enfin, nous relevons que pour le modèle de coût l'erreur RMSE de chaque maille est croissante avec le nombre de classes sur la base d'apprentissage et décroissante sur la base de test. Parmi les différentes classifications testées, la classification à 9 classes à la maille contrat est la plus appropriée sur la base de test. L'évolution opposée de l'erreur RMSE entre la base d'apprentissage et celle de test peut s'expliquer par un sur-apprentissage de la classification au-delà de 9 classes. Ce résultat peut également provenir d'une carte des véhicules impropre pour le modèle de coût. Pour le modèle de fréquence, l'erreur RMSE est décroissante avec les

4. Le gradient boosting, développé par FRIEDMAN (2002) est un cas particulier de boosting appliqué aux arbres dans lequel chaque nouvel arbre est une version adaptative du précédent cherchant sa performance. Les arbres construits sont ensuite agrégés par une moyenne pondérée des prédictions

nombres de classes testées aussi bien pour la maille contrat que véhicule. Parmi les classifications construites, celle ayant 20 classes à la maille contrat est la plus appropriée.

C- Comparaisons deux approches et interprétations

Quelle est l'approche la plus appropriée à nos données ?

Il est intéressant de noter qu'aucune approche ne s'est révélée unanimement préférable. En effet, la comparaison des erreurs RMSE sur la base de test met en évidence que la première approche, combinant arbre et crédibilité de Bühlmann-Straub, est nettement plus appropriée pour le modèle de coût alors que pour le modèle de fréquence, c'est la seconde approche combinant Random Forest et carte des véhicules qui se révèle, la plus adéquate.

Nous remarquons en outre que, mise à part la remarque sur les résultats non intuitifs du modèle de coût faite précedemment, les Random Forest sont plus appropriés que les arbres même après l'étape de crédibilité.

L'idée de crédibilité dans chaque approche s'avère-t-elle pertinente dans la pratique ?

L'idée de crédibilité amenée par ce mémoire semble porter ses fruits dans les deux approches.

La selection des véhicules par crédibilité dans son ensemble tend à une erreur RMSE après l'étape de crédibilité inférieure à celle des arbres initiaux respectifs. Nous notons un bémol sur la diminution de la quantité des données. La quantité de données à la maille véhicule après crédibilité est faible pour le modèle de coût à la maille véhicule ce qui expliquerait une erreur plus importante par rapport à l'arbre initial.

Le lissage spatial sur la carte des véhicules améliore à l'évidence la segmentation. Les voisinages de la carte paraissent dans l'ensemble satisfaisants. Il est toutefois difficile de déterminer si les performances de la seconde approche pour le modèle de coût sont liées ou non à la carte des véhicules et donc de juger de l'adéquation de la carte. Une analyse des pics aux figures 3.35 met en exergue une limite de cette carte. Le lissage le plus fort du modèle de coût est subi par la routière sportive 600 XJ6 Diversion de Yamaha alors que, tout comme la plupart de ses voisins, ce véhicule a une crédibilité forte de l'ordre de 0,80. Ce même véhicule a un coût prédit de -2010 (prédiction du Random Forest à la maille contrat) alors que la moyenne des coûts prédits de son voisinage est de 130 pour un écart type de 1878. Ainsi, bien que le voisinage de ce véhicule se justifie du point de vue caractéristiques, son coût prédit par le Random Forest est très différent de celui de son voisinage.

Du reste, le seuil de crédibilité pour la sélection des variables est délicat à arbitrer et peut conduire à la construction d'arbres moins performants que l'arbre initial si le seuil est trop élevé.

Quelles conclusions pouvons-nous tirer sur la maille à utiliser ?

Dans l'ensemble, la maille la plus pertinente s'est révélée être la maille contrat donnant d'ailleurs la meilleure classification de chaque approche et chaque modèle.

Le choix final du $V_{Groupe-new}$ et de la $V_{Classe-new}$?

Le véhiculier final sera composé du $V_{Groupe-new}$ dont la classification est celle du modèle rf-freq-mc-w20 et la $V_{Classe-new}$ constituée de la classification obtenue par le modèle arbre-cm-mc-c25-bis.

3.3.2 Véhiculiers construits versus véhiculiers à dire d'experts existants : comparaison de la performance et de la segmentation

A - Quel est le véhiculier le plus performant sur nos données ?

Nous constatons tout d'abord que, selon le critère RMSE, la classification des experts C14 qui ont une connaissance des profils assurés est plus appropriée que la classification SRA pour le modèle de fréquence. Cependant pour le modèle de coût la classification SRA prime (cf figure 3.41). Ce résultat prolonge les précédentes remarques sur le modèle de coût : la forte hétérogénéité des montants de sinistres rend difficile l'apprentissage aussi bien des modèles mathématiques que la classification des experts. L'association SRA qui dispose d'une quantité plus large d'informations parvient à une segmentation du coût des véhicules plus pertinente bien que non spécifique au portefeuille étudié.

	Nom du modèle	Nombre de groupes	RMSE sur base d'apprentissage	RMSE sur base test
C14	V_{Classe}	9	7504	6668
	V_{Groupe}	12	0,618220	0,711054
SRA	$V_{Classe-SRA}$	11	7500	6651
	$V_{Groupe-SRA}$	14	0,619084	0,712170
Véhiculier construit	$V_{Classe-new}$	18	7448	6643
	$V_{Groupe-new}$	20	0,609127	0,704430

FIGURE 3.41 – Tableau récapitulatif de l'erreur RMSE des véhiculiers existants et du véhiculier construit

Le second constat est que, malgré les difficultés *a priori* (toute garantie, multiplicité des modèles de motos, faible sinistralité, et l'étendue des informations dont disposent les experts), **le véhiculier à dire de machines parvient à classer les véhicules de manière à diminuer l'erreur RMSE par rapport au véhiculier à dire d'experts véhiculier C14 et véhiculier SRA**.

Ce résultat, plus qu'encourageant est néanmoins à prendre avec certaines précautions dans la mesure où l'effet véhicule peut être surestimé dans les modèles GLM utilisés en raison du choix des variables candidates. De plus, l'ensemble de ces résultats est évalué selon un unique critère. Ainsi, dans le prolongement de cette l'étude, il pourrait être intéressant de tester le véhiculier construit sur le modèle tarifaire en cours selon plusieurs critères de comparaison. De plus, bien

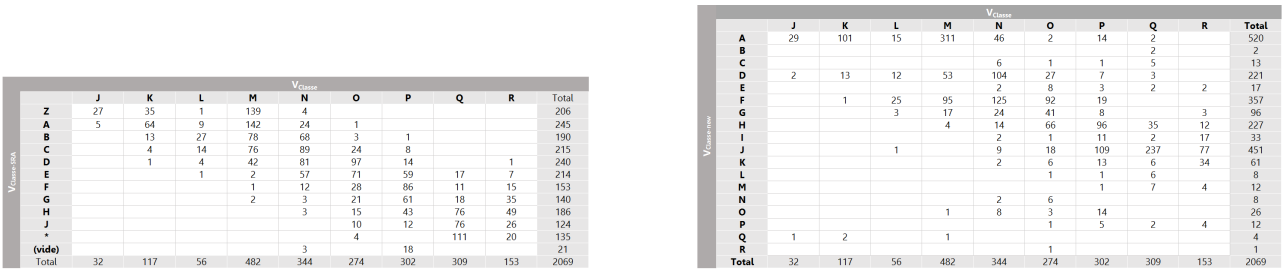
que préciser que les résultats sont intrinsèquement liés aux données utilisées et aux hypothèses retenues, soit en quelque sort tautologique, il convient de tenir compte de l'ensemble de ces éléments dans l'analyse des résultats. Notamment, parmi les hypothèses faites, il conviendrait de vérifier les hypothèses requises pour l'application de la théorie de Bühlmann-Straub.

B - Quel véhiculier est le plus segmentant ?

Nous nous intéressons par cette question à la manière dont les véhicules sont segmentés entre les différents véhiculiers.

Classification pour le modèle de coût :

Une première comparaison des véhiculiers C14 et SRA met en relief une ressemblance frappante entre les classifications pour le modèle de coût (cf figure 3.42).



$V_{Classe-SRA}$ \ V_{Classe}	J	K	L	M	N	O	P	Q	R	Total
Z	27	35	1	139	4					206
A	5	64	9	142	24	1				245
B		13	27	78	68	3	1			190
C		4	14	76	89	24	8			215
D		1	4	42	81	97	14		1	240
E			1	2	57	71	59	17	7	214
F				1	12	28	86	11	15	153
G				2	3	21	61	18	35	140
H					3	15	43	76	49	186
J						10	12	76	26	124
*						4		111	20	135
(vide)					3		18			21
Total	32	117	56	482	344	274	302	309	153	2069

$V_{Classe-new}$ \ V_{Classe}	J	K	L	M	N	O	P	Q	R	Total
A	29	101	15	311	46	2	14	2		520
B								2		2
C					6	1	1	5		13
D	2	13	12	53	104	27	7	3		221
E					2	8	3	2	2	17
F		1	25	95	125	92	19			357
G			3	17	24	41	8		3	96
H				4	14	66	96	35	12	227
I					2	1	11	2	17	33
J			1		9	18	109	237	77	451
K					2	6	13	6	34	61
L						1	1	6		8
M							1	7	4	12
N					2	6				8
O				1	8	3	14			26
P						1	5	2	4	12
Q	1	2		1						4
R						1				1
Total	32	117	56	482	344	274	302	309	153	2069

FIGURE 3.42 – Comparaisons de $V_{Classe-SRA}$ et V_{Classe} (à gauche) et de V_{Classe} et $V_{Classe-new}$ (à droite)

La classification V_{Classe} définie par les experts C14 est donc très fortement liée au coût du véhicule qui pour rappel est l'unique variable utilisée pour la classification $V_{Classe-SRA}$. *A contrario*, les classifications V_{Classe} et $V_{Classe-new}$ sont foncièrement différentes. Ainsi nous constatons que la classe A [5] de $V_{Classe-new}$ est composée de véhicules classés par les experts C14 dans des classes V_{Classe} sur une amplitude important (allant de J à Q). Un tel résultat montre que des véhicules dont le prix moyen est très différent peuvent avoir un coût de sinistres très proche. Ce résultat doit également tenir compte que l'étude est faite toutes garanties confondues : ainsi il est possible, voire probable, que cette similitude de coût ne soit plus valable lors d'une analyse plus fine au niveau de la garantie.

La figure 3.42 montre que les classes sont inégalement réparties en termes de nombre de véhicules. Certaines ne sont composées que de quelques véhicules quand d'autres sont composées de plusieurs centaines de véhicules. La comparaison des coefficients GLM de chaque classe de $V_{Classe-new}$ met en évidence que certaines classes notamment B, Q et R, pourraient être remises en cause dans la mesure où elles contiennent un nombre de véhicules extrêmement faible et que leurs coefficients GLM entre la base d'apprentissage et la base de test sont volatiles.

5. Les classes ont été triées et renommées par ordre croissant en fonction de leur coût moyen prédit.

Ainsi, au vu du RMSE, si la $V_{Classe-new}$ améliore dans l'ensemble la segmentation actuelle et donc la tarification, elle est de toute évidence perfectible. Par exemple, certaines classes nécessiteraient une analyse approfondie afin d'effectuer d'éventuels regroupements de véhicules.

Nous avons analysé les classifications des scooters et contre toute attente la classification construite ne permet pas de créer des groupes de scooters plus homogènes en termes de taille de classe. Ainsi, aussi bien dans la classification à dire d'experts que dans la classification construite par des modèles d'apprentissage statistique, la majorité des scooters sont regroupés dans une même classe. L'utilisation de la même méthodologie sur les seuls scooters serait un moyen d'améliorer la segmentation de cette catégorie. Notre étude laisse donc à penser qu'une partie des scooters ont une charge moyenne des sinistres des coûts proches en toutes garanties.

Classification pour le modèle de fréquence

Nous constatons tout d'abord qu'à l'inverse des conclusions du modèle du coût, les classifications V_{Groupe} et $V_{Groupe-SRA}$ sont relativement dissemblables pour le modèle de fréquence (cf figures 3.43)

$V_{Groupe-SRA}$ \ V_{Groupe}	5	6	30	31	32	33	34	35	36	37	38	Total
4	5	1	35	44	401	28	24	5	1	1		545
5			4	2	67	109	5	3	1	1		192
7					4	22	34	3	2	1	1	67
8					7	6	53	12	5	1	1	85
10							4	2	1			7
Total	5	1	39	46	479	165	120	25	10	4	2	896

$V_{Groupe-new}$ \ V_{Groupe}	5	6	30	31	32	33	34	35	36	37	38	Total
1			2	3	22	4	1					32
2				1	12	5	2					20
3	1		7	5	55	12	2	1				83
4					3							3
5					18	2						20
6			1	5	39	15	4				1	65
7				1	13	32						46
8	1		4	4	46	3	4	3				65
9		1	3	4	45	14	4					71
10	1		8	7	56	16	20	4		1		113
11			3		13	5	8	2	1			32
12			2	2	36	11	12	3	1	1		68
13			2	6	44	4	6	1				63
14				3	17	21	21	1	3	1		67
15			6	3	21	8	10	4	2			54
16	1		1	1	19	4	8	1	1	1		37
17	1			1	11	9	11	2			1	36
18					7		6	2	2			17
19					2		1	1				4
Total	5	1	39	46	479	165	120	25	10	4	2	896

FIGURE 3.43 – Comparaisons de $V_{Classe-SRA}$ et V_{Classe} pour la catégorie des scooters (à gauche) et de V_{Classe} et $V_{Classe-new}$ pour la catégorie des scooters (à droite)

Dès lors, nous pouvons en déduire que la classification V_{Groupe} intègre de toute évidence d'autres éléments que le ratio puissance/poids et la catégorie qui sont pour rappel les variables utilisées pour la classification $V_{Groupe-SRA}$. La classification $V_{Groupe-new}$ est également éloignée de V_{Groupe}.

D'autre part pour le modèle de fréquence, la seconde approche s'est révélée plus performante que la première. Faut-il en déduire que la classification pour le modèle de fréquence est plus performante en ne forçant pas un même classement pour deux véhicules aux caractéristiques similaires ? Autrement dit, la classification serait-elle plus pertinente en classant les véhicules après crédibilité uniquement par rapport à leur valeur de crédibilité et non plus par rapport à leur similarité ? Il est difficile de conclure sur cette question au vu des informations à notre disposition, mais il serait intéressant de mener une étude comparative dans ce sens.

Par ailleurs, si nous nous attardons sur le cas des scooters, la nouvelle classification répartit de façon plus homogène les véhicules dans les différents groupes par rapport à la classification à dire d'experts C14. Toutefois, certains groupes de $V_{Groupe-SRA}$ nécessiteraient une analyse approfondie pour éventuellement rapprocher certains groupes contenant peu de véhicules.

3.3.3 Quelle est la place de l'avis d'experts

Cette étude a permis la construction d'un véhiculier améliorant la tarification et donc dans ce sens plus performant que le véhiculier à dire d'experts. Ainsi les informations disponibles grâce aux données SRA et SIV ont permis de construire des modèles mathématiques tirant des informations sur les véhicules et leur sinistralité qui se sont avérés selon le RMSE sur la base de test plutôt fiables, et ce malgré les contraintes fortes au départ. Cependant, ce résultat ne remet aucunement en cause le rôle central des experts du deux roues vis-à-vis du véhiculier moto. S'il parait difficile pour des experts de refondre totalement un véhiculier, c'est-à-dire de reclasser l'intégralité des véhicules les uns par rapport aux autres, il semble opportun voire indispensable de concilier l'avis d'experts aux modèles mathématiques construits pour bâtir un véhiculier toujours plus compétitif. Ces interactions entre experts et modèles pourraient alors prendre différentes formes.

Premièrement, les résultats des classes construites mettent en évidence que certaines classes contenant peu de véhicules nécessitent d'être analysées en détail afin d'effectuer d'éventuels rapprochements. Si la taille de la base de données de sinistres s'est révélée insuffisante, entrainant des difficultés d'apprentissage des méthodes de Machine Learning, les experts ont une connaissance des véhicules qui leur permet de trancher sur ces classes atypiques. Il est bien évidemment possible d'avoir des éléments de réponse en comparant véhicule par véhicule la base de test et celle d'apprentissage. Néanmoins, l'avis d'experts s'impose pour les véhicules absents de la base de test, notamment pour la base de sinistres où les cas de véhicules peu représentés n'ayant pas de sinistres dans la base test ne sont pas anecdotiques.

Deuxièmement, si une limite de la carte des véhicules a déjà été mise en exergue à la section 3.2.3, une autre limite peut être émise ici. La carte des véhicules telle que construite ne tient compte que de certaines caractéristiques, omettant ainsi un ensemble d'informations, notamment le duo-conducteur, mais également des éléments plus facilement modélisables tels que l'allure du véhicule. En effet, les variables catégorie et marque sont les seules à fournir de l'information sur la structure visuelle du véhicule, mais il va sans dire qu'elles sont insuffisantes pour distinguer la carrure des véhicules dans une même catégorie. Dès lors, la carte ne tient pas compte de la similarité esthétique des véhicules. Une telle information pourrait pourtant être une indication riche quant au profil du risque du véhicule et notamment être un indicateur de la cible marketing de certains véhicules. Cet examen a mis en évidence un phénomène doublement intéressant. Tout d'abord, nous avons montré que deux véhicules "crédibles" ayant des caractéristiques très similaires et donc définis comme voisins sur la carte des véhicules, peuvent avoir des fréquences de sinistres résiduelles extrêmement différentes et peuvent être visuellement assez dissemblables. C'est notamment le cas des deux véhicules présentés aux figures 3.44 et 3.45. Ce constat met tout d'abord en lumière les limites de la carte des véhicules et émet la possibilité d'affiner la suppression des liens aberrants en éliminant tous liens entre

les différentes catégories. Plusieurs axes d'amélioration peuvent alors être proposés. L'expert pourrait intervenir dans la définition des liens aberrants et dans le choix des variables que ce soit en input ou pour corriger des liens. Il pourrait également intervenir dans la sélection des variables en amont. L'ajout de la sous-catégorie rétro pourrait, par exemple, être envisagé pour mieux prendre en compte l'allure du véhicule.

FIGURE 3.44 – La routière basique Bonneville T100 BLACK de Triumph

FIGURE 3.45 – Le scooter GP 800 de Gilera

Ensuite, le modèle de fréquence tel que construit ne formule aucune règle pour introduire les futurs véhicules à venir. Des règles proxys peuvent être définies en appliquant un arbre de décision sur la classification $V_{Groupe-new}$ construite. Cependant, un tel modèle aurait nécessairement un taux d'erreur et l'interaction captée par le Random Forest pourrait également ne pas être vérifié sur les nouveaux véhicules.

Enfin, si le modèle de coût construit permet à l'aide des règles de décisions explicites de classer les véhicules, l'avis d'experts semble irremplaçable pour classer les véhicules atypiques et pour raffiner la classification mathématique.

Conclusion

Après une confrontation entre l'homme et la machine sur le terrain de la tarification moto, le score ne peut être aussi explicite que celui au jeu Go. Dans ce contexte, l'homme expert et la machine gagneraient plutôt à joindre leurs efforts pour le plus grand gain de l'assureur.

En effet, la construction d'un véhiculier se basant uniquement sur des modèles mathématiques a montré de nombreux avantages selon le RMSE mais également des insuffisances, qui pourraient vraisemblablement être comblées par l'intervention ponctuelle d'un expert.

Dans le cadre de ce mémoire, nous proposions de comparer les capacités des véhiculiers à dire de machines à ceux à dire d'experts. Pour cela, deux approches originales de véhiculiers à dire de machines ont été développées en enrichissant les travaux existants et leurs limites. L'outil central de ces deux approches repose sur une méthode de Machine Learning à laquelle a été greffée une étape de crédibilité visant améliorer ses prédictions. La première approche de l'étude est basée sur un algorithme CART (Classification and Regression Tree) où une approche singulière de la théorie de la crédibilité a également été introduite afin d'améliorer la classification des véhicules tout en conservant le pouvoir prédictif de l'arbre de décision. La seconde approche a été développée à partir d'un modèle de Random Foret auquel a été aggregé un lissage spatial des prédictions en détournant le concept de carte des véhicules développé par Axa Global P&C.

L'étude a montré que l'étape de crédibilité suggerée dans cette étude améliore la classification des véhicules. Ainsi, d'un côté, l'omission des véhicules les moins représentés dans le portefeuille permet un meilleur apprentissage des régressions CART. De l'autre, pour des véhicules aux caractéristiques proches, l'utilisation des observations des comportements des véhicules très présents dans le portefeuille a permis d'obtenir de l'information pertinente pour les véhicules sous-représentés.

Par ailleurs, la comparaison de différentes mailles a mis en évidence la sensibilité du véhiculier et de ses performances au choix de la maille utilisée. De manière générale, pour la régression CART, la maille contrat s'est révélée plus adaptée que la maille véhicule, et ce, même pour le modèle de coût construit à partir d'une base de données de petite taille caractérisée par une forte volatilité entre les observations.

Nous notons également un résultat intéressant qui s'est révélé suite à l'application de l'approche de crédibilité : sur la base de données pour le modèle de fréquence, la sélection des observations selon leur fiabilité a généré des régressions CART à la maille contrat et la maille

véhicule quasiment identiques. Ceci laisse supposer que la classification des véhicules à partir d'informations crédibles en nombre suffisant est peu sensible au choix de la maille utilisée.

L'ensemble de ces méthodes et résultats a eu pour objectif de départager l'Homme et la Machine dans leur duel. Avec un premier avantage pour l'algorithme d'apprentissage qui a permis de réduire la part de variance expliquée par rapport à la classification à dire d'experts. Les scores se ressèrent cependant en constatant que le véhiculier a également fait apparaitre de nombreuses insuffisances au niveau de la précision de la classification. Ces insuffisances pourraient notamment être comblées en intégrant l'avis d'experts dans la modélisation.

D'autres élements de modélisation pourraient être envisagés afin de pallier certaines limites de cette étude. La carte des véhicules qui s'est révélée pertinente pour le modèle de fréquence ne parvient pas, compte tenu des données disponibles, à considérer l'allure/structure du véhicule dans la constitution des véhicules voisins. En complément ou remplacement de l'avis d'experts, ce dernier point pourrait être solutionné en utilisant les avancements de ces dernières années dans le domaine de la reconnaissance d'image. En effet, les méthodes de Machine Learning sont aujourd'hui capables, dans d'autres domaines, de comparer et d'identifier des images. Ainsi, la récupération de photos de véhicule par web-scraping permettrait de créer un score de ressemblance entre deux images de véhicules qui pourrait alors être incorporé pour perfectionner la carte des véhicules.

De manière plus large, le recours à la reconnaissance d'image offre vraisemblablement de nouvelles pistes dans l'implémentation de véhiculiers à dire de machines.

Bibliographie

[1] TUFFERY S. *Data mining et statistique decisionnelle.* Editions TECHNIP, 2010.

[2] CHARPENTIER A. & DENUIT M. *Mathematiques de l'Assurance Non-Vie.* Paris : Economica, 2004.

[3] SIPULSKYTE R. Development of a motor vehicle classication scheme for a new zealand base. *New Zealand Society of Actuaries Conference*, 2012.

[4] AXA GLOBAL P&C. Methodology of vehicule classification. 2014.

[5] AFAILAL H. Modelisation de la valeur contrat pour le pilotage du portefeuille moto, memoire club 14 / axa france. *Memoire DAUPHINE*, 2014.

[6] VIDAL I. Prediction de l acte de resiliation et tarification de la garantie rc des jeunes conducteurs. *Memoire ISFA*, 2015.

[7] PAGLIA A. Tarification des risques en assurance non-vie, une approche par modele d apprentissage statistique. *Memoire EURIA*, 2011.

[8] KUSNIK V. Tarification en assurance automobile de particulier. *Memoire ISUP*, 1981.

[9] GEY S. Bornes de risque, detection de ruptures, boosting : trois themes statistiques autour de cart en regression. *These Paris 11, Orsay*, 2002.

[10] GENUER R. Forets aleatoires : aspects theoriques, selection de variables et applications. *These Universite Paris Sud XI*, 2010.

[11] MANO C. & RASA E. Use of classification analysis for grouping multi-levels rating factors. 2006.

[12] WENZEL T.P. & ROSS M. The effects of vehicle model and drivers behivior on risk. *Accident Analysis and Pretension, Vol. 37*, 2004.

[13] OHLSSON E. Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, 2008.

[14] BROCKMAN M.HL & WRIGHT T.S. Statistical motor rating : Making effective use of your data. *Journal of the Institute of Actuaries 199*, 1992.

[15] KIM & al. Factors associated with automobile accidents and survival. *Accident Analysis and Prevention*, 2006.

[16] LAWRENCE B. Significant developments in motor insurance in thailand. *Risk Management and Insurance Review, Vol. 4, No. 1*, 2001.

[17] ROOSVELT C. Estimating claim settlement values using glm. *Casualty Actuarial Society*, 2004.

[18] WANG & al. Modelling different types of bundled automobile insurance choice behaviour : the case of taiwan. *The Geneva Papers on risk and Insurance â“ Issues and Practice, Vol. 35*, 2010.

[19] YEO & al. Clustering technique for risk classification and predictions of claim costs in automobile insurance industry. *International Journal of Intelligent Systems in Accounting, Finance and Management, Vol. 10*, 2001.

[20] BREIMAN L. & CUTLER A. Random forrest methods. 2001.

Table des figures

Rappels théoriques

A - Triangulation de Delaunay

Définition 2 (*Triangulation*)

Soit $P = \{p1, \ldots., pn\}$ un ensemble de points placés dans l'espace euclidien à deux dimensions. La triangulation est définie comme un réseau interconnecté de points dans lequel aucune nouvelle connexion en ligne droite entre deux points ne peut être faite sans couper / perturber les lignes existantes.

Il existe dans la littérature plusieurs méthodes de triangulation. À la figure 46 ci-dessous, les points sont à la même position mais la triangulation de gauche (a) et celle de droite (b) sont différentes. Quelle triangulation doit-on privilégier ?

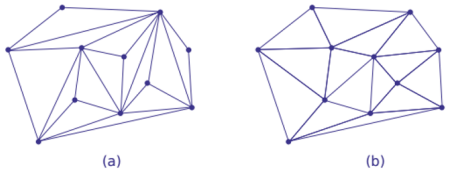


FIGURE 46 – Choix de la triangulation

Définition 3 (*Triangulation de Delaunay*)

La triangulation de Delaunay est une triangulation $DT(P)$ telle qu'aucun point de P ne soit à l'intérieur d'un triangle de $DT(P)$. Cette définition est équivalente à dire qu'une triangulation est de Delaunay si aucun point de P n'est à l'intérieur du cercle circonscrit d'un des triangles de $DT(P)$.

Généralisation en dimension supérieure

Pour un ensemble de points P dans l'espace euclidien de dimension n, la triangulation DT(P) est de Delaunay si aucune hypersphère circonscrite du simplexe de DT(P) ne contient de point de P .

En dimension deux, une façon de détecter si un point D appartient au cercle circonscrit ABC où les points ABC sont positionnés dans le sens inverse des aiguilles est de calculer le déterminant

ci-dessous. Ce déterminant est positif si et seulement si le point D se trouve à l'intérieur du cercle circonscrit.

$$\begin{vmatrix} A_x & A_{x^2}+A_{y^2} & 1 \\ B_x & B_{x^2}+B_{y^2} & 1 \\ C_x & C_{x^2}+C_{y^2} & 1 \\ D_x & D_{x^2}+D_{y^2} & 1 \end{vmatrix} = \begin{vmatrix} A_x - D_x & (A_{x^2}-D_{x^2})\,(A_{y^2}-D_{y^2}) \\ B_x - D_x & (B_{x^2}-D_{x^2})\,(B_{y^2}-D_{y^2}) \\ C_x - D_x & (C_{x^2}-D_{x^2})\,(C_{y^2}-D_{y^2}) \end{vmatrix} > 0$$

Les avantages de la triangulation de Delaunay sont :

- Méthode largement utilisée et implémentée sous de nombreux logiciels, notamment R.
- Méthode finie et résultante d'une triangulation unique sous réserve que les points ne soient pas alignés (pour plus de 4 points).
- Cette méthode maximise l'angle minimum de tous les angles de tous les triangles des triangulations. Les triangles sont ainsi aussi équiangulaires que possible, réduisant dès lors les potentiels problèmes de précision numérique dus à des triangles avec de longues arêtes.

B - Modèle de Bühlmann-Straub

Définition 4 (*Modèle de Bühlmann-Straub*)

Le modèle de Bühlmann-Straub est une généralisation du modèle de Bühlmann intégrant une pondération.

Les notations :

Soient

- *les variables aléatoires modèles de véhicules θ_i ;*
- *X_{ij} l'image j associée au véhicule i ;*
- *w_{ij} le nombre d'années d'exposition associé à l'image X_{ij}.*

Les hypothèses :

- *Les couples (θ_i, X_i) sont indépendants ;*
- *Pour tout j, $E(X|\theta_i) = \mu(\theta_i)$ et $Var(X_{ij})w_{ij} = \sigma^2\theta_i$;*
- *Les X_i sont des variables aléatoires indépendantes et identiquement distribuées.*

Sous ces hypothèses, le modèle s'écrit :

$$\widehat{\mu(\theta_i)} = Z_i X_i + (1 - Z_i)\mu_0$$

où

$$\mu_0 = \sum_{i=1}^{I} \frac{Z_i}{Z_\bullet} X_i$$

$$Z_i = \frac{w_{i\bullet}}{w_{i\bullet} + \frac{\sigma^2}{\eta^2}}$$

$$w_\bullet = \sum_{i=1}^{I} Z_i$$

Les paramètres de structure :

$$\sigma^2 = \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij} \left(X_{ij} - X_i\right)^2$$

$$\eta^2 = \frac{w_{\bullet\bullet}}{w_{\bullet\bullet}^2} \left\{ \sum_{i=1}^{I} w_{i\bullet} \left(X_i - \overline{X}\right)^2 - (I-1)\,\sigma^2 \right\}$$

Compléments sur les méthodes de classifications des véhicules

A - R. SIPULSKYTE

Dans le cadre de son mémoire de master à l'université Aarhus au Danemark, **R. SIPULSKYTE** a développé en 2012, un système de classification de véhicules à moteur pour une compagnie d'assurances basée en Nouvelle-Zélande à l'aide des deux modèles mathématiques, un modèle permettant d'isoler l'"effet conducteur/géographique", un second expliquant l'"effet véhicule". La première étape utilise les GLM, et le second, un modèle de Machine Learning. Le but de l'étude, assez proche de celui de la présente étude, était de classer les véhicules en groupe sur la base à la fois de leurs caractéristiques techniques et du coût supporté par l'assureur.

Cette étude a été réalisée pour le produit auto d'un assureur néo-zélandais. Il est à noter que contrairement à de nombreux pays, l'assurance automobile n'est pas obligatoire en Nouvelle-Zélande. De plus, les données utilisées pour cette étude sont celles d'une assurance tous risques ("Comprenhensive insurance") couvrant uniquement les dommages ou pertes liés au véhicule et la responsabilité civile pour les dommages aux véhicules d'autrui.

Les hypothèses de l'étude

Les hypothèses et les choix faits dans l'étude sont les suivants :

- Périmètre de l'étude : la base de données utilisée contient les contrats et sinistres automobiles de 2007 à 2011 (1424076 assurés, 33640 véhicules automobiles distincts, dont 12583 accidents au cours de la période analysée sinistre) hors des sinistres graves. Seuls les modèles de véhicules ayant au moins un sinistre dans la base de données d'étude ont été gardés, soit 58% de la base de données initiale.
- Modélisation GLM : la modélisation du coût moyen est réalisée selon un modèle fréquence x coût en distinguant pour chacun, les dommages faisant suite à une collision des autres dommages et en excluant les variables concernant le véhicule autre que la somme assurée et l'âge du véhicule. Les quatre modèles GLM sont ensuite combinés afin d'obtenir un unique modèle de prime pure à l'aide du logiciel Emblem.
- Extraction de l'effet véhicule : La valeur étudiée est définie comme la partie non expliquée par l'effet non véhicule (ie. conducteur et géographie). Elle inclut donc le bruit du GLM. Le ratio du coût moyen estimé sur le produit des coefficients est calculé pour

chaque image. L'effet véhicule est alors le coût relatif obtenu.

- Arbre de décision : Le critère d'arrêt de l'arbre CART est le suivant : chaque noeud final doit contenir au moins 125 observations, soit 125 modèles de véhicules différents.
- Fiabilisation des groupes : Un sous-groupe peut être déplacé vers le groupe ayant le coût moyen le plus proche s'il est considéré comme fiable (ie. avec un facteur de crédibilité supérieur à 0,55) et si son coût estimé (coût lié à l'"effet véhicule") est 5% inférieur ou supérieur à celui de son groupe, le sous-groupe.

Les résultats et limites

Après l'étape de fiabilisation, 18 groupes sont obtenus. L'étape de validation de la méthode, consistant à analyser les coefficients GLM de la nouvelle variable de classement, met en avant le sur apprentissage de classification construite, comme l'indique la figure 47. Autrement dit, la classification construite s'ajuste sur les données au dépend du pouvoir de généralisation.

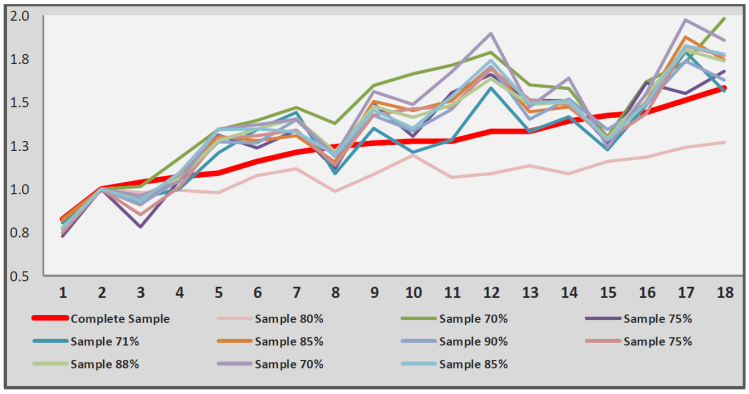


FIGURE 47 – Validation de la classification des véhicules de l'étude de R. SIPULSKYTE (source R. SIPULSKYTE)

Plusieurs explications possibles sont proposées dans l'étude. Tout d'abord, l'introduction de trop nombreuses variables dans le modèle GLM initial peut être à l'origine du sur-apprentissage observé. L'overfitting selon l'étude pourrait également être lié au bruit introduit dans l'analyse.

L'étude conclut que la classification construite est meilleure que celle construite à l'aide de méthode de classification automatique. La justification se base sur l'écart entre les groupes aux extrêmes (groupe ayant le plus petit coût vs celui ayant le plus élevé) passe de 48,78% dans le nouveau classement contre 19% dans le classement à refondre. Cependant, l'étude ne précise pas qu'il est possible que cette amélioration provienne de l'augmentation du nombre (5 classes supplémentaires dans la nouvelle classification). Ensuite, la pertinence de certains groupes est discutable dans la mesure où les coefficients sont très proches comme l'indique la figure 47. Il pourrait être intéressant de calculer la variance inter et intra des classes et de comparer leur coût moyen afin de déterminer le nombre de classes optimal.

Le pouvoir explicatif des classes par l'arbre de décision semble être statistiquement très faible avec un $R^2 = 0.03$ uniquement.

Différentes remarques peuvent être formulées à propos de ces travaux. Il est intéressant de noter que la crédibilité mise en place a considérablement amélioré les groupes finaux, la différence entre le groupe le plus élevé et le moins élevé passe de 48% à 91%. Cependant, elle modifie la classification de véhicules formée par l'arbre perdant ainsi l'intérêt principal des arbres de décision. En effet, après la mise en place de la crédibilité, il n'est plus possible de prédire la classe des modèles de véhicule en suivant les règles de décisions de l'arbre. Une alternative pour contourner ce problème pourrait être la mise en place d'un modèle de crédibilité modifiant les règles de décisions de l'arbre en introduisant des véhicules crédibles.

Par ailleurs, on note que le critère d'arrêt choisi ne permet pas d'isoler les modèles atypiques. Un tel choix de modèle semble devoir être justifié en analysant l'absence de modèles atypiques dans l'étude. Dans le cas présent, l'hétérogénéité entre les véhicules semble trop importante pour envisager ce critère.

B - Axa Global P&C

Cette section présente et analyse la méthode de classification des véhicules développée par AXA Global P&C dans l'objectif d'une possible implémentation à l'échelle du groupe au niveau international. Le document étudié recense l'étape de la mise en place de cette méthode dans le désir de créer une variable de classement des véhicules expliquant la variance résiduelle.

Les notations :

- Observé : il s'agit de la Fréquence réellement observée ;
- Ajusté : il s'agit de la Fréquence modélisée par le modèle incluant les variables internes et externes ;
- Observé standardisé : il s'agit de la Fréquence observée, neutralisée de l'effet des variables non-véhicule ;
- Ajusté standardisé : il s'agit de la Fréquence modélisée, neutralisée de l'effet des variables non-véhicule ;
- Résidu : il s'agit du rapport suivant :

$$\text{Résidu pour un modèle de véhicule "A"} = \frac{\sum_{j \in A} \text{Observé}j}{\sum_{j \in A} \text{Ajusté}j}$$

- Bruit : part du résidu lié à des variations aléatoires entre les observations ;
- Signal : part du résidu que le modèle n'arrive pas à expliquer, mais qui n'est pas aléatoire entre les observations.

La mise en place de la méthode, les hypothèses faites

Les principales hypothèses et les choix faits dans l'étude sont les suivants :

- Périmètre de l'étude : le produit automobile d'AXA France sur l'année 2011 à 2012, pour les dommages accidentels uniquement. La base de données contient 3,6 millions

d'observations, ce qui correspond à 77013 uniques modèles de véhicules et 36 variables concernant les caractéristiques véhicule provenant des bases de données SRA.

- Projections des véhicules : La carte de véhicule est construite, en projetant à l'aide d'une AFDM, les modèles de véhicules dans un espace en 3 dimensions. L'AFDM repose sur l'idée de créer une variable fictive pour chaque variable qualitative en convertissant la variable qualitative en « 0/p » valeurs où le « p » est calculé à partir de la fréquence d'une catégorie concernée.
- Analyse et correction des voisins : les 5% plus longs liens et les liens ayant plus ou moins 3 groupes d'écarts par rapport au classement SRA $V_{Groupe-SRA}$ ou $V_{Classe-SRA}$ ont été supprimés.
- Extractions de l'ajusté standardisé et du résidu : Les extractions des différentes notations présentées précédemment est réalisée à la maille modèle de véhicules à l'aide du logiciel Classifier.
- Lissage et corrections spatiales : Le résidu est lissé à l'aide une méthode de lissage spatiale selon le voisinage de la carte des voisins basé sur une méthode bayésienne. Le degré de lissage dépend de l'exposition ou du nombre de sinistres de chaque modèle de véhicule. Une méthode similaire est appliquée pour corriger l'ajusté standardisé.
- Partitionnement des véhicules : Les modèles de véhicules sont ensuite partitionnés. La méthode de "poids égaux" est la seule méthode testée et celle choisie. Cette méthode consiste à trier les modèles de véhicules par ordre croissant selon leur valeur lissée, puis de former des groupes d'exposition équivalente.

Les résultats et limites

Les informations dans le document étudié sont insuffisantes pour émettre un avis sur les résultats de l'étude. Par exemple, aucune information n'est donnée sur la qualité des données, les retraitements effectués, les discrétisations réalisées pour le modèle GLM, les variables sélectionnées dans le modèle GLM, l'étude des corrélations de celles-ci ou encore des lois choisies. L'étude conclut que les résultats sont relativement satisfaisants dans le cas de la fréquence dommages matérielles et plus mitigés pour d'autres garanties.

Les variables véhicule utilisées pour la carte des véhicules sont les mêmes que celles candidates dans le modèle GLM. L'hypothèse forte faite ici est que la carte des véhicules construite réussira à capter une information supplémentaire non captée par le modèle GLM. Il y a risque non négligeable qu'aucun signal ne soit capté par la carte. De plus, la carte de véhicule construite dans un espace en 3 dimensions ne représente que 16,28% de l'information de la base initiale des véhicules. Une analyse de la signification des axes et de la pertinence de la carte semble indispensable.

L'un des principaux avantages de la carte des véhicules, après celui de fiabiliser les valeurs par modèles de véhicules, est l'intégration de l'ensemble des véhicules recensés par SRA et non uniquement ceux dans le portefeuille. Cela va permettre de classer les véhicules absents de la base d'étude. Cependant, la pérennité du modèle n'est pas assurée pour autant. L'une

des limites de cette méthode est qu'elle ne propose aucune manière de classer les nouveaux modèles de véhicules. D'autant plus que l'intégration de nouveaux véhicules dans la carte de véhicule n'est pas une option envisagée dans la mesure où la triangulation en serait modifiée. Une solution envisageable serait de créer des règles proxy expliquant les groupes formés afin de classer les nouveaux véhicules, qui pourrait par exemple se faire à l'aide d'arbres de décision.

Par ailleurs, la méthode choisie pour le partitionnement des véhicules favorise l'homogénéité des classes vis-à-vis du nombre de véhicule par groupe au dépend de l'homogénéité des variables véhicule au sein de chaque groupe. L'utilisation de la méthode Ward permet de former des groupes de manière à minimiser la variance intra-groupe.

Compléments de la base de données

Les bases de données utilisées sont les suivantes :

- **Données internes sur les assurés** : la base des assurés regroupe les caractéristiques des assurés en contrat image [6]. Chaque ligne correspond à un risque homogène d'un contrat donné. Une année police correspond à la durée d'exposition de l'image. Ainsi, le nombre d'années polices [7], construit sur une année glissante, prend ses valeurs entre 0 et 1. Afin d'obtenir le niveau de granularité souhaité notamment en termes d'information sur le véhicule (marque, modèle), la base assurés a été construite en rapprochant deux bases : une base RT contenant des tables en contrat image et une base ACN contenant des tables contrats.
- **Données internes sur les sinistres** : la base sinistres répertorie l'ensemble des sinistres survenus par année. À chaque mouvement du sinistre, une nouvelle image est créée à une date d'observation unique dans la table concernée. Pour obtenir la base sinistres d'une année N, il est nécessaire de fusionner et rendre compatible 5 tables sinistres à l'aide de la variable numéro de sinistres.
- **Données externes sur les véhicules - base SIV** : remplaçant le Fichier National des Immatriculation (FNI) depuis 2009, le Système d'Immatriculation des Véhicules (SIV) a pour objet la gestion des pièces et opérations administratives liées à la circulation des véhicules sur les voies publiques. La Base SIV contient les informations renseignées lors de l'immatriculation de chaque véhicule. Au moment de l'étude, la seule base d'immatriculation à notre disposition est la base de mai 2015 qui recense l'ensemble des véhicules en circulation à cette date. Les données disponibles dans cette base sont présentées figure 48.
- **Données externes sur les véhicules - base SRA** : Une des principales missions de l'association SRA est de fournir à ses adhérents des informations sur les caractéristiques techniques et commerciales des véhicules par l'intermédiaire de base de données. Les caractéristiques techniques des véhicules données par SRA sont présentées figure 49.

6. Une image est une ligne de la base de données correspondant à un contrat de profil de risque constant. Lorsqu'un assuré change par exemple d'adresse, l'exposition est modifiée et une nouvelle image est créée.

7.
$$\text{Nombre d'années polices} = \frac{\text{date de début d'image} - \text{date de fin d'image}}{\text{Nombre de jours de l'année}}$$

Carrosserie	Énergie	N° immatriculation
Carrosserie CE	Genre	N° réception
Catégorie	Marque	Poids vide
Classe environnement CE	Masse F1	Puissance administrative
CNIT	Masse F2	Puissance maximale
CO^2	Masse F3	Ratio puissance masse
Couleur	Masse F4	Type réception
Cylindrée	Mentions techniques	Usages
Date immatriculation	Nb places assises	Variante modèle
Date mise en circulation	Nb places debout	VIN
Dénomination commerciale	Niveau sonore	Vitesse moteur

FIGURE 48 – Variables de la base des immatriculations

ABS	Date début commercialisation	Poids-à-sec
Aide à la conduite	Date dernier tarif	Pollution
Alimentation	Date fin commercialisation	Puissance fiscale
Anti-démarrage	Date tarif origine	Puissance réelle CEE
Assistance freinage urgence	Date dernier tarif connue	Puissance réelle CEE libre
Capacité huile moteur	Dernier tarif connue	Puissance réelle DIN
Capacité liquide refroidissement	Freinage couple	Puissance réelle DIN libre
Capacité réservoir essence	Groupe retenu	Sous-catégorie
Catégorie	Homologué	Tarif origine
Classe origine	Identifiant	Types Mines
Classe retenue	Marque	Version
Cylindrée	Modèle	Vitesse maxi
Cylindrée réelle	Nombre cylindres	
Date classement origine	Numéro CNIT	

FIGURE 49 – Variables de la base SRA

Les principales difficultés de ce rapprochement sont les suivantes :

- Des écritures différentes des modèles de véhicules (voire des marques) entre les bases ;
- Des catégorisations différentes entre SRA et C14 pour un même véhicule ;
- Une multiplicité des identifiants SRA pour un même code moto C14 [8] ;
- Une variabilité des codes motos dans le temps.

Les moyens mis en œuvre pour répondre aux difficultés sont les suivants :

- Le problème d'écriture entre les bases a nécessité dans un premier temps, le remplacement dans les clés de la variable modèle par d'autres variables. Puis, dans un second temps nous avons regardé si le nom du modèle de la base assurés/SIV était contenu dans le modèle SRA (base dans laquelle le modèle est souvent écrit de façon plus complète).
- La multiplicité des identifiants SRA a été gérée en cherchant dans un premier temps à rapprocher les immatriculations où il n'y avait pas d'ambiguïté sur l'identifiant SRA (i.e. identifiant SRA unique). Dans un deuxième temps, des hypothèses ont été faites veillant à ne pas introduire un biais trop important dans l'étude.
- La variable catégorie de véhicule a été utilisée pour vérifier et valider la cohérence du rapprochement.
- Pas d'utilisation des codes motos pour le rapprochement

8. Dans le véhiculier C14, chaque véhicule est identifié par un code moto.

Les hypothèses faites aux différentes étapes pour le rapprochement sont les suivantes :

À chaque étape, les hypothèses ont été définies avec soin suite à une analyse des modèles non rapprochés et tenant compte de l'éventuel biais. Les premières hypothèses ont été faites afin de récupérer les informations les plus précises possibles, dans le cas où ces hypothèses étaient insuffisantes, des hypothèses plus larges ont été définies (dernier point) dans la mesure où le biais introduit n'était pas trop grand.

- Dans le cas un modèle SIV correspondont à plusieurs modèles SRA, si le nombre de versions de véhicules est d'au plus 4 et si l'écart de prix entre les versions de véhicule est d'au plus 20%, la version choisie est celle ayant le prix le plus élevé considéré.
- Dans le cas où une information est manquante dans la base SIV / contrat, l'information des contrats pour lesquels le rapprochement a été réussi est utilisée. Dans le cas d'ambiguïté entre plusieurs modèles, le modèle le plus répandu (en termes d'années police) est choisi à condition que la somme des années d'exposition associée à ce modèle soit supérieure à 5. Les modèles/versions de véhicules dont l'exposition est inférieure à ce seuil, ont une information disponible insuffisante pour retirer de l'information pour rapprocher les autres véhicules.

Compléments sur les modèles GLM

A - Détail et complément sur la loi de probabilités pour le coût moyen

Dans un premier temps, une analyse graphique a été utilisée afin de comparer l'adéquation des lois à notre échantillon. Pour cela, l'histogramme de l'échantillon a été superposé aux fonctions de répartition des lois à tester. Ce premier test semblait indiquer un meilleur ajustement de la loi gamma (cf figure 50)

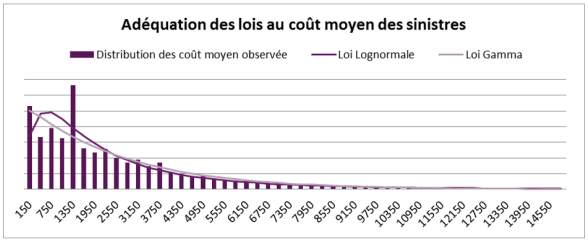


FIGURE 50 – Adéquation des lois au coût moyen des sinistres

Remarque Le pic observé sur le graphique 50 provient de la convention IRSA (Indemnisation règlement des sinistres automobiles) qui règlemente les rapports assureurs pour l'indemnisation de leurs assurés. Le montant du recours forfaitaire de la convention IRSA est de 1236€ avant le 1er janvier 2014 et de 1308€ à partir de cette date.

Dans un deuxième temps, les QQ-plots ont été tracés afin d'avoir une deuxième indication sur la loi qui s'ajuste le mieux aux données (cf figure 51). L'étude des QQ-plot laisse à penser que la loi log-normale s'ajuste mieux aux données et que la loi gamma possède une queue de répartition inadaptée aux données. Sur ce critère, la loi log-normale semblerait donc plus s'adpatée aux données.

Ensuite, un test de Kolmogorov-Smirnov a été mis en place (cf dans la figure 52). Ce test, consistant à mesurer l'écart maximum entre les fonctions de répartition aux données et de la loi à tester, conclut au rejet des deux lois testées.

Ce double rejet peut sembler un peu décourageant. Cependant, il doit être pris avec un certain recul, le test de Kolmogorov-Smirnov est sévère et rejette facilement les lois de par sa construction et sa sensibilité aux maximums.

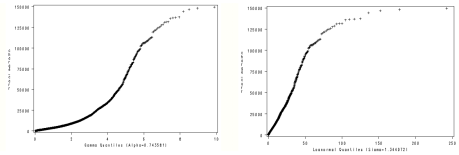


FIGURE 51 – QQ-plots lois Gamma et Log-normale

	Statistique D	P-value= Pr>D	Conclusion
Loi log-normale	0,049	<0,001	Rejet de H0
Loi Gamma	0,036	<0,001	Rejet de H0

FIGURE 52 – Résultats des test Kolmogorov-Smirnov

En plus des éléments présentés dans le corps du mémoire, une autre solution a été testé. Celle-ci consistait à utiliser une modélisation GLM avec un lien logarithme et une loi d'erreur gaussienne. Dans un tel cas, le logarithme de l'espérance conditionnelle suit une loi gaussienne. C'est alors chaque case tarifaire qui suit une log-normale. La loi du coût est log-normale conditionnellement au fait d'être dans une case tarifaire. Lorsque que le coût est déconditionné et agrégé, celui-ci suit un mélange de lognormale avec pour loi de mélange le poids des différentes cases tarifaires. Un tel modèle a été testé, mais l'analyse des erreurs a montré que les hypothèses de modèles GLM avec un tel modèle n'étaient pas respectées sur nos données. Pour ces raisons, ce modèle a donc été mis de côté.

B - Détail des calculs sur la loi de probabilité pour le modèle de fréquence

Le test du χ^2 va être employé pour tester l'adéquation de la fréquence à une distribution discrète.

Test d'adéquation à la loi de Poisson

La moyenne empirique est calculée afin d'estimer le paramètre de la loi.

$$\widehat{\lambda} = \overline{N} = \frac{1}{n}\sum_{i=1}^{n} N_i = 2,33\%$$

La loi Poisson est définie ainsi :

$$P(N = k) = \frac{e^{-\lambda}\lambda^k}{k} \text{ avec } E(N) = \lambda \text{ et } V(N) = \lambda$$

La statistique du χ^2 est de $927,90$. Le quantile d'ordre 99% (respectivement 95%) de la loi du χ^2 à 4 degrés de liberté est de $13,28$ (respectivement $9,49$). L'hypothèse H_0 est alors rejetée. Le nombre de sinistres ne s'ajuste pas à une loi de Poisson.

Test d'adéquation à la loi de Binomiale négative

Soit la loi Binomiale négative, de paramètres ν et $\frac{\nu}{\nu+\lambda}$. On note λ l'espérance de cette loi et $\lambda(1-\frac{\lambda}{\nu})$ sa variance.

Soit $\Gamma(\lambda)$ la valeur de la fonction gamma en λ. On a alors :

$$P(N=k) = \frac{\Gamma(k+\nu)}{\Gamma(k+1)\Gamma(\nu)} \times \left(\frac{\nu}{\nu+\lambda}\right)^{\nu} \times \left(\frac{\lambda}{\lambda+\nu}\right)^{k}$$

Nous estimons l'espérance de la loi par la moyenne empirique et la variance par la variance empirique de l'échantillon :

$$E(N) = \overline{N} = \frac{1}{n}\sum_{i=1}^{n} N_i = \widehat{\lambda} = 2,33 \times 10^{-2}$$

$$V(N) = S_N^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(N_i - \overline{N}\right)^2$$

$$\widehat{\nu} = \frac{\overline{N}^2}{S_N^2 - \overline{N}} = 4,63 \times 10^{-1}$$

La statistique du χ^2 est de $1,99$. Le quantile d'ordre 99% (respectivement 95%) de la loi du χ^2 à 2 degré de liberté est de $9,21$ (respectivement $5,99$). L'hypothèse H_0 est acceptée. Au seuil 1%, le nombre de sinistres s'ajuste à une loi binomiale négative.