Institut de Science Financière et d'Assurances
ISFA

MÉMOIRE POUR L'OBTENTION DU TITRE D'ACTUAIRE

Addactis

Réalisation d'un véhiculier à l'aide d'outils de Machine Learning

Novembre 2019

Étudiant :
François-Xavier Chamoulaud

Tuteur :
Guillaume Rosolek

Résumé

L'assurance automobile est un marché tendu et concurrentiel, l'assureur doit donc élaborer des modèles de tarification plus précis et à la pointe de la technologie pour approcher correctement le risque et rester dans la course au tarif. Les variables relatives aux conducteurs et à la géographie du risque sont les plus utilisées par les assureurs. Ainsi, comme dans le cadre de l'élaboration d'un zonier géographique, l'assureur peut s'intéresser à l'analyse de ses résidus pour proposer une segmentation homogène du risque en fonction du véhicule, appelée véhicule, afin de compenser les biais de structure du portefeuille. Le risque véhicule est difficile à approcher, car beaucoup de véhicules sont peu exposés. La classification proposée par la Sécurité et Réparation Automobile (SRA) permet de segmenter le risque, mais cette classification est-elle optimale pour notre portefeuille automobile issu d'un assureur partenaire ?

Dans une première partie, nous avons réalisé un *clustering* des véhicules, en ignorant la sinistralité observée. Deux approches ont été développées, une première en réalisant une carte des véhicules, puis une seconde en utilisant la notion de distance entre véhicules.

Dans une seconde partie, nous avons extrait le résidu d'un modèle GLM sans variable véhicule afin d'élaborer un véhiculier pour la garantie dommage, séparément pour la fréquence et le coût moyen. La technique du Krigeage est utilisée pour lisser le résidu grâce à une carte des véhicules. Le résidu lissé est ensuite modélisé à l'aide d'un *Gradient Boosting Machine* permettant de capter les variables explicatives du risque véhicule.

Les nouveaux véhiculiers sont alors intégrés séparément dans une modélisation GLM afin d'évaluer leurs efficacités respectives et de conclure sur leurs éventuels apports.

Mots clés : Tarification, Véhiculier, Apprentissage non supervisé, Clustering, Assurance automobile, Apprentissage supervisé, Résidu, Krigeage, Gradient Boosting Machine

Abstract

Car insurance is a complex and highly competitive market ; thus insurers have to elaborate thinner pricing methodology and always be at the edge of innovation to remain a significant market participant. Driver-related and geographic factors used to be the top interest variables for insurers. Following what can be done for zoning variables, insurers could analyze residuals derived for car-related features to build a strong and homogeneous segmentation and reduce bias induced by portfolio structure. It is generally challenging to estimate vehicle risk given the number of vehicles with low exposure. The *Sécurité et Réparation Automobile* (SRA) provides a classification which enable the insurer to make a first segmentation, however is this classification optimal regarding an insurer portfolio ?

Firstly, we conducted an unsupervised vehicle clustering regardless of the claim frequency or severity. Two clustering approaches have been developed, one using a vehicle map and the other using vehicle distance measures.

Secondly, we extracted the residuals from a GLM model to develop a vehicle classification for damage warranty, for both the frequency and the average cost. The Kriging method was used to smooth residuals thanks to a vehicle map. Finally, a Gradient Boosting Machine was fitted for modeling smoothed residuals, which allowed us to determine explanatory variables for vehicle-linked risk.

New vehicles classifications are tested separately in a GLM model in order to assess respectively their own performances and to conclude on their potential benefits.

Keywords : Pricing, Vehicle classification, Unsupervised learning, Clustering, Car insurance, Supervised learning, Residual, Kriging, Gradient Boosting Machine

Remerciements

Je souhaite remercier en premier lieu Guillaume Rosolek, responsable du pôle *Pricing & Data (P&C)* et Anne-Charlotte Bongard, associée du cabinet, pour m'avoir laissé l'opportunité de réaliser mon alternance au sein du cabinet et de m'avoir accordé leurs confiances tout au long de la réalisation de ce mémoire. Je tiens aussi à remercier Nabil Rachdi, responsable de la partie *data*, qui m'a permis d'interpréter mes résultats et à les concilier avec le point de vue métier.

Mes remerciements les plus sincères à Jean-Louis Rullière, tuteur académique de ce mémoire, pour son suivi et les conseils qu'il m'a donnés pour la rédaction de ce mémoire.

Je tiens aussi à mentionner les collaborateurs suivants, qui m'ont beaucoup apporté lors de mon parcours au sein du cabinet, que ce soit lors de mon alternance ou lors de la réalisation de ce mémoire : Victor Bouton, Aubin Chauveaux, Félix Hébert, Claire Lamon, Annabelle Longo, Victoria Delavaud, Pierre Chatelain, Romain Gauchon, Xuan-Quang Do et Montassar Ben Laiba.

Enfin, je tiens à remercier ma famille et mes amis, avec une mention spéciale pour mes camarades de promotion ISFA, Manon Dal Pont et Guillaume Gillot.

Table des matières

Introduction

L'assurance automobile est un secteur très concurrentiel. Tout conducteur est dans l'obligation de souscrire à minima la garantie responsabilité civile. Les assureurs utilisent ce levier pour en faire un produit d'appel. Ainsi, ce marché se livre une véritable guerre des prix, souvent revus à la baisse, face à des revalorisations tarifaires toujours à la hausse et des directives de plus en plus contraignantes pour l'activité d'assurance : changement des conditions de résiliations (loi Hamon), transparence sur la distribution des produits d'assurances (Directive sur la Distribution d'Assurances). D'autant plus que l'avènement des comparateurs d'assurances entraîne une forte concurrence des tarifs entre les différents acteurs du marché. Un assureur doit être compétitif pour tirer son épingle du jeu.

Le principe de base de l'assurance est la mutualisation, mais pour répondre à ces contraintes commerciales, un assureur doit jouer sur plusieurs paramètres, dont celui de l'affinage de la segmentation de son risque. L'assureur va donc chercher à créer des groupes homogènes de risque, en prenant garde aux risques de concentration et les risques d'antisélection.

Dans ce mémoire, nous étudions la segmentation des véhicules assurés, communément appelée véhiculier. La pertinence de la réalisation d'un véhiculier réside dans la quantité d'informations disponible sur les véhicules. Celles-ci sont accessibles par la base SRA, une base de données commune aux assureurs fournis par l'association *Sécurité et Réparation Automobile* . Néanmoins, les assureurs n'exploitent pas la totalité de leurs données. Une classification des véhicules est réalisée par la SRA. La segmentation proposée par cette dernière est la suivante : une classe de risque pour expliquer les évolutions de coût du véhicule et un groupe de risque pour expliquer la dangerosité intrinsèque du véhicule. Ces variables sont très utilisées sur le marché français pour les modèles de fréquence et de coût. Pourtant, elles agrègent de l'information dont l'assureur n'a pas forcément le contrôle et le véhiculier SRA n'est pas toujours adaptée pour la modélisation de la fréquence. Ainsi, les assureurs tendent à réaliser leur propre véhiculier pour affiner le tarif.

L'objectif de l'assureur est alors de mieux connaître son risque, dans un contexte où la concurrence est très rude à cause des évolutions réglementaires. Il est aussi confronté à de nombreuses contraintes opérationnelles liées au véhiculier et à des problèmes de fiabilité des bases externes à sa disposition.

Première partie

Contexte

Le produit automobile est un produit d'appel pour l'assureur non-vie. Celui-ci permet d'attirer le client dans son portefeuille. La difficulté se situe dans la rétention du client. En effet, l'évolution du marché fait que la concurrence est rude dans ce secteur. Ainsi, depuis une dizaine d'années, nous observons sur le marché une hypersegmentation des portefeuilles automobile, malgré de nombreuses contraintes réglementaires.[1]

Depuis quelques années, la segmentation se poursuit sur le véhicule. La classification SRA est utilisée par un grand nombre d'acteurs de l'assurance, mais ne s'appuie que sur la responsabilité civile. Qu'en est-il des garanties vol et dommage ? La classification SRA n'est pas construite spécifiquement pour ce type de garantie. Les assureurs se sont intéressés à analyser les résidus de leurs modèles pour repenser la classification pour les autres garanties. Néanmoins cette méthodologie pose plusieurs problèmes :

— **Stabilité des résidus** : la taille du portefeuille influe sur la convergence des algorithmes. Ainsi, plus le portefeuille est grand, plus la loi les des grands nombres est applicable et la stabilité des résidus observable. Qu'en est-il alors d'un portefeuille de taille intermédiaire ?

— **Données externes** : comment pouvons-nous utiliser des données externes pour améliorer notre segmentation ?

— **Choix de la variable véhicule** : l'abondance des variables liées au véhicule, face à celles du conducteur, fait qu'un choix s'impose pour intégrer le véhicule dans l'équation tarifaire. Il faut trouver les variables les plus pertinentes pour capter le risque.

Dans notre problématique de réalisation d'un véhiculier, nous devons comprendre les éléments associés à sa réalisation. Ainsi, nous allons dans un premier temps étudier le marché de l'automobile, pour ensuite nous focaliser sur le marché de l'assurance automobile dans sa globalité, jusqu'à s'approcher des raisons incitant l'assureur à réaliser un véhiculier. Nous explorerons enfin les différentes méthodologies qui ont été adoptées dans la littérature actuarielle afin de répondre à ce besoin croissant sur le marché français.

1. Application de *la Gender Directive* au secteur de l'assurance depuis 2012

Chapitre 1

Le marché du secteur automobile en France

Le secteur automobile est un secteur en bonne santé, que ce soit en termes de production de véhicules ou de ventes de véhicules en France ou à l'étranger. Ce secteur est à la pointe de l'innovation, avec un budget de recherche et développement représentant 15% du budget global de la R&D, soit la deuxième branche en France.

Les chiffres du Comité des Constructeurs Français d'Automobiles (CCFA) permettent d'entrevoir les spécificités du marché automobile. Même si la crise financière de 2007 avait impacté ce secteur, l'industrie automobile a réussi à se relever. Le groupe Renault et le groupe PSA sont les deux mastodontes de l'industrie française, représentant respectivement en 2017 27,5% et 39,9% du parc automobile des ménages français.

1.1 Le parc automobile des ménages français

Le parc automobile français se dénombre à environ 39 millions de véhicules d'après la CCFA pour un âge moyen d'environ 9 ans . 74% de ces véhicules sont utilisés quotidiennement, majoritairement pour des déplacements domicile-travail.

Le vieillissement du parc est lent et régulier, hormis dans des moments forts comme au début des années 2000 ou lors de la mise en place de la prime à la casse en 2009. Cette prime est issue d'un programme français nommé "Aide à l'acquisition des véhicules propres". Cette aide a été initialement accordée pour les véhicules vendus entre le 4 décembre 2008 et le 31 décembre 2009. L'objectif de ce programme est de faciliter le remplacement des vieux véhicules, en offrant une prime de 1 000 € pour les véhicules de plus de 10 ans. Cet âge est décompté à partir de la date de première mise en circulation jusqu'au jour de facturation du nouveau véhicule. Ce nouveau véhicule devait respecter certaines conditions sur les émissions de dioxyde de carbone.[1] Ce programme a ensuite été prolongé en 2010 avec une réduction de prime. Les conséquences d'une telle aide peuvent être multiples : mutation du parc automobile assuré, augmentation des coûts de réparation du fait de la sophistication des pièces automobiles.

En plus de la prime à la casse, il existe un dispositif fiscal incitant les ménages à l'achat d'un véhicule à faible émission de CO_2 : **la prime à la conversion**. Ce dispositif prend sa source dans un texte réglementaire issu du décret n°2014-1672 du 30 décembre 2014. Ce texte a été codifié dans la partie réglementaire du code de l'énergie (décret n°2015-1823 du 30 décembre

1. inférieures ou égale à 160 grammes de CO_2 par kilomètre

2015). Treize conditions sont à respecter pour être éligible et le montant de la prime peut atteindre jusqu'à 2000 €. En 2019, les modalités et le montant de la prime à la conversion ont changé. Le seuil de CO_2 passant à 122 grammes de CO_2 par kilomètre et le montant de la prime pouvant aller jusqu'à 5000 € pour les foyers non imposables, dits modestes, pour l'achat d'un véhicule hybride ou électrique. Ainsi, ces dispositifs fiscaux peuvent entraîner dans les années à venir des changements dans la composition du portefeuille.

L'âge du parc automobile n'est pas le seul critère à observer. La durée de rétention du véhicule est une information importante dont l'assureur doit être conscient. Les Français conservent leurs véhicules plus longtemps qu'avant, atteignant une moyenne de 5.6 années en 2017.

Enfin, le kilométrage d'un véhicule du parc atteint en moyenne 106 000 kilomètres. Néanmoins, en réalisant une ventilation sur le type d'alimentation du véhicule, nous remarquons qu'un véhicule diesel roule en moyenne 125 700 km contre 77 800 km pour un moteur essence. Le type d'alimentation entraîne un usage différent du véhicule assuré.

1.2 Le secteur de l'entretien et de la réparation automobile

L'entretien-réparation de véhicules a connu une baisse de son activité suite à la crise de 2008, mais elle a connu une relance sur ces dernières années (+4.4% en 2017). Le vieillissement du parc automobile français vient soutenir la relance de ce secteur. En effet, un véhicule nécessite d'être entretenu pendant toute sa durée de vie, d'après l'enquête *Kantar TNS Parc Auto*, une voiture du parc automobile français subi en moyenne, deux opérations d'entretien- réparation par an.

Les prix des pièces d'un véhicule sont pris en compte dans l'élaboration de la classe de prix du véhiculier SRA. En effet, les constructeurs et concessionnaires mettent en place des stratégies sur les pièces d'un véhicule pour réaliser des marges importantes. Ces pièces peuvent être divisées en deux catégories, les pièces d'usures et les pièces dites esthétiques.

Ce secteur subit aussi des refontes législatives. En effet, un amendement à la loi d'orientation des mobilités (LOM) devrait mettre fin au monopole des constructeurs sur la vente des pièces détachées automobiles dites "visibles". Ces pièces sont protégées au titre du droit des dessins et modèles, ce qui donne aux constructeurs automobiles un monopole de droit à la fois sur le processus de commercialisation et de fabrication. Cet amendement commencera à s'appliquer sur les rétroviseurs, les phares et les vitrages à compter du 1er janvier 2020. Il s'étendra ensuite sur les pièces de carrosserie, qui représente un fort coût dans le panier des pièces automobiles.

Chapitre 2

Marché de l'assurance automobile en France

2.1 Les chiffres du marché automobile en France

Le produit automobile représente à lui seul 55,6% des cotisations d'assurance pour le marché des particuliers sur la branche non-vie. L'assurance automobile est un produit d'appel permettant à l'assureur de faire souscrire à ses clients des contrats annexes. Le ratio combiné oscille généralement autour de 100% sur ces dernières années, d'après la Fédération Française de l'Assurance (FFA), ce qui montre qu'il est difficile pour l'assureur de faire un quelconque profit sur ce type d'assurance.

Le parc des véhicules assurés augmente d'année en année, atteignant 42.2 millions de véhicules de première catégorie en 2017. De fait, nous observons simultanément une évolution des cotisations, que ce soit pour la branche responsabilité civile ou dommage. Le cumul atteignant un solde de 21 291 millions d'euros pour l'année 2017.

2.2 Évolution globale de la sinistralité sur les garanties automobile

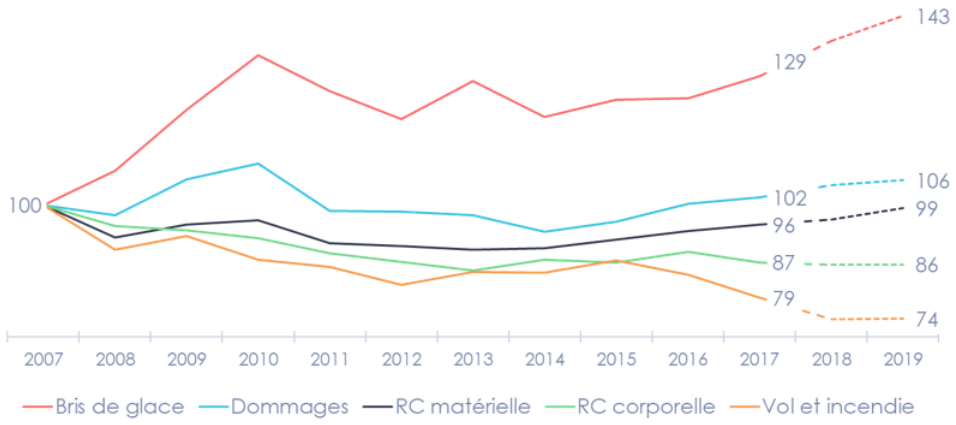

FIGURE 2.1 – Évolution des primes pures par garantie sur le produit automobile et prévisions pour 2019 (Source : FFA / Addactis)

L'évolution sur le graphique 2.1 de la prime pure est portée par une augmentation des coûts moyens liée à une augmentation des prix des pièces automobiles, que ce soit les pare-brises, ou les pièces détachées. Cette augmentation est plus forte que la baisse de la sinistralité.

2.3 Le dilemme entre mutualisation et segmentation

Un des principes fondamentaux de l'assurance est la mutualisation. Dans un contexte concurrentiel, l'assureur s'est tourné dans un premier temps vers les variables géographiques (élaboration de zonier) et les variables conducteurs (modélisation comportementale) pour pouvoir affiner son tarif face à un marché exerçant une pression de plus en plus forte : offre promotionnelle, mise en avant des conditions de la loi Hamon, boitier télématique ... De plus, les solutions innovantes proposées par les *InsurTech*, voulant améliorer le parcours client, tendent à réduire drastiquement le nombre de questions à poser à l'assuré.

En moyenne, un assureur pose une trentaine de questions en assurance automobile afin de proposer un tarif. Ces questions vont de l'expérience de conduite, jusqu'à la situation matrimoniale du conducteur. L'assureur demande aussi la date de mise de circulation et la date d'acquisition du véhicule pour en déduire l'âge du véhicule ou la durée de détention de celui-ci. Ensuite, à l'aide de la plaque d'immatriculation, il est possible de retrouver le véhicule assuré. Les assureurs ont alors accès a de nombreuses informations grâce aux bases de données auxquelles ils ont accès afin de retrouver le véhicule, sans poser beaucoup de questions à l'assuré. Ce point est un réel atout pour l'assureur, sans pour autant allonger le parcours client. En effet, les questions posées à l'assuré sont un frein à la souscription d'un produit d'assurance et l'amélioration de la satisfaction client est au cœur des enjeux de demain. Néanmoins, cela nécessite une fiabilisation des bases de données.

Il existe aujourd'hui différentes méthodes actuarielles de tarification. L'intégration de toutes les modalités caractérisant un véhicule perturbe la calibration et les capacités prédictives du modèle. Ainsi, le recours à une variable captant l'effet à expliquer est une alternative souvent envisagée (comme le zonier pour le risque géographique).

Le véhiculier est un réel gain pour l'assureur : il permet de poser moins de questions, tout en exploitant la richesse de ses données à des fins d'optimisation tarifaire.

2.4 L'objet d'intérêt du mémoire : le véhicule

2.4.1 Typologie d'un véhicule

Lorsque nous nous intéressons à un véhicule, à l'aide de la base de données SRA, nous avons accès à de nombreux éléments caractérisant le véhicule :

- **La carrosserie** représente l'enveloppe métallique recouvrant un véhicule motorisé. Généralement en aluminium, certains véhicules particuliers (sport, voiture de luxe) utilisent d'autres matériaux pour les réaliser. De fait, la carrosserie est un excellent prédicteur pour le prix des réparations des véhicules.
- **L'alimentation** désigne le carburant utilisé. Nous observons une mutation du marché, avec l'apparition des **voitures hybrides** (moteur thermique et électrique) et des **voitures électriques** (moins de bruit et forte accélération).
- **La puissance du moteur** : La puissance s'explique par de nombreux facteurs. Celle-ci est à différencier de la puissance fiscale d'un véhicule qui elle prend en compte le rejet de

CO^2 en cycles normalisés. La cylindrée, le nombre de cylindres, la suralimentation (turbo ou compresseur), l'injection sont tant de facteurs qui peuvent expliquer ce chiffre.

— **La taille du véhicule** : De nombreuses variables vont décrire cette caractéristique : longueur, largeur, hauteur, empattement (distance entre les roues).

— **Prix des pièces - panier SRA** : La SRA réalise à partir d'un catalogue des constructeurs un panier de pièces dont il suit l'évolution. Cette variable permet ensuite de mettre à jour la classe de réparation.

— **Prix du véhicule neuf** : La SRA communique sur le dernier prix observé à neuf du véhicule. Cette donnée pourrait être mêlée avec le prix du véhicule d'occasion (données externes à collecter).

2.4.2 La classification SRA

- Actuellement, les assureurs segmentent leurs véhicules grâce au modèle crée par la SRA, la classification s'opère au sein de 3 classes :

— **Le groupe**, captant la dangerosité du véhicule, est souvent utilisé pour déterminer la responsabilité civile. Elle est comprise entre 20 et 50. Des variables comme la puissance fiscale, le poids, peuvent expliquer cette segmentation.

— **La classe de prix** est fixée par la valeur à neuf TTC du véhicule (hors option et hors remise) par tranches de prix. Elle peut-être utilisée pour la garantie remplacement à neuf ou perte totale. Les valeurs oscillent entre A et ZA + HC (Hors classe)

— **La classe de réparation** est définie par le coût HT du panier de pièces SRA. Cette variable est mise à jour par la SRA qui va capturer l'évolution du prix d'un panier de pièces au fil du temps. Elle permet de donner une idée sur le prix de réparation des véhicules. Néanmoins, le calcul du panier SRA ne fait pas la distinction entre les pièces d'usure (liées aux concessionnaires) et les pièces d'accident (liées aux assureurs).

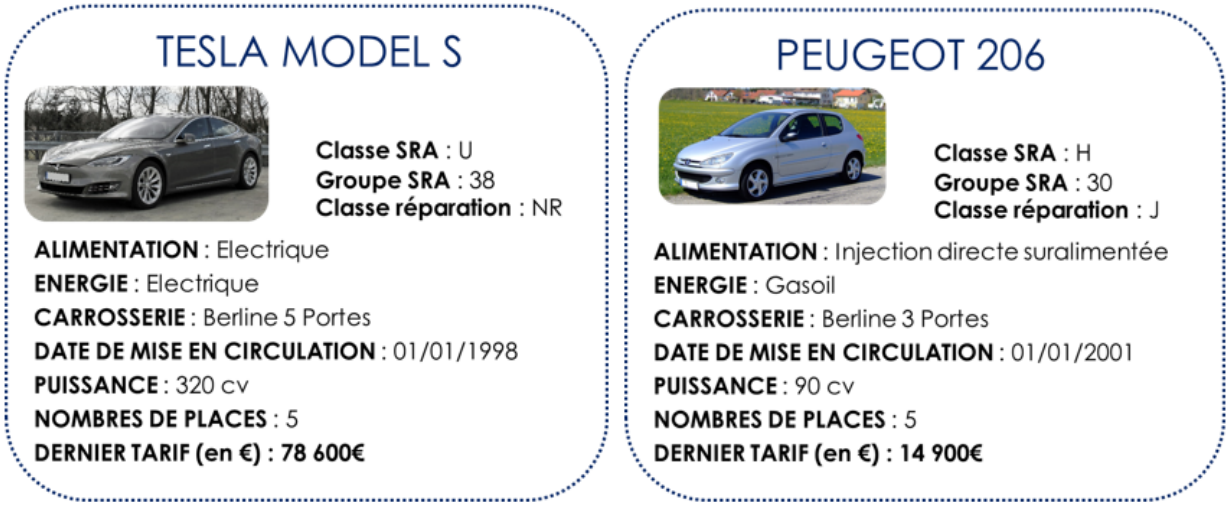

FIGURE 2.2 – Cartes d'identités de deux véhicules présents dans notre base

Sur la Figure 2.2, nous retrouvons les cartes d'identité de la Tesla Model S et de la Peugeot 206 présentes dans la base SRA. Il existe de nombreux dérivés de ces deux modèles. La base véhicule créée dans le cas pratique doit capter ces variations. Les informations sont issues de la base SRA.

Généralement, la classe de prix et la classe de réparation sont agrégées en une seule classe. Nous observons souvent des classes de réparation à la hausse du fait de la hausse des prix des pièces.

Néanmoins, avec *la libéralisation des pièces détachées automobiles*, nous pourrons voir à l'avenir une future évolution à la baisse de la classe de réparation, d'où l'intérêt du travail réalisé sur le calcul des paniers de pièces SRA. Cette libéralisation s'opérera en deux temps :

— Ouverture du marché des rétroviseurs, phares et vitrages : cet impact pourra se voir directement sur la base SRA.

— Ouverture du marché de la carrosserie.

D'après l'UFC Que Choisir, cette libéralisation permettra évidemment à l'assuré de réaliser des économies sur sa prime d'assurance.

2.5 L'intérêt d'un véhiculier

Le véhiculier SRA permet d'agréger correctement les différentes informations des véhicules et est fortement corrélé avec les principales variables représentatives de la sinistralité. Pourtant, ce véhiculier ne s'applique pas à toutes les garanties d'un contrat automobile.

De plus, dans une logique concurrentielle, il est primordial pour l'assureur d'internaliser ses modèles. N'ayant pas la main sur la classification SRA, il lui est alors impossible de modifier cette classification pour prendre en compte un critère commercial par exemple. De plus, comme évoqué précédemment, ce véhiculier ne permet pas de capter le risque intrinsèque à une garantie spécifique.

Les assureurs veulent réaliser (ou affiner) leurs véhiculiers pour différentes raisons :

— **Amélioration de la segmentation SRA** : L'assureur cherche à améliorer le critère développé par la SRA, à partir de variables explicatives de notre choix et pour de multiples garanties.

— **Amélioration du processus de tarification** : l'objectif de cette nouvelle classification est d'optimiser nos classes et notre tarif en segmentant le risque relativement à notre portefeuille (et non pas au marché).

— **Lissage du tarif** : l'objectif premier d'un véhiculier est d'éviter des sauts de tarifs rendant la commercialisation plus difficile (par exemple, si un client se trompe de modèle, l'écart du tarif doit être interprétable pour le commercial qui va vendre le contrat).

— **Intégration de la sinistralité et/ou pilotage marketing** : un véhiculier permet d'intégrer la sinistralité observée sur le portefeuille pour chaque garantie. Il permet aussi, si celui-ci est de nature commerciale, de considérer des éléments marketing qu'un assureur voudrait appliquer sur son portefeuille.

— **Avis d'expert** : le véhiculier doit pouvoir prédire un score ou un groupe pour les nouveaux véhicules et mettre à jour les véhicules avec une nouvelle typologie. Le véhiculier pourra être un outil pour vérifier que la classe donnée par une saisie manuelle est adéquate.

— **Stabilité de la classification** : l'objectif est d'avoir un véhiculier stable, que l'assureur pourra réviser à un rythme fixé.

2.6 Panorama des méthodologies relatives au véhiculier

La théorie sur la création d'un véhiculier a été peu étudiée dans la littérature actuarielle. Ce sont généralement des processus internes à l'entreprise se rapprochant brièvement des techniques de lissage géospatial, souvent utilisées dans le cadre de la création de zonier.

Pourtant, il manque l'aspect géographique pour pouvoir appliquer cette théorie et avoir une notion de voisinage. Le passage par une représentation spatiale de nos véhicules est nécessaire, c'est à dire une projection de notre base initiale sur un nouvel espace. Une nouvelle méthodologie pour représenter une carte des véhicules dans un espace de faible dimension sera proposée. (3 axes maximum)

Avant d'expliciter la démarche, nous allons faire un retour sur les travaux réalisés pour tenter de substituer l'approche établie par la SRA, utilisée majoritairement par le marché.

Deux références sont présentes dans la littérature actuarielle : un mémoire de R. Sipulskyte [19] sur le développement d'une classification de véhicule en Nouvelle-Zélande. Son approche s'articule d'une part sur l'extraction de l'effet véhicule à l'aide d'un Modèle Linéaire Généralisé *(General Linear Model)*, qui est capturé dans les résidus du modèle, puis par un lissage de crédibilité.
J. Lavenu [12] a ensuite repris son approche sur un portefeuille pour le produit deux roues. Son mémoire a permis d'illustrer la réalisation d'une carte des véhicules avec une approche par AFDM (Analyse Factorielle de Données Mixtes) pour projeter sur un espace la variation expliquée dans le jeu de données, puis a relié les voisins par triangulation de Delaunay. [1] Cette carte est ensuite utilisée pour réaliser un lissage spatial sur la sortie de son modèle d'apprentissage supervisé qui tente de modéliser les résidus issus de son GLM.

La triangulation de Delaunay n'est pas une méthode optimale du fait de la concentration des points autour du centre de la carte des véhicules, nous explorerons une nouvelle piste dans le cadre de ce mémoire en réalisant des cartes de véhicules à partir d'une matrice de dissimilarité.

Néanmoins, l'approche par AFDM est un outil très adapté pour explorer notre base de véhicules. Nous l'utiliserons pour réaliser une première classification et pour explorer nos véhicules d'une manière plus poussée que des analyses univariées, tout en essayant de détecter des cases de tarification qui ne sont pas observables à l'aide d'analyse univariée.

1. Méthodologie développée par AXA Global P&C

Deuxième partie

Méthodologie et aspects théoriques

L'objectif de cette partie est de faire un état de l'art des méthodes utilisées actuellement en tarification dans le secteur de l'assurance (principalement en Non-Vie) et de soulever quelques problématiques rencontrées en pratique sur le traitement de la donnée. En effet, l'actuaire est garant de la qualité des données qu'il manipule, une attention particulière a été portée sur cette problématique.

Pour ce faire, nous avons utilisé des outils de *Machine Learning* pour réaliser une alliance entre les GLM et la *data science* à des fins d'amélioration du modèle tarifaire. Ainsi, nous présenterons les méthodes suivantes :

- **Les modèles linéaires généralisés** et ses applications dans notre problématique de véhiculier supervisé.
- Le principe du *boosting* et un algorithme de régression supervisé : ***le Gradient Boosting Machine***.
- Les méthodes de *clustering* permettant de créer des groupes de véhicules en analyse non supervisée.
- Les méthodes de **réduction de dimension** permettant une exploration intuitive des données et la création d'un nouvel espace d'observation pour nos véhicules.
- Une méthode de lissage géospatial pour lisser le résidu moyen agrégé contenant l'effet véhicule : **le Krigeage**.

Chapitre 3

Les fondements de la tarification en IARD Automobile

Les principaux aspects de la tarification IARD sont tirés du livre d'A. Charpentier [3] et du cours de X. Milhaud [14]. Nous reprendrons point par point les éléments essentiels de la constitution du tarif impactant l'efficacité du véhiculier ainsi créé.

Le principe fondamental de la tarification en assurance est la mutualisation.Elle consiste à répartir le risque solidairement entre plusieurs têtes. Cette mutualisation repose sur la loi des grands nombres :

Néanmoins, la segmentation d'un tarif peut remettre en cause cette loi asymptotique et il faudra alors surveiller que les classes créées présentent un effectif conséquent.

3.1 La composition d'une prime d'assurance

Une prime d'assurance se regroupe en plusieurs parties.

- **La prime pure** correspond au prix que l'assuré doit payer pour se couvrir seulement contre le risque, calculée grâce à notre modélisation.
- **La prime de risque** constitue la prime pure complétée d'un chargement de sécurité.
- **Les frais de gestion et administratifs** sont tous les frais relatifs à la gestion des contrats.
- **Les frais d'acquisition** désignent les frais nécessaires pour les affaires nouvelles.
- **Les frais de distribution et de commissionnement** (courtage, agents de ventes) sont les frais inhérents à la rémunération des courtiers et des agents généraux.

Notre étude portera uniquement sur la composition de la prime pure. Ainsi notre véhiculier impactera dans un premier temps la prime pure. Néanmoins, il pourrait être envisagé de créer un véhiculier commercial.
La modélisation de la prime pure est réglementée par des directives. L'assureur ne peut pas utiliser toutes les variables auxquelles il a accès pour tarifer le risque et n'a pas accès à toutes les variables observables.

3.2 L'intégration des contraintes réglementaires dans le tarif

Le processus de tarification en assurance est impacté au fil du temps par des modifications réglementaires en France ou par l'Union européenne.

Un premier exemple est la directive de 2012 rendant impossible l'intégration de la variable explicative relative au sexe de l'individu dans le processus tarifaire.

Un second exemple est la loi Hamon (2014) permettant à l'assuré de résilier son contrat dès qu'il le souhaite, après le premier anniversaire de son contrat. L'assuré est alors en capacité de résilier son assurance au bout d'un an. De fait, les évolutions tarifaires devront prendre en compte ce point pour éviter de faire fuir une partie du portefeuille sensible à une hausse des prix.

3.3 L'approche par le modèle collectif

Le modèle collectif est usité en assurance pour répartir le coût total des sinistres remboursés par l'assureur parmi les assurés en leur faisant payer une prime d'assurance.

Soit (N_t) le processus de comptage des sinistres sur la période [0,t]
Soit (Y_i) le montant du i-ème sinistre.

La perte totale sur la période [0,t] s'écrit donc :

$$S_t = \begin{cases} \sum_{i=1}^{N_t} Y_i \\ 0 \ si \ N_t = 0. \end{cases} \tag{3.1}$$

La prime pure annuelle π est égale à l'espérance de S_1, soit :

$$\pi = \mathbb{E}(S_1)$$

Si nous supposons que les sinistres sur la première année sont indépendants et que les (Y_i) sont i.i.d et que le nombre de sinistres est indépendant des montants, alors la prime annuelle est donnée par

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

Néanmoins, l'assureur observe dans son portefeuille de nombreux individus, avec des caractéristiques plus ou moins différentes. Nous sommes alors dans un contexte d'hétérogénéité que l'assureur doit prendre en compte. Il observe alors un vecteur X d'informations, relatif dans notre cadre aux données du conducteur, de la zone géographique et du véhicule assuré.

L'équation précédente devient donc, sous hypothèses d'indépendance par rapport au vecteur d'information X

$$\pi = \mathbb{E}(N|X) \times \mathbb{E}(Y|X)$$

L'objectif de la section suivante est d'illustrer la mise en place des GLMs pour estimer les deux quantités suivantes :

— $\mathbb{E}(N|X)$: La fréquence des sinistres sur l'année pour les assurés possédant les critères X.

— $\mathbb{E}(Y|X)$: Le montant moyen des sinistres pour les assurés possédant les critères X.

En résumé

La tarification en IARD se repose sur **le principe de mutualisation** et sur **le modèle collectif**. De nombreux facteurs sont à prendre en compte, parmi les évolutions réglementaires et les caractéristiques des assurés pour approcher le plus fidèlement possible les quantités d'intérêts à modéliser : la fréquence et le coût moyen.

Chapitre 4

Les modèles de régression

4.1 Les Modèles Linéaires Généralisés

Le modèle linéaire généralisé (*General Linear Model* - GLM) est une méthode statistique permettant d'étudier la relation entre une variable à expliquer Y appartenant à la famille exponentielle et un ensemble de variables explicatives $(X_1, ...X_n)$ grâce au lien d'une fonction g dérivable et inversible :

$$g(\mathbb{E}(Y|X = x)) = X\beta$$

Les paramètres β sont estimés par maximum de vraisemblance. Le lecteur intéressé pourra se référer au livre d'A. Charpentier [3] ou au livre de P. De Jong et al [16].

Ce modèle est beaucoup utilisé dans les compagnies d'assurance et dans les problématiques générales de régression. C'est une généralisation du modèle linéaire.

L'objet d'intérêt d'une des méthodologies proposées dans ce mémoire et d'extraire l'effet véhicule en réalisant un modèle GLM sans les variables véhicules. L'information non spécifiée se retrouve dans le résidu, mais quel résidu choisir pour notre modélisation ?

4.1.1 Les différents résidus d'un GLM

L'analyse des résidus est une étape importante pour valider la modélisation GLM. Hormis les résidus additifs $\hat{\epsilon_i} = Y_i - \hat{Y_i}$ qui présentent un intérêt limité, il existe d'autres résidus dont l'étude est primordiale et le choix d'un type de résidu par rapport à un autre va avoir un impact crucial sur notre modélisation du risque véhicule.

Les **résidus de Pearson** sont définis par :

$$\hat{\epsilon_{P,i}} = \frac{Y_i - \hat{Y_i}}{\sqrt{V(\hat{Y_i})}}$$

Ces résidus permettent de corriger le caractère hétéroscédastique du résidu additif classique.

Les **résidus de déviance** sont définis par :

$$\hat{\epsilon_{D,i}} = \pm\sqrt{d_i} \text{ et } D = \sum_{i=1}^{n} d_i$$

La quantité D est appelée déviance et permet de mesurer la qualité d'ajustement du modèle. En effet, $D = 2(\ ln(L_{sat}) - ln(L_{modele}))$ où L_{sat} représente la vraisemblance du modèle saturé et L_{modele} celle de notre modèle.

L'élément d_i dépend de la spécification de la loi de Y. Ainsi dans le cadre poissonien :

$$\epsilon_{\hat{D},i} = \pm\sqrt{|y_i \ log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)|}$$

Les **résidus d'Anscombe** se basent sur la différence $h(y_i) - h(\hat{y}_i)$ où la fonction h est choisie de telle sorte que la variable $h(y)$ est approximativement gaussienne. La fonction h est telle que $h'(y) = V(y)^{-1/3}$ où $V(.)$ est la fonction de variance de la réponse y, ainsi dans le cadre de Poisson : $V(\mu) = \mu$. Le résidu est enfin réduit en divisant par l'écart-type de $h(y)$ dont l'approximation est donnée par $h'(y)\sqrt{V(y)}$. Ainsi, le résidu d'Anscombe est défini par

$$\epsilon_{A,i} = \frac{h(y_i) - h(\hat{y}_i)}{h'(\hat{y}_i)\sqrt{V(\hat{y}_i)}}$$

Dans le cadre du modèle de poisson, on obtient alors la formule analytique suivante :

$$\epsilon_{A,P,i} = \frac{3}{2}\frac{y_i^{2/3} - \hat{y_i}^{2/3}}{\hat{y_i}^{1/6}}$$

4.1.2 Utilisation du GLM dans la réalisation d'un véhiculier prédictif

Le cours de X. Milhaud [14] permet de poser les bases des principes relatifs à la théorie sur les zoniers et à l'extraction d'un effet porté par une variable. Dans le cadre de ce mémoire, l'effet véhicule est la variable d'intérêt de notre problématique.

Dans notre exemple, nous nous intéresserons à un véhiculier de fréquence, mais il est également possible de réaliser un véhiculier de coût. Comme pour les variables d'intérêts géographiques, il existe une certaine maille de précisions pour les véhicules :

— La marque du véhicule pourrait s'assimiler à un pays

— Le modèle du véhicule pourrait s'assimiler à une région

— La version du véhicule pourrait s'assimiler à un département

Une première étape est de trouver la maille idéale pour modéliser le facteur véhicule et isoler l'effet véhicule qui ne dépend alors que d'un seul facteur.

Soit $\mathbf{X} = (X_1, ..., X_p)$ l'ensemble des variables explicatives. Supposons que X_1 est le critère véhicule. β_1 est donc l'effet véhicule.

La modélisation du nombre de sinistres à partir des p variables explicatives et de l'introduction d'un offset (l'exposition dans le cas d'un modèle de fréquence) est donnée par :

$$n_i = \text{offset}_i * e^{\beta_0 + \beta_1 X_1^i + ... + \beta_p X_p^i} + \epsilon_i$$

La méthodologie suivante est appliquée pour extraire l'effet véhicule :

1/ Modélisation de n en excluant X_1 du GLM, nous obtenons alors $\hat{\beta}_2, ..., \hat{\beta}_p$ permettant d'obtenir l'estimateur $\hat{n_i}$ sans l'effet véhicule.
Nous avons sous l'hypothèse de spécification correcte ($\hat{\beta}_2 = \beta_2, ..., \hat{\beta}_p = \beta_p$)

$$n_i = \hat{n_i} \times e^{\beta_0 + \beta_1 X_1^i} + \epsilon$$

2/ Lorsque l'exposition d'un modèle est différente de 0, posons le résidu multiplicatif donné par :

$$r_i = \frac{n_i}{\text{offset}_i e^{\hat{\beta}_2 X_2^i + ... + \hat{\beta}_p X_p^i}} = e^{\beta_0} \times e^{\beta_1 X_1^i} \times \frac{e^{\beta_2 X_2^i}}{e^{\hat{\beta}_2 X_2^i}} \times ... \times \frac{e^{\beta_p X_p^i}}{e^{\hat{\beta}_p X_p^i}} + \epsilon' \tag{4.1}$$

Sous l'hypothèse où $\hat{\beta}_2 = \beta_2, ..., \hat{\beta}_p = \beta_p$, nous définirons le risque véhicule par :

$$r_i = e^{\beta_0} e^{\beta_1 X_1^i} + \epsilon'$$

Lorsque l'exposition est nulle, nous prendrons $r_i = 0$

3/ Plaçons-nous dans le cas où k représente la maille de véhicule choisi. Nous introduisons désormais l'estimateur du résidu au niveau de la maille de véhicule k par :

$$\hat{r_k}^c = \frac{\sum_{i=1}^{l^k} e_i \hat{r_i}}{\sum_{i=1}^{l^k} e_i}$$

où e_i représente l'exposition (introduit en offset), l^k le nombre d'assurés possédant le véhicule k et $\hat{r_i}$ l'estimateur du risque véhicule résiduel, donné par l'équation (4.1)

4/ L'objectif est désormais d'extraire le signal de l'effet véhicule contenu dans le résidu à travers un ensemble de variables $\mathbf{X}_1 = (X_1^1, ..., X_1^l)$ décrivant les caractéristiques relatives aux véhicules.

Nous souhaitons expliquer l'estimateur $\hat{r_k}^c$ à l'aide de ces variables. Ainsi, nous cherchons une fonction de la forme :

$$\hat{r_k}^c \simeq f(X_{1,k})$$

Le résultat de cette modélisation permet de prédire un résidu final, comportant l'information portée par les véhicules.

A partir de ces étapes, nous créons une nouvelle base de données avec nos véhicules, puis nous construisons un modèle prédictif à l'aide de ces résultats. Ce modèle est expliqué par les variables véhicules présentes dans la base d'étude.

L'avantage de cette méthodologie est de réaliser des prédictions sur des véhicules peu présents dans notre portefeuille, mais également les véhicules dont la sinistralité est peu significative (du fait d'une faible exposition par exemple). Néanmoins, cette méthode nécessite un historique profond pour être fiable. Certaines garanties ne peuvent pas être modélisées à l'aide de cette méthode.

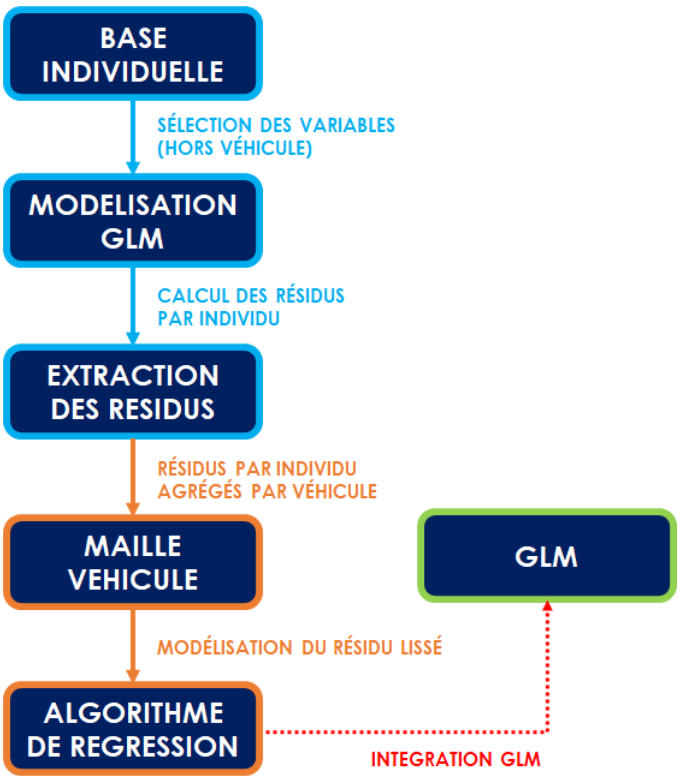

FIGURE 4.1 – Méthodologie employée pour la réalisation du véhiculer supervisé

Dans la méthodologie précédente, le choix du résidu multiplicatif réside dans l'élimination de l'*offset*. Néanmoins, d'autres résidus introduits précédemment, comme le résidu d'Anscombe, peuvent être plus adaptés dans notre problématique de réalisation d'un véhiculier.

4.2 Le *Gradient Boosting Machine* (GBM)

4.2.1 Principe du boosting

Le *boosting* est une méthode ensembliste se basant sur l'entrainement et l'agrégation séquentielle de *weak learners*[1] $f_m, m \in (1, ..., M)$ corrigeant le modèle précédent, afin de créer un estimateur robuste F_m où M est le nombre de modèles de modèle simple à agréger.

$$F_m(x) = \sum_{m=1}^{M} f_m(x)$$

L'objectif du *boosting* est de réduire le biais et la variance de l'estimateur $\hat{y} = F_M(x)$ crée. La particularité du *boosting* réside dans l'aspect séquentiel : le choix de f_m n'altère pas les anciens modèles. Le choix du nombre d'itérations M est un paramètre de *tuning* et plus celui-ci croit, plus nous pouvons espérer de la précision, mais réaliser en contrepartie du surapprentissage. Une visualisation simple du processus itératif est présentée sur l'équation 4.2. Nous rajoutons au modèle le *weak learner* f_m à l'itération m.

$$F_m(x) = F_{m-1}(x) + f_m(x) \tag{4.2}$$

Le principe du *boosting* en lui-même ne précise pas comment choisir f_m et cela dépend particulièrement de l'algorithme utilisé. Néanmoins, ce choix se repose sur l'optimisation du modèle en réalisant une descente de gradient.

4.2.2 La descente de gradient

Par la suite, la fonction de coût L n'est pas spécifiée. La fonction de coût associée au modèle pour prédire y sur l'échantillon d'apprentissage est :

$$L(F) = \sum_{i=1}^{N} L(y_i, F(x_i))$$

L'objectif est de minimiser $L(F)$ par rapport à $\mathbf{F} = (F(x_1), F(x_2), ..., F(x_N))$:

$$\hat{\mathbf{F}} = \underset{F}{\operatorname{argmin}}\ L(\mathbf{F}) \tag{4.3}$$

Les paramètres $\mathbf{F} \in \mathbb{R}^{\mathbb{N}}$ sont les valeurs de la fonction $F(x_i)$ en chaque point de la base d'apprentissage. Les méthodes numériques résolvent (4.3) comme une somme de vecteurs découlant de la somme télescopique (4.2) :

$$\mathbf{F}_m = \sum_{m=1}^{M} \mathbf{f}_m$$

où $\mathbf{F}_0 = \mathbf{f}_0$ est le modèle initial et chaque $\mathbf{F}_m$ est déterminé grâce aux paramètres $\mathbf{F}_{m-1}$.

A chaque itération m de l'algorithme, nous avons $L(F_m) = L(F_{m-1} + f_m)$. La recherche d'un *optimum* se détermine à l'aide d'une descente de gradient, ce qui permet de choisir $\mathbf{f}_m = -\rho_m g_m$ où ρ_m est un scalaire et $g_m = \nabla L(F)_{\mathbf{F}=\mathbf{F}_{m-1}}$ le gradient de $L(F)$ évalué en $\mathbf{F} = \mathbf{F}_{m-1}$.

1. Un *weak classifier* est un modèle légèrement meilleur qu'un modèle aléatoire

La longueur du pas ρ_m est solution de l'équation

$$\rho_m = \underset{\rho}{\operatorname{argmin}}\ L(\mathbf{F}_{m-1} - \rho g_m)$$

La solution actuelle est donc mis à jour :

$$\mathbf{F}_m = \mathbf{F}_{m-1} - \rho_m g_m$$

Le procédé est ensuite répété pour l'itération suivante. La descente de gradient est une stratégie efficace, puisque $-g_m$ est la direction dans $\mathbb{R}^N$ pour laquelle la fonction de coût $L(F)$ décroit le plus rapidement.

4.2.3 Présentation de l'algorithme GBM

Le GBM se fonde sur les arbres de régression (CART) et sur la méthode du *boosting*. La théorie sur le modèle CART n'est pas rappelée dans ce mémoire, mais l'utilisateur intéressé peut se référer à l'article de R. Timofeev [20].

A partir de la méthode de la descente de gradient évoquée précédemment, l'algorithme GBM suit la démarche suivante, présentée dans l'ouvrage de HASTIE et al.[7] :

1 Initialisation avec $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \gamma)$

Dans le cas où $L(y_i, \gamma) = (y_i - \gamma)^2$, nous avons $F_0 \equiv \bar{y}$ c'est à dire la moyenne de y.

2 Pour m allant de 1 à M :

a Pour i allant de 1 à N, calculer

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}$$

b Entrainer un arbre de régression sur r_{im} donnant des noeuds terminaux R_{jm} , $j = 1, 2, ..., J_m$

c Pour j allant de 1 à J_m, calculer

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

d Mise à jour : $F_m(x) = F_{m-1}(x) + \sum_{i=1}^{J_m} \gamma_{jm} 1_{x \in R_{jm}}$

3 Résultat : $\hat{F}(x) = F_M(x)$

L'implémentation en pratique et le *tuning* des paramètres seront évoqués dans la partie pratique. Nous pouvons tout de même identifier les différents types de paramètres à calibrer.

— Les paramètres relatifs au processus de *boosting* : le nombre d'arbres et le taux d'apprentissage ν (*shrinkage* ou *learning rate*). Ce dernier permet de régulariser la mise à jour du modèle à chaque itération en définissant un rythme d'apprentissage :

$$F_m(x) = F_{m-1}(x) + \nu \times f_m(x) \quad , \ 0 < \nu \leqslant 1$$

— Les paramètres relatifs aux arbres CART.

En résumé

Les modèles de régression présentés ci-dessous ont été utilisés dans notre modélisation pour approcher plusieurs quantités :

— **Le GLM** permet d'approcher le risque véhicule résiduel issu d'un modèle GLM sans les variables explicatives relatives aux caractéristiques techniques du véhicule.

— **Le GBM** permet de modéliser le risque véhicule résiduel à l'aide des caractéristiques techniques du véhicule.

Chapitre 5

Les méthodes de *clustering*

5.1 La problématique liée aux traitements des données quantitatives et qualitatives

Les bases de données rencontrées en assurance automobile possèdent deux types de variables : quantitatives et qualitatives. De nombreux cas d'usages sont confrontés à cette même problématique relative aux données mixtes.

Comment procéder lorsque ces deux types de variables sont mélangés ? En pratique, nous pouvons proposer deux approches. La première est la binarisation des variables catégorielles, ce qui créé une augmentation de l'espace de travail (méthode choisie par les GLMs ou par l'AFDM par exemple). La seconde est l'utilisation d'une métrique différente de la distance euclidienne pour transformer la base initiale en une matrice de dissimilarité, qui va donner un score de ressemblance entre chaque individu. Cette démarche se retrouve très souvent dans l'analyse du génome, où des notions de distances entre gênes sont utilisées. [18]

Revenons sur le cas du traitement des variables catégorielles par un GLM. Les GLMs permettent cette option, en binarisant la variable qualitative, entraînant alors l'augmentation des coefficients estimés en p-modalités. Cette option est préférable pour capter une évolution par tranche d'âge par exemple. Néanmoins, à plus grande échelle, cela peut entrainer **un problème de dimension**, appelé fléau de la dimension. En effet, en plus d'augmenter le nombre de paramètres à estimer, la dimension du problème se retrouve augmentée. Il faut alors plus de données pour représenter les segments vides de l'espace de travail ou procéder à des regroupements.

Pour des problématiques de *clustering*, la binarisation ou la création d'une variable polytomique, va créer des sauts dans nos distances entre les individus, en plus d'agrandir la taille de la matrice de *design*.

Il existe différentes méthodes pour analyser et modifier des données quantitatives et qualitatives, nous allons étudier les méthodes qui seront exploitées au sein de ce mémoire.

5.2 La distance de Gower

Cette distance introduite par Gower [5] en 1971 permet de calculer la distance entre deux individus possédant des attributs mixtes, c'est-à-dire un mélange de modalités quantitatives et qualitatives. La notion de matrice de dissimilarité est utilisée pour mesurer la différence entre nos véhicules.

Soit d_{ij} la dissimilarité partielle comprise entre [0,1] et f la variable dont on calcule le

coefficient intermédiaire. Par définition, la distance de Gower est la moyenne des dissimilarités partielles. Cette quantité se calcule en fonction du type de la variable (f) :

— Pour une variable de nature quantitative :

$$d_{ij}^{(f)} = \frac{|x_i^f - x_j^f|}{max(x^f) - min(x^f)}$$

— Pour une variable de nature qualitative :

$$d_{ij}^{(f)} = \mathbf{1}_{(x_i^f = x_j^f)}$$

La distance de Gower entre deux individus caractérisés par p variables se calcule à l'aide de la formule suivante :

$$d(x_i, x_j) = \frac{1}{p} \sum_{f=1}^{p} d_{ij}^{(f)}$$

Pour calculer la matrice de dissimilarité, nous utilisons la fonction *daisy* du package R **cluster**. Cette fonction permet d'introduire un vecteur de poids pour chaque variable. Ainsi, l'utilisateur peut affecter plus ou moins d'importances à une variable. La distance de Gower devient alors :

$$d(x_i, x_j) = \frac{\sum_{f=1}^{p} \omega_f \times d_{ij}^{(f)}}{\sum_{f=1}^{p} \omega_f}$$

Plusieurs méthodologies seront explorées pour déterminer un paramètre optimal des poids affectés aux variables dans le cas pratique.
Pour illustrer l'impact que cette métrique peut avoir dans une problématique de *clustering*, nous calculons deux matrices de dissimilarité.

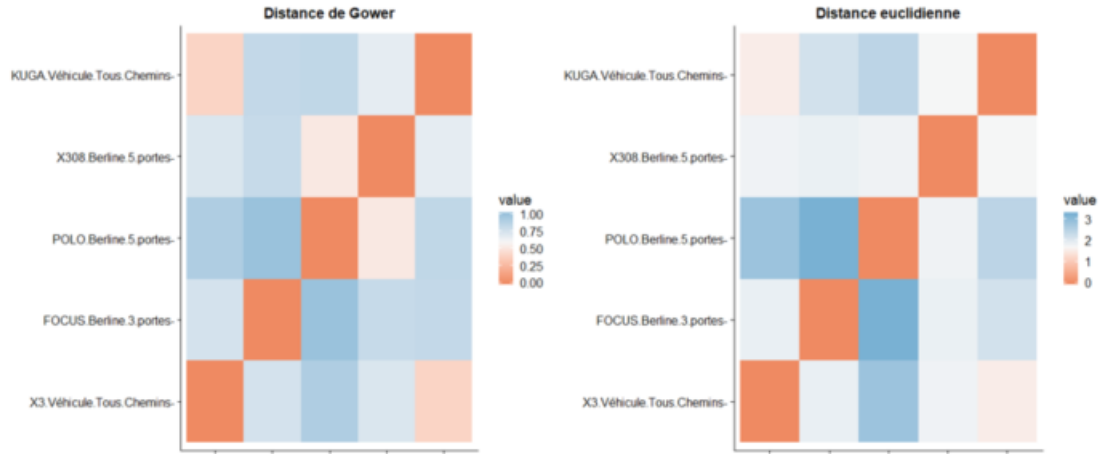

FIGURE 5.1 – Comparaison de la distance de Gower et de la distance Euclienne sur quelques véhicules.

Sur la Figure 5.1 a été simulé un exemple avec 5 véhicules choisis aléatoirement dans la base de véhicule à disposition. Parmi les variables constituant cette base, nous sélectionnons une variable numérique : la puissance du véhicule et deux variables catégorielles : le modèle du véhicule et le type de carrosserie. Les deux mesures ne renvoient pas les mêmes résultats. Ainsi, le *one-hot encoding*[1] a un impact sur la matrice de dissimilarité. Par conséquent, un

1. Le one-hot encoding est le terme se référant à la binarisation d'une variable catégorielle en plusieurs indicatrices

algorithme de *clustering* va renvoyer deux résultats totalement différents. Si en plus le nombre de variables quantitatives et qualitatives augmente, nous serons confrontés au problème du fléau de la dimension.

5.3 L'algorithme PAM : *Partitioning Around Medoids*

L'algorithme PAM, aussi appelé k-medoids, est une version plus robuste de l'algorithme *K-Means*. Cet algorithme itératif permet d'affecter à chaque observation de notre jeu de données un groupe à l'itération i. Le nombre de groupes est un paramètre à optimiser par l'utilisateur. L'objectif de l'algorithme est de trouver pour chaque individu le groupe lui correspondant le plus selon un critère illustré par la suite.

L'avantage de PAM est qu'il se généralise pour une matrice de dissimilarité alors que le k-Means fonctionne uniquement pour des données numériques. Ces deux algorithmes appartiennent à la famille des *clusterings* par partitionnement, en opposition au *clustering* hiérarchique (un exemple très connu et le *clustering* hiérarchique ascendant).

Avant de rentrer dans des considérations algorithmiques, il faut définir les différentes notions suscitées par cette méthode.

5.3.1 Notion d'inertie

Théorème 1. *Théorème d'Huygens - Relation fondamentale*

Soit $G' = e_i, i = (1, ..., n)$ un groupe d'individus, de centre de gravité G, que nous souhaitons partitionner en K classes $(G'_1, ..., G'_K)$ de centre de gravité $(G_1, ..., G_K)$. Soit d une distance entre nos individus. [2]

Alors

$$\underbrace{\sum_{i=1}^{n} d^2(x_i, G)}_{\textit{Inertie totale}} = \underbrace{\sum_{k=1}^{K} n_k d^2(G_k, G)}_{\textit{Inertie interclasse}} + \underbrace{\sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(x_i, G_k)}_{\textit{Inertie intraclasse}}$$

L'inertie intraclasse est l'indicateur de concentration des classes, tandis que l'inertie interclasse représente la séparation des classes.

La classification pourrait avoir comme objectif de minimiser l'inertie intraclasse. Nous retrouvons alors l'objectif de l'algorithme PAM : créer des groupes avec un fort degré de similitude entre les individus, tout en vérifiant que ces individus sont différents des individus appartenant aux autres groupes.

5.3.2 Notion de médoïde

Contrairement à un barycentre qui est un point artificiel, un médoïde est un point réel de la classe G'_k. Le médoïde est défini comme le point qui minimise sa distance avec l'ensemble de tous les points, c'est à dire :

$$M = \underset{m \in 1,...,n}{\arg\min} \sum_{i=1}^{n} d(x_i, m)$$

2. Euclidienne dans le cas numérique, Gower dans notre contexte

5.3.3 Comparaison avec l'algorithme k-Means

L'algorithme *k-Means* va utiliser le barycentre (μ) au lieu d'utiliser la notion de médoide. Le critère suivant doit être minimisé :

$$\arg\min_{G'} \sum_{i=1}^{K} \sum_{x_j \in G'_i} d^2(x_j, \mu_i)$$

L'algorithme PAM va minimiser le critère suivant :

$$\arg\min_{G'} \sum_{i=1}^{K} \sum_{x_j \in G'_i} d^2(x_j, M_i)$$

5.3.4 Description de l'algorithme PAM

Nous possédons en entrée des observations représentées par une matrice de distance et k désigne le nombre de *clusters*.

L'algorithme PAM, comme illustré dans l'article de L. Kaufman et J.Rousseeuw [10], s'article autour de deux phases, une première phase de construction (*build*) et une seconde phase d'échange (*swap*).

1/ *BUILD*

A/ Choisir k individus aléatoirement pour devenir les médoïdes. (sauf si ces individus ont été définis par l'utilisateur)

B/ Assigner à chaque individu son médoïde le plus proche.

2/ *SWAP*

C/ Pour chaque groupe, chercher si un individu dans le groupe diminue la moyenne des dissimilarités. Si c'est le cas, sélectionner l'individu qui réduit le plus cette quantité pour ce groupe et l'affecter comme médoïde du groupe.

D/ Si au moins un médoïde a changé, retour à l'étape B, sinon l'algorithme est terminé.

L'initialisation de l'algorithme peut influencer la convergence finale de l'algorithme. Ainsi, nous devrons vérifier la robustesse de la classification en réalisant plusieurs *clusterings* pour vérifier que les groupes ne changent pas.

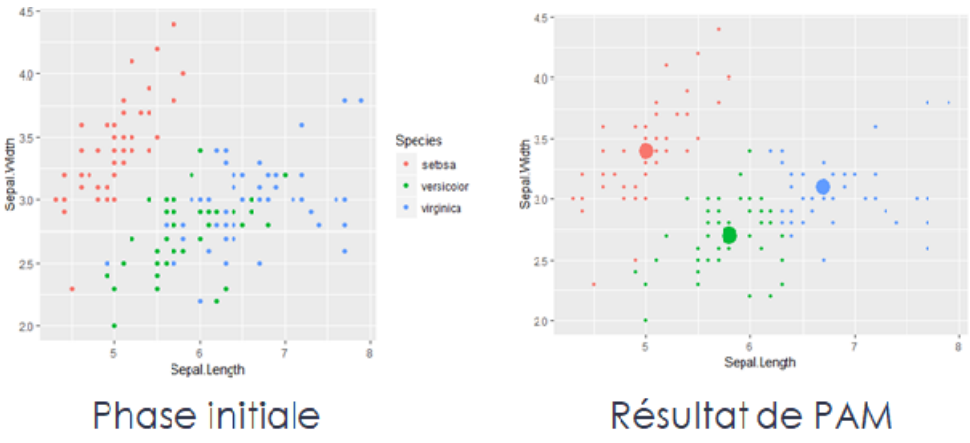

FIGURE 5.2 – Application de l'algorithme PAM sur le jeu de données *iris*

Le jeu de données *iris* est un jeu de données multivariées comprenant 150 échantillons de trois espèces d'iris (*setosa*, *virginica* et *versicolor*). Nous avons segmenté les variables relatives à la longueur (*length*) et la largeur (*width*) des pétales pour retrouver la structure initiale. PAM a identifié 3 médoïdes (Figure 5.2) et permet d'approcher la segmentation observable empiriquement.

5.3.5 Statistiques de validation de PAM

Le coefficient de silhouette [17] est retenu dans notre étude pour valider ou réfuter le résultat de notre *clustering*. C'est une mesure de qualité d'une partition, permettant de savoir si nos clusters classent correctement nos individus.

Ce coefficient se définit pour un individu i appartenant au groupe G'_k. La première quantité à calculer est la distance moyenne du point à son groupe :

$$a_i = \frac{1}{Card(G'_k)} \sum_{j \in G'_k, i \neq j} d(x_i, x_j)$$

La seconde quantité est la distance moyenne du point à son groupe voisin le plus proche :

$$b_i = \min_{k' \neq k} \frac{1}{Card(G'_{k'})} \sum_{j \in G'_{k'}} d(x_i, x_j)$$

Le coefficient de silhouette du point i permet de savoir si le point i est éloigné des clusters voisins, plus ce coefficient est proche de 1, plus il est éloigné. Ce coefficient est donné par :

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Pour vérifier la globalité du *clustering*, nous pouvons moyenner les coefficients de silhouette sur l'ensemble des individus, ce qui nous donne :

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{Card(G'(k))} \sum_{i \in G'_k} s(i)$$

5.3.6 Version bootstrap de PAM pour segmenter une base avec de nombreux individus. (CLARA)

CLARA (*Clustering on LARge Applications*) est une version *bootstrap* de l'algorithme PAM introduit précédemment. L'algorithme PAM n'est pas adapté pour segmenter des jeux de données possédant beaucoup d'individus.

CLARA se repose sur une approche par échantillonnage. Au lieu de se focaliser sur la totalité du jeu de données, CLARA tire un sous-échantillon et applique PAM afin de générer des médoïdes sur ce sous-ensemble de la base initiale. Ensuite, chaque individu n'appartenant pas à un sous-ensemble est affecté au médoïde le plus proche. La totalité du jeu de données possède un groupe à ce stade. Une mesure de la qualité du clistering est obtenue en calculant la dissimilarité moyenne entre les individus de la base et son médoïde le représentant.

Ces étapes sont répétées m fois (m=5 dans notre cas) et nous choisissons le *clustering* minimisant la mesure de qualité définit précédemment.

En résumé

Le contexte des données mixtes rend le travail plus difficile pour l'interprétation des groupes. Le choix d'une métrique sur une autre influe sur les résultats observés. La distance de Gower permet de combiner la présence des variables quantitatives et qualitatives dans des problématiques de *clustering*.

L'algorithme PAM s'adapte à notre problématique de données mixtes en supportant une matrice de dissimilarité en entrée. Nous l'avons utilisé pour réaliser une classification des véhicules dans un cadre non supervisé. A partir des variables relatives aux véhicules, nous avons calculé la distance entre chaque véhicule pour ensuite créer les groupes de véhicules et assigner à chaque individu un groupe.

La vérification des résultats est réalisée en réalisant un graphique de silhouette, permettant de savoir si un véhicule est classé dans le bon groupe.

Dans le cas où le nombre d'individus est grand, l'utilisation de CLARA est privilégiée pour résoudre les problèmes de complexité relative à PAM, en utilisant des échantillons issus de la base initiale

Chapitre 6

Les méthodes de réduction de dimension

Les données exploitées en assurance possèdent de nombreuses caractéristiques et il est parfois préférable de réduire la dimension du problème pour pouvoir les observer. Réduire la dimension par une ACP (Analyse en Composantes Principales) permet par exemple de voir si des variables sont redondantes, ou bien si une combinaison de variables contribue à expliquer un phénomène.

Néanmoins, l'ACP représente des variables quantitatives qui sont représentées par des vecteurs de longueur 1. Tandis que l'ACM (Analyse en Composantes Multiple) va représenter une variable qualitative par un nuage de points. L'idée est d'avoir une méthode permettant de représenter des données mixtes dans un nouvel espace, comme le réalise les deux précédents algorithmes, tout en accordant autant de poids aux variables quantitatives que qualitatives.

L'assemblage des points de vue de l'ACP, de l'ACM et de l'AFM (Analyse Factorielle Multiple) a permis de créer une nouvelle méthodologie : l'AFDM (Analyse Factorielle de Données Mixtes). Cet algorithme permet de capter **la structure globale** de la base de données et de réaliser une carte des véhicules.

D'autres méthodologies ont été développées récemment, comme l'algorithme t-SNE qui permet de projeter une base de données ou une matrice de dissimilarité sur un espace en 2D ou en 3D, pour être facilement visualisable. Cet algorithme a l'avantage de garder la **structure locale** de la base et d'être un excellent outil de visualisation pour vérifier le caractère local des clusters. Nous illustrerons aussi une version des cartes de Kohonen adaptées à notre problématique de données mixtes représentées par le biais d'une matrice de dissimilarité.
L'intérêt de ces outils réside dans l'exploration des données et permet de capter des structures que les statistiques descriptives ne pourraient pas capter à elles seules.

Nous allons présenter les algorithmes suivants dans ce chapitre :

— Analyse Factorielle de Données Mixtes (AFDM)

— *t-distributed Stochastic Neighbor Embedding* (t-SNE)

— Les cartes de Kohonen (SOM)

6.1 Analyse Factorielle des Données Mixtes

Nous rappellerons les enjeux principaux de l'algorithme AFDM, l'article de J.Pages reprenant les principales étapes pour la mise en place de cet algorithme. [15]

L'AFDM est préférable lorsque nous souhaitons garder nos variables quantitatives, dans plusieurs cas :

— Le nombre d'individus est faible (ce n'est pas le cas dans une problématique de tarification)

— **Le nombre de variables qualitatives est faible par rapport au nombre de variables quantitatives**

Nous sommes dans la seconde situation, pour notre problématique de classification des véhicules. En effet, nous avons beaucoup de données relatives à la taille du véhicule, à la puissance, à leurs prix. Il est alors préférable de ne pas séparer le problème quantitatif et qualitatif.

6.1.1 Hypothèses du modèle

Nous supposons que nous avons :

— K_1 variables quantitatives, qui sont **centrées-réduites**.

— Q variables qualitatives, chacune notée K_q avec q modalités soit

$$\sum_q K_q = K_2$$

— Soit $K = K_1 + K_2$ le nombre total de variables quantitatives et de variables indicatrices

Poids des individus
Chaque individu se voit un poids D assigné, défini par :

$$D(x_i, x_j) = \begin{cases} p_i \ si \ i = j \\ 0 \ sinon \end{cases}$$

Généralement, avec Id_n la matrice identité de dimension n :

$$D = \frac{1}{n} Id_n$$

Principe de l'AFDM
L'algorithme repose sur la maximisation du critère W suivant par rapport à la direction ν dans R^I, l'espace des individus :

$$W = \sum_{k \in K_1} r^2(k, \nu) + \sum_{q \in Q} \eta^2(q, \nu)$$

En d'autres termes, nous souhaitons résoudre le problème suivant :

$$\max_{\nu \in R^I} W$$

— Les variables quantitatives K étant réduites, nous avons $r(k, \nu) = cos(\theta_{k\nu})$ où $\theta_{k\nu}$ est l'angle entre k et ν .

— ν est centré : $\eta^2(q, \nu) = cos^2(\theta_{q\nu})$ où $\theta_{q\nu}$ est l'angle entre ν et sa projection sur l'espace E_q

Nous obtenons alors

$$W = \sum_{k \in K_1} cos^2(\theta_{k\nu}) + \sum_{q \in Q} cos^2(\theta_{q\nu})$$

6.1.2 Représentation des individus dans R^K

La distance entre un individu i et j s'écrit par :

$$d^2(x_i, x_j) = \sum_{k \in K_1} (x_{ik} - x_{jk})^2 - \sum_{q \in Q} \sum_{k \in K_q} \frac{1}{p_{k_q}} (x_{ik_q} - x_{jk_q})^2$$

Le résultat qui nous intéressera est la distance de l'individu i par rapport à l'origine O. Grâce à l'hypothèse des variables centrées, nous obtenons :

$$d^2(x_i, O) = \sum_{k \in K_1} x_{ik}^2 - \sum_{q \in Q} \sum_{k \in K_q} (\frac{x_{ik_q}}{p_{k_q}} - \sqrt{p_{k_q}})^2$$

La première quantité est en relation avec les variables quantitatives, qui auront été mises préalablement à la même échelle, tandis que la seconde quantité fait référence aux variables qualitatives, qui va introduire des sauts dans la distance.

6.1.3 Interprétation de l'AFDM

Comme l'ACP ou l'ACM, l'AFDM permet d'explorer les différents espaces à l'aide de graphique. Nous pourrons alors représenter les trois éléments suivants :

— Le nuage des individus par sa projection sur ses axes d'inerties

— Les variables quantitatives par leur coefficient de corrélation

— Les modalités des variables qualitatives par les centres de gravité des individus correspondants.

6.1.4 Utilisation de l'AFDM dans notre problématique de réalisation d'un véhiculier

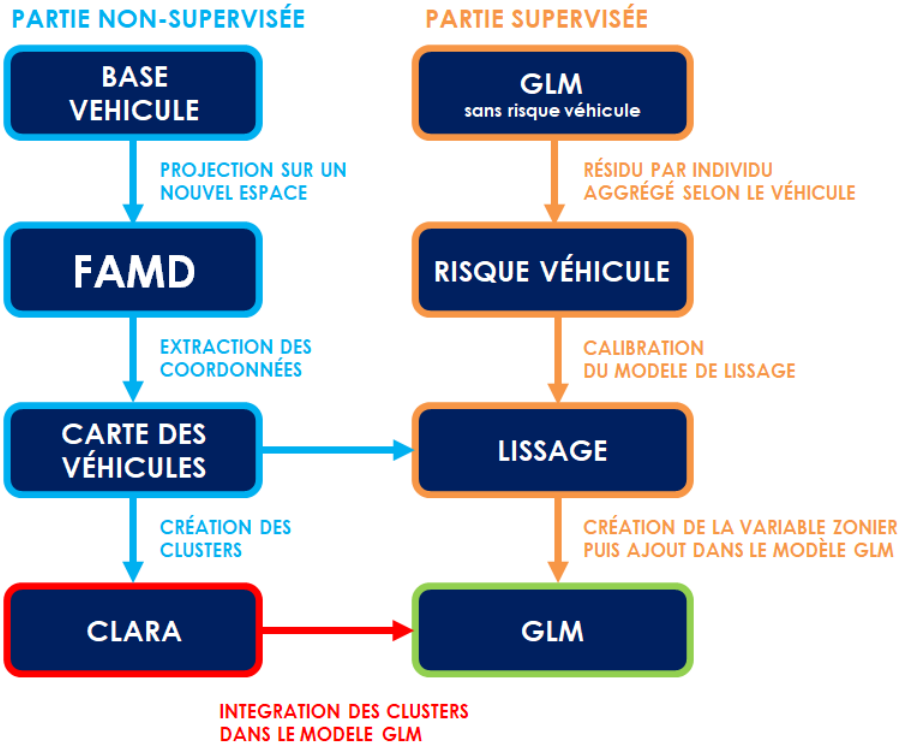

FIGURE 6.1 – Utilisation de la projection AFDM dans notre problématique de réalisation d'un véhiculier.

En résumé

L'Analyse Factorielle de Données Mixtes joue un double rôle dans notre approche. A partir d'une base véhicule comportant les caractéristiques de chaque véhicule, nous pourrons :

- **Créer un espace continu**, appelé dans ce mémoire **carte des véhicules**.
- **Explorer nos données** et détecter des poches tarifaires non observées dans nos analyses descriptives univariées.

La carte des véhicules est la sortie principale de l'AFDM et permettra de créer des premiers groupes de véhicules à l'aide de CLARA, une version *bootstrap* de l'algorithme PAM. Cette carte est aussi utilisée dans la partie supervisée afin de lisser notre risque véhicule résiduel extrait par la modélisation GLM sans risque véhicule.

6.2 L'algorithme t-SNE : *t-distributed stochastic neighbor embedding*

6.2.1 Présentation de l'algorithme

L'algorithme t-SNE est une technique de réduction de dimension développée en 2018 par G. Hinton et L. van der Maater [13]. Cet algorithme permet de réduire des espaces avec de nombreuses variables pour les représenter dans un nouvel espace en 2D ou 3D.

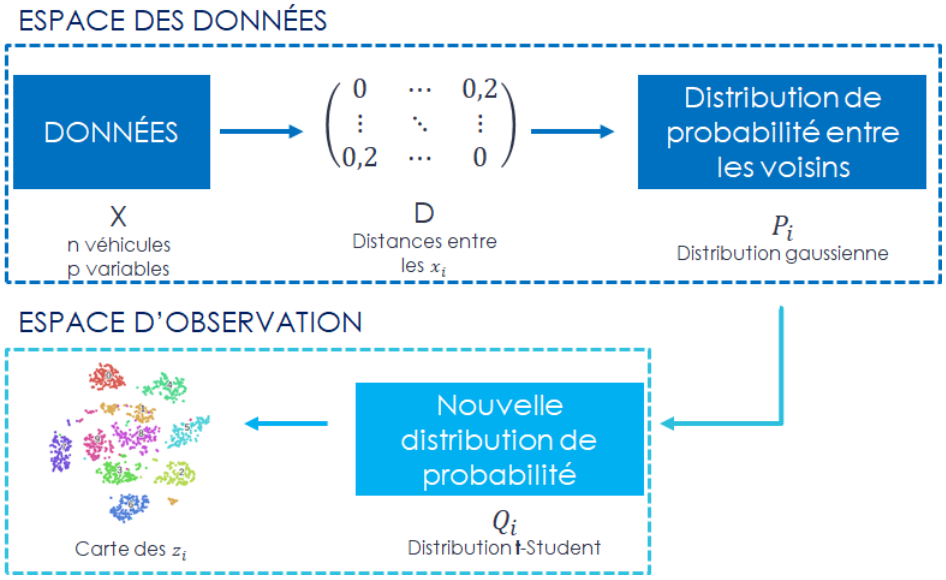

FIGURE 6.2 – Récapitulatif de la méthodologie t-SNE

Il faut fournir en entrée une matrice de distance D, la métrique par défaut est **la distance euclidienne**. Nous verrons dans la partie *clustering* comment spécifier une distance dans le cas de données mixtes. Nous nous plaçons alors dans le cas euclidien.

A partir d'une matrice de distance D entre les individus , nous calculons un score de similarité dans **l'espace initial** $p_{j|i}$. Ce score de similitude entre les points x_i et x_j peut-être vu comme la probabilité conditionnelle que x_i est voisin avec x_j si les voisins étaient pris proportionnellement à la densité sous une gaussienne centrée en x_i :

$$p_{j|i} = \frac{\exp(-||x_j - x_i||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Avec :

— σ_i est la variance d'une distribution normale centrée sur le point x_i.

— $p_{i|i} = 0$

Ainsi, si une partie de l'espace présente un amas de points, la valeur de σ_i sera plus petite et donc le score de similarité $p_{j|i}$ sera plus grand.

Dans l'**espace d'observation** (de dimension 2 ou 3), une distribution t de Student à un degré de liberté (soit une loi de Cauchy) est utilisée pour la distribution des distances entre les individus voisins :

$$q_{j|i} = \frac{\exp(-||z_i - z_j||^2)}{\sum_{k \neq i} \exp(-||z_i - z_k||^2)}$$

Encore une fois, $q_{i|i} = 0$ car nous souhaitons modéliser les similarités entre les pairs.

Si les points de la carte z_i et z_j modélisent correctement la similitude entre les points x_i et x_j, les quantités $p_{j|i}$ et $q_{j|i}$ seront quasiment égales. Une manière de mesurer la fidélité de la valeur de $q_{i|j}$ par rapport à $p_{i|j}$ est **la mesure de divergence de Kullback-Leibler** . L'algorithme t-SNE utilise une version améliorée de l'algorithme SNE original, qui consiste à minimiser la somme des divergences de Kullback-Leibler sur tous les points en utilisant une descente de gradient.

La fonction de coût C est donnée par :

$$C = \sum_i KL(P_i|Q_i) = \sum_i \sum_j p_{j|i} log(\frac{p_{j|i}}{q_{j|i}})$$

Où P_i désigne la distribution de probabilité entre tous les points sachant le point x_i et Q_i représente la distribution entre tous les points de la carte sachant z_i. Cette fonction de coût permet de capturer la structure locale des points de la carte.

Le paramètre de perplexité spécifié par l'utilisateur permet d'estimer le paramètre σ_i. Ce paramètre est tel que la distribution P_i va avoir une perplexité fixe (qui est un paramètre de l'utilisateur).

La perplexité est définie par $Perp(P_i) = 2^{H(P_i)}$ où $H(P_i)$ est l'entropie de Shannon, qui correspond à la quantité d'information contenue :

$$H(P_i) = -\sum_j p_{j|i} log_2(p_{j|i})$$

La perplexité s'interprète alors comme une mesure de lissage sur le nombre de voisins. Il faudra faire varier ce paramètre pour vérifier la robustesse des résultats donnée par l'algorithme t-SNE. Les performances de l'algorithme sont très dépendantes de ce paramètre, il faudra le faire varier sur un intervalle compris entre 5 et 50 d'après l'auteur.

La minimisation de la fonction de coût est réalisée par une descente de gradient. Le gradient a la forme suivante :

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

La descente de gradient est initialisée en échantillonnant les points avec une distribution gaussienne avec une faible variance et centrée à l'origine de la carte. Pour accélérer l'optimisation et pour éviter les problématiques de minimum local, un terme supplémentaire est rajouté au gradient, appelé *momentum*.

La mise à jour du gradient est donnée à l'étape t par :

$$\zeta^{(t)} = \zeta^{(t-1)} + \eta\frac{\delta C}{\delta \zeta} + \alpha(t)(\zeta^{(t-1)} - \zeta^{(t-2)})$$

Où $\zeta^{(t)}$ désigne la solution à l'itération t, η correspond au taux d'apprentissage et $\alpha(t)$ représente le *momentum* à l'itération t.

6.2.2 Utilisation dans des contextes hors assurance

L'algorithme t-SNE, en complément d'une ACP pour retirer du bruit ou diminuer l'espace du travail, a été beaucoup utilisé dans des domaines comme la biologie [18] dans des problématiques de détection de gênes et a prouvé son efficacité.

La possibilité de réduire un espace de très grande dimension à un espace facilement visualisable est un atout majeur, permettant de détecter des structures dans les données. Néanmoins, la construction de la matrice de dissimilarité reste une étape primordiale et la pondération qui est attribuée à chaque variable. Celle-ci impactera plus ou moins les résultats du t-SNE dans notre cas.

D'autres applications intéressantes, comme la détection d'image et la reconnaissance d'écriture, ont montré la performance de cet algorithme pour regrouper des éléments. Son interprétation est sensible car nous pouvons avoir des résultats différents en fonction de la perplexité. Ainsi, de nombreux tests sont à effectuer pour éviter d'aller dans une mauvaise direction.

6.2.3 Utilisation de t-SNE pour la visualisation du *clustering* donnée par PAM

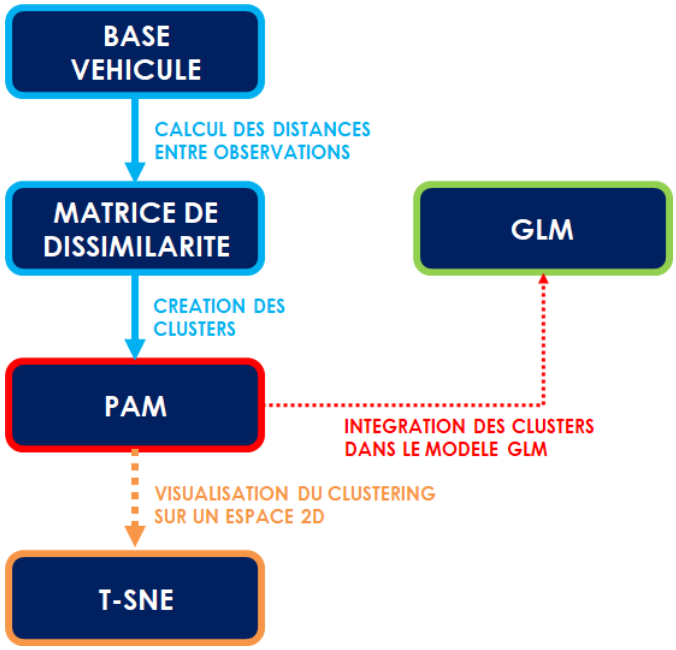

FIGURE 6.3 – Utilisation de la projection t-SNE dans notre problématique de réalisation d'un véhiculier.

En résumé

La projection réalisée par t-SNE permet de vérifier à partir de la matrice de dissimilarité issue des caractéristiques des véhicules, le caractère local des groupes de véhicule réalisés par l'algorithme PAM.

6.3 Les cartes de Kohonen

Les cartes auto-organisatrices [11], créées par Teuvo Kohonen, permettent de réaliser une segmentation non supervisée d'un jeu de données, comme le permet le *K-means* dans le cas de données numériques. Cependant, cette méthode permet aussi de réaliser une réduction de dimension en visualisant une base initiale sur un nouvel espace topologique.

La répartition des observations sur la grille n'est pas calculée de manière directe, mais elle est approchée étape par étape en minimisant une fonction de coût. Enfin, l'espace topologique créé par cette méthode permet d'avoir une carte dont la topologie est proche des données d'origines.

6.3.1 Le cas classique des cartes de Kohonen : le cas numérique

Soit n observations dans $\mathbb{R}^d$ et X la matrice représentant nos données, de dimension n$\times$ d. L'algorithme va projeter les n observations sur une grille de dimension 2, qui est composée de U neurones (1, ... ,U) et d'une distance entre neurones. Chaque neurone possède un représentant dans l'espace de départ, appelé prototype (noté p_i). Ces représentants doivent être initialisés au début de l'algorithme. La projection f : $x_i \mapsto f(x_i) \in (1, ..., U)$ est effectuée de manière à préserver **la topologie de l'espace initiale**. Cette projection est approchée par itération, en minimisant le critère suivant, appelé **énergie** :

$$\varepsilon = \sum_{i=1}^{n} \sum_{u=1}^{U} h_t(f(x_i), u) ||x_i - p_u||^2$$

Où h est **une fonction de voisinage** décroissante en fonction de la distance entre les neurones sur la grille.

Les variables à optimiser pour minimiser l'énergie sont les suivantes :

— La projection de x_i à l'étape t, notée $f^t(x_i)$

$$f^t(x_i) \leftarrow \underset{u \in (1,...,U)}{\arg\min} ||x_i - p_u^{t-1}||$$

— La mise à jour des prototypes à l'étape t, p_u^t

La version stochastique de l'algorithme peut-être retrouvée dans le package R **SOMBrero**.

6.3.2 Visualisation des *clusters* et de la grille

L'avantage majeur des cartes de Kohonen, comme la méthode de l'AFDM, est qu'elle permet de visualiser la contribution de chaque variable, qu'elles soient quantitatives avec une *heatmap*, ou bien qualitatives à l'aide de diagramme en boite. Il est possible de savoir comment notre espace topologique a été créé à partir des éléments explicatifs de départ.

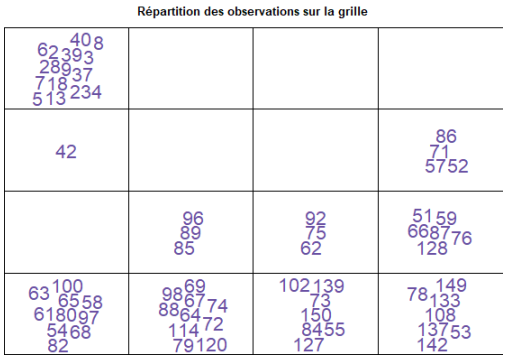

FIGURE 6.4 – Observations du jeu de données *iris* réparties sur la grille

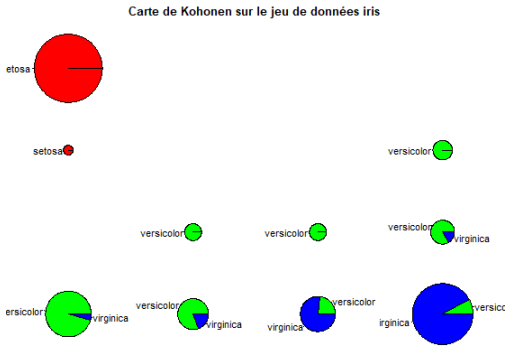

FIGURE 6.5 – Répartition des 3 espèces de la base *iris* sur la grille

Pour apporter une première observation de la méthodologie, nous avons réalisé une carte de Kohonen sur le jeu de données *iris*. La topologie de la grille (Figure 6.4) est carrée et possède une fonction de voisinage h de type gaussienne et la distance utilisée entre les individus est la distance euclidienne, où les variables numériques ont été centrées-réduites avant modélisation. Nous pouvons voir (Figure 6.5) que la grille segmente correctement l'espèce d'iris *setosa* mais à des difficultés à différencier les deux espèces restantes, *virginica* et *versicolor*.

6.3.3 Adaptation pour une matrice de dissimilarité

Nous supposons que les données sont décrites par une matrice de dissimilarité D (calculée à l'aide de la distance de Gower par exemple). Il est supposé que les prototypes peuvent être écrits symboliquement comme des **combinaisons convexes** des observations.
Soit $\gamma_{u,i} \geq 0$, $\sum_{i=1}^{n} \gamma_{u,i} = 1$, nous avons avec l'hypothèse précédente :

$$\forall u \in 1, ..., U, p_u = \sum_{i=1}^{n} \gamma_{u,i} x_i$$

Dans le cas où la matrice de dissimilarité D est symétrique et définie positive et que nous notons $\gamma_u = (\gamma_{u,i})_{i \in (1,\dots,n)}$:

$$||x_i - p_u||^2 = (D\gamma_u)_i - \frac{1}{2}\gamma_u^T D\gamma_u$$

Cette alternative est celle qui se rapproche le plus de la méthode euclidienne dans le cas numérique. Les modifications de l'algorithme s'opèrent donc seulement sur la notion de distance entre les individus.

6.3.4 Validation du clustering

Le package **SOMBrero** [2] offre la possibilité de valider le *clustering* grâce à deux critères

— **L'erreur topographique** mesure la continuité de la grille par rapport à l'espace d'origine, en comptant le nombre de fois que le second élément le plus proche de la grille se trouve dans le voisinage de la meilleure unité choisie par l'algorithme. Ainsi, une erreur topographique proche de 0 signifie que tous les seconds individus se trouvent dans le voisinage de l'unité choisie et que la topologie originale de la base de données a été préservée.

— **L'erreur de quantification** quantifie la qualité des clusters en calculant la quantité :

$$\frac{1}{n}\sum_{u=1}^{U}\sum_{f(x_i)=u}[(D\gamma_u)_i - \frac{1}{2}\gamma_u^T D\gamma_u]$$

Néanmoins, le second indicateur est une fonction décroissante de la taille de la carte. Ainsi, plus la grille créée est grande, plus la qualité des clusters est bonne. Il faut fiabiliser cette mesure à l'aide d'un critère graphique, où il serait souhaitable que la carte soit remplie en évitant des zones topologiques où aucun individu n'est présent.

Enfin, pour les données de dissimilarité, une ANOVA de dissimilarité est réalisée. L'idée de cette ANOVA est d'utiliser un dérivé de la statistique du F-Ratio de Fisher pour permettre de tester l'hypothèse H_0 où il n'y aurait pas de différence entre les groupes. Ainsi, plus le F-Ratio est élevé, plus nous sommes dans la capacité de rejeter l'hypothèse H_0 et donc de conclure qu'il y a une différence entre nos groupes [1].

6.3.5 Utilisation des cartes de Kohonen dans notre problématique

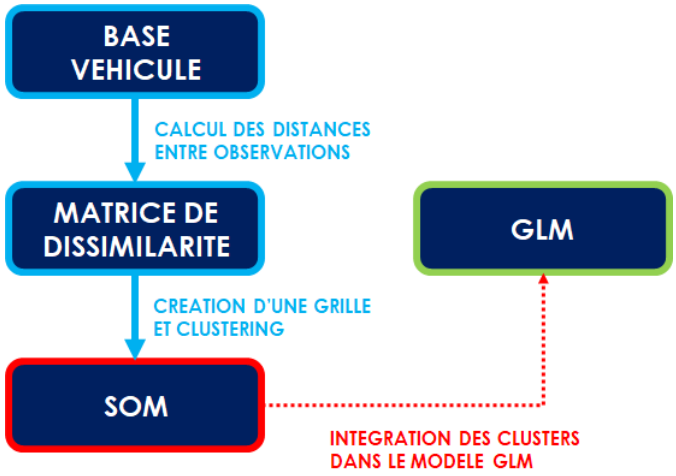

FIGURE 6.6 – Utilisation des cartes de Kohonen dans notre problématique de réalisation d'un véhiculier.

En résumé

Les cartes de Kohonen permettent à partir d'une matrice de dissimilarité (calculée à l'aide de l'indice de Gower dans notre situation) de créer une grille dont la taille est spécifiée par l'utilisateur.

Cette grille est composée de neurones, et chaque neurone se voit attribuer un représentant appelé prototype. Ainsi, chaque véhicule appartiendra à un groupe issu d'un prototype.

L'avantage de la carte réside dans la visualisation des données par l'aspect réduction de dimension associé. Il est possible d'obtenir des statistiques descriptives pour les variables caractérisant les véhicules pour chaque neurone de la grille.

Chapitre 7

Les méthodes de lissage géospatial

7.1 Motivation pour l'utilisation d'une méthode d'interpolation spatiale

Les méthodes de lissage sont très souvent utilisées en assurance pour de nombreuses problématiques afin de corriger les irrégularités qu'un actuaire observe en pratique. Ainsi, dans le cadre de l'élaboration d'un zonier, le lissage du résidu est une alternative permettant d'améliorer la compréhension et la fiabilité du modèle.

Dans notre situation, nous possédons **une carte des véhicules continue** réalisée à l'aide d'une AFDM. Une fois que l'effet véhicule a été capté à l'aide d'un GLM, nous devons lisser ce résidu pour gommer les signaux que nous n'avons pas pu capter dans notre modélisation du fait de l'anonymisation de notre base (Sexe du conducteur) ou des erreurs de spécification du modèle. Ainsi, après lissage, nous garderons un résidu comportant une forte part du signal portée par les véhicules.

Nous décidons d'utiliser le Krigeage pour lisser notre résidu et capter la globalité de notre effet véhicule. Cette méthode a été retenue pour tester une nouvelle approche différente du lissage par crédibilité, employée par F. Hébert pour la réalisation d'un zonier [8]. Ce lissage corrige le résidu en fonction de la distance et de son exposition, mais exige de calculer la distance entre plusieurs individus, opération très lourde en termes de calcul, d'autant plus que la notion de voisinage entre nos véhicules est tangible du fait de la concentration des individus.

7.2 Théorie autour du Krigeage

Le Krigeage est une méthode d'interpolation spatiale utilisée dans de nombreux domaines, allant de la météorologie jusqu'à l'électromagnétisme. Cette méthode est adaptée aux corrélations spatiales ainsi qu'aux tendances particulières des données. C'est pour cela que cette méthode a prouvé son efficacité première pour la prospection minière ou en géologie. [6]

L'interpolation spatiale consiste à estimer une fonction $z(\mathbf{x})$, où $\mathbf{x} = (x, y)$, en un point x_p du plan à partir des valeurs observées de F autour de ce même point. Notons m le nombre de points autour de x_p, alors :

$$z(x_p) = a + \sum_{i=1}^{m} \omega_i \times z(x_i)$$

Les poids w_i sont choisis à l'aide de la corrélation spatiale entre les points, calculés à l'aide d'un variogramme que nous définirons dans la partie suivante. Ces poids permettent d'obtenir une prévision non biaisée et de variance minimale.

Le modèle du Krigeage comporte une structure similaire au modèle de régression classique, mais les erreurs possèdent cette fois-ci une dépendance spatiale :

$$Z(x_p) = \mu(x_p) + \delta(x_p)$$

— $\mu(.)$ est la structure déterministe

— $\delta(.)$ est une fonction aléatoire stationnaire, d'espérance nulle et de structure de dépendance connue.

La spécification de la forme de la tendance $\mu(.)$ permet de préciser le type de Krigeage utilisé, nous en distinguons trois :

— Le Krigeage simple : $\mu(x) = m$ (constante connue)

— Le Krigeage ordinaire : $\mu(x) = \mu$ (constante inconnue)

— Le Krigeage universel : $\mu(x) = \sum_{i=0}^{p} \beta_j f_j(x)$ (combinaison linéaire de fonctions de la position x)

La structure de dépendance $\delta(.)$ est déterminée en pratique à partir des données à l'aide d'un variogramme. La détermination des poids ω_i permettent alors de spécifier le modèle obtenu et ceux-ci respectent la condition de non-biais : $E[\hat{Z(x_p} - Z(x_p)] = 0$ et la condition de minimisation de la variance de l'erreur de prévision $V[\hat{Z(x_p)} - Z(x_p)]$.

7.2.1 La fonction de covariance et le variogramme

Le semi-variogramme empirique est calculé pour les $n(h)$ points de coordonnées (x_i, y_i) de la manière suivante :

$$\hat{\gamma(h)} = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (Z(x_i + h) - Z(x_i))^2$$

Avec $n(h) = Card(\{(x_i, x_j) / \ \ |x_i - x_j| \sim h\})$

Pour rappel, dans le cas d'un processus stationnaire, nous avons :

— $Cov(\delta(x), \delta(y)) = \gamma(x - y)$

— $Cov(\delta(x), \delta(x + h)) = \gamma(h)$

Le Krigeage consiste à calculer les poids ω_i à l'aide des valeurs de la fonction $\gamma(h)$ correspondant aux m points choisis. Les poids intègrent par construction la corrélation spatiale entre les points.

Le choix et l'ajustement d'une fonction au semi-variogramme sont primordiaux pour la spécification du modèle de Krigeage. L'allure du variogramme permet de visualiser la relation des données et permet d'émettre des premières hypothèses sur le fonctionnement du modèle.

Pendant la phase d'apprentissage du modèle de Krigeage, une fonction de covariance doit-être paramétrée pour prendre en considération la structure de dépendance entre les points et l'hypothèse de stationnarité de ceux-ci. Celle-ci peut se déterminer en analysant le semi-variogramme réalisé précédemment. La famille de fonctions de covariances usuelles peut-être choisie en émettant des hypothèses sur la régularité du processus gaussien composant le modèle. Les paramètres de la fonction de covariance sont estimés par maximum de vraisemblance.

Nom	Expression	Classe
Gaussien $K_g(x,x')$	$\sigma^2 exp\left(-\frac{(x-x')^2}{2\theta^2}\right)$	C^∞
Matèrn 5/2 $K_{m_{\frac{5}{2}}}(x,x')$	$\sigma^2\left(1+\frac{\sqrt{5}\lvert x-x'\rvert}{\theta}+\frac{5\lvert x-x'\rvert}{3\theta^2}\right)\exp\left(-\sqrt{5}\frac{\lvert x-x'\rvert}{\theta}\right)$	C^2
Matèrn 3/2 $K_{m_{\frac{3}{2}}}(x,x')$	$\sigma^2\left(1+\frac{\sqrt{3}\lvert x-x'\rvert}{\theta}\right)\exp\left(-\sqrt{3}\frac{\lvert x-x'\rvert}{\theta}\right)$	C^1
Exponentiel $K_e(x,x')$	$\sigma^2 exp\left(-\frac{(x-x')}{\theta}\right)$	C^0
Power exponentiel $K_{pe}(x,x')$	$\sigma^2 exp\left(-\left(\frac{\lvert x-x'\rvert}{\theta}\right)^p\right)$ 0<p<2	C^0

FIGURE 7.1 – Exemple de fonctions de covariances pour la détermination de la structure de corrélation spatiale.

Nous retrouvons sur la Figure 7.1 les différentes fonctions de covariance pour la détermination de la structure spatiale. Cette structure se rapproche de la fonction de covariance en remarquant que $\gamma(h) = K(x, x+h)$ où $K(.)$ désigne la structure du noyau.

Dès lors que la structure du noyau est spécifiée à l'aide d'un semi-variogramme , nous avons tous les éléments nécessaires pour réaliser le modèle.

7.2.2 Analyse du variogramme

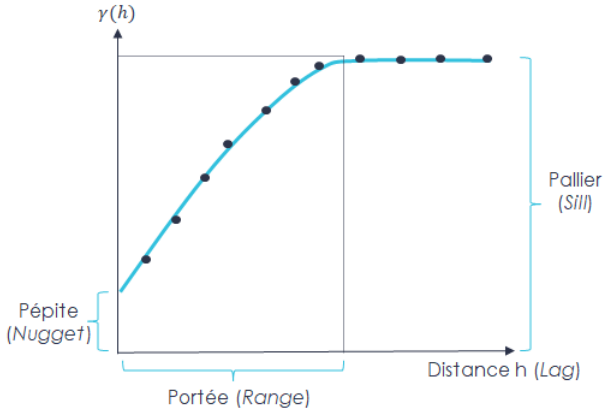

FIGURE 7.2 – L'allure d'un semi-variogramme et les éléments clés à détecter.

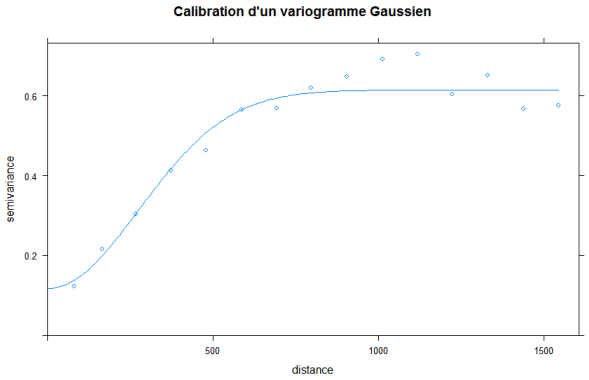

FIGURE 7.3 – Calibration d'un semi-variogramme gaussien sur le jeu de données *meuse*

Plusieurs éléments sont à détecter sur le semi-variogramme (Figure 7.2)

— La structure de noyau K(.) en comparant la structure du semi-variogramme avec les semi-variogrammes spécifiques à chaque structure. La Figure 7.3 présente l'allure d'un variogramme gaussien.

— L'effet pépite *(nugget)* est la limite du variogramme en zéro. Cet effet représente la variation entre deux individus infiniment proches et nous alerte sur la présence d'erreur de mesures entre les observations.

— Le pallier *(sill)* est la valeur à partir de laquelle $\gamma(h)$ devient constant avec l'évolution de h.

— La portée *(range)* est la distance à partir de laquelle $\gamma(h)$ atteint le palier. Au-délà de la portée, les observations possèdent une covariance nulle.

En résumé

Le Krigeage est une méthode d'interpolation spatiale captant la corrélation spatiale entre les points de notre jeu de données.

Chaque véhicule est associé à un risque véhicule résiduel capté par la modélisation GLM sans la variable véhicule et possède des coordonnées sur les trois premiers axes retenus par l'AFDM. Ces éléments permettront de calculer le semi-variogramme et de calibrer les paramètres nécessaires au modèle de Krigeage.

L'utilisation de ce lissage dans notre contexte réside dans **l'aspect continu** de notre carte des véhicules et dans sa capacité à généraliser des phénomènes observés.

Troisième partie

Cas pratique : réalisation d'un véhiculier pour un assureur

Cette partie illustre les outils théoriques introduit précédemment dans le cadre de la réalisation d'un véhiculier sur un portefeuille automobile d'un assureur partenaire.

Le véhiculier est réalisé pour chaque garantie (pertinente en termes de volumétrie) pour le coût moyen et la fréquence séparément. Ainsi, le véhiculier construit a pour objectif d'être un substitut à la classification SRA utilisée. Ce véhiculier capte précisément les risques inhérents à la garantie responsabilité civile matérielle. Cependant, il n'est pas forcément adapté pour d'autres garanties. Ainsi, nous avons décidé de réaliser des véhiculiers pour les garanties vol et dommage. **La garantie vol est présentée dans la partie non supervisée, tandis que la garantie dommage est modélisée dans la partie supervisée.**

Nous consacrerons une première partie au retraitement des données, qu'elles soient internes ou externes. Le retraitement de la base de données est la première étape primordiale pour adapter nos informations en portefeuille à la problématique.
Nous effectuerons ensuite une phase d'exploration de notre base de donnée relative à nos véhicules. Cette étape permet de réaliser plusieurs *clustering* de nos véhicules à l'aide des caractéristiques techniques à disposition.
Dans un second temps, nous extrairons l'effet véhicule des résidus par GLM afin de les modéliser par apprentissage statistique supervisé. Cette étape nous permet de mettre en valeur, s'il est présent, l'effet véhicule de notre base.

Finalement nous comparerons les différentes statistiques relatives à chaque véhiculier ainsi que les différentes méthodologies employées. Ces résultats permettront de vérifier la pertinence de l'intégration du véhiculier dans l'équation tarifaire.

La méthodologie mise en place dans ce mémoire se présente de la manière suivante. (Figure 7.4)

FIGURE 7.4 – Démarche globale du mémoire

Chapitre 8

Collecte, préparation et analyse des données

L'objectif de ce chapitre est de présenter le jeu de données utilisé au sein de l'étude et de présenter les méthodes mises en œuvre pour exploiter leurs potentiels : traitement des données manquantes, enrichissement de la base à l'aide de données externes et utilisation des données assureur.

Dans l'ère du ***Big Data*** et de l'**augmentation exponentielle** des données, nous utiliserons pleinement les données SRA fournies par l'assureur partenaire pour approcher le plus fidèlement possible le risque intrinsèque du véhicule.

En effet, les données SRA sont une mine d'or pour un assureur, car elles apportent déjà une information primordiale sur les véhicules du portefeuille, **la quantité de variables explicatives sur le véhicule étant non-négligeable**.

Dans la suite, nous distinguerons les données externes et internes par cette différenciation :

— **Interne** : Toutes les données issues de l'assureur, c'est-à-dire les informations clients, les informations sur le véhicule, sur la couverture de l'objet assuré, etc ...

— **Externe** : Toutes les informations permettant de compléter les informations collectées par l'assureur. Par exemple, le type mine [1] ou le CNIT permet d'identifier un véhicule.

Pour obtenir une modélisation cohérente, l'étape de pré-traitement est cruciale. Une mauvaise connaissance de la base d'étude entraînant forcément un modèle moins efficace. Un temps considérable a été accordé sur le retraitement de la base fournie par l'assureur partenaire.

8.1 Les données internes

Les données de l'étude sont issues d'un assureur partenaire, qui a préalablement anonymisé les données. L'assureur a fourni deux tables :

— *Table contrat* : recense les informations en portefeuille (date de souscription, formule, franchise, etc...)

— *Table sinistre* : recense les informations des sinistres, rattachés à un contrat présent dans la table contrat (date du sinistre, coût du sinistre, est-ce que le sinistre est clôturé, garantie impactée, etc ...).

1. Le type-mine et le Code National d'identification du Véhicule (CNIT) sont des identifiants du véhicule se trouvant sur la carte grise du véhicule.

Dans un premier temps, nous gardons toutes les garanties, puis nous analyserons dans la suite les garanties les plus pertinentes à modéliser dans le cas de réalisation d'un véhiculier (de coût ou de fréquence).

Nos sinistres sont survenus sur **la période du 01/01/2012 au 31/12/2016**. Ceux-ci doivent être revalorisés à une base monétaire 2016 (mise en as-if des données). Pour ce faire, nous avons convenu de revaloriser les sinistres grâce aux évolutions de prime pure annuelle communiquées par la Fédération Française de l'Assurance.

8.2 Les données externes

Les données externes se regroupent en deux catégories :

— **Les données SRA** : Base qui regroupe toutes les informations composant les véhicules en circulation (type de moteur, carrosserie, etc). Cette base propose également une segmentation des véhicules.

— Les données issues de méthode de ***Webscraping*** : ce sont des données non hiérarchisés, collectées sur un site internet par exemple. Ces données sont ensuite réorganisées pour les traiter et les inclure dans des modèles.

8.2.1 Les données SRA

Variable	Type	Exemple	Nombre de modalités
Date de mise en circulation	Date	01/01/2015	
Groupe SRA	Qualitative ordinale	30	32
Classe SRA	Qualitative ordinale	B	24
Classe réparation SRA	Qualitative ordinale	B	33
Alimentation	Qualitative	Injection	14
Cylindrée	Quantitative	1500	
Energie	Qualitative	Électrique	7
Carrosserie	Qualitative	Break 5 Places	36

TABLE 8.1 – Exemple de variables SRA présentent dans notre base d'étude

Les modalités relatives au véhicule montrent que nous sommes en présence de **données mixtes** : mélange de variables quantitatives et qualitatives. Nous devons alors réduire la dimension du problème en diminuant le nombre de modalités dans les variables qualitatives. Cela améliorera considérablement les performances des algorithmes, en évitant **le fléau de la dimension**. [2]

Traitement des valeurs manquantes (pour la base SRA)

Les spécificités techniques des véhicules ont évolué au fur et à mesure du temps. Ainsi, nous observons des valeurs manquantes pour certaines variables. Par exemple, la puissance fiscale est issue d'une formule qui est en place depuis juillet 1998. [3] Avant cette date, deux autres formules étaient d'usage. Ainsi, plus le véhicule est ancien, plus nous sommes susceptibles de remarquer

2. Par exemple, la classification SRA réalise une catégorie pour les camionnettes et une autre pour les fourgonnettes, mais d'un point de vue assurantiel, ces deux modalités peuvent être regroupés en une seule suite à des justifications statistiques au regard de notre base.

3. Suite à la mise en place de l'article 62 de la loi numéro 98 – 546

des valeurs manquantes dans la base SRA.

Nous pouvons à l'aide de forets aléatoires, ou d'AFDM itératif, compléter notre base de données. Néanmoins, pour éviter l'introduction de biais supplémentaire dans notre modèle, nous avons décidé de ne conserver que les véhicules dont la ligne est complète. D'un point de vue opérationnel, la carte des véhicules permet de rattacher un véhicule non complet à un véhicule complet (utilisation du croisement *Marque* x *Modele* ou de variables qualitatives pour en déduire les variables quantitatives).

Afin d'éviter d'introduire du biais, les véhicules présentant des valeurs manquantes seront retirés de la base de véhicules pour la réalisation de notre classification [4]

8.2.2 *Webscraping* d'un site de vente de véhicule

Le *scraping* permet d'extraire des informations à partir d'un site internet. Les informations présentes sur un site web sont très souvent désorganisées et le programme permettant le *scraping* ordonne ces données pour ensuite les exploiter.
Dans notre contexte, nous n'avons aucun moyen de retracer le prix d'un véhicule en fonction de sa vétusté avec les informations à disposition. Nous allons collecter le prix des véhicules d'occasions. Ces informations nous permettront d'obtenir une première information sur l'impact de la vétusté dans notre classification.

Nous avons extrait les informations d'un site de vente de véhicules d'occasion à l'aide du package **rvest** sous R. Les éléments récupérés sont les suivants :

— Le marque du véhicule

— Le modèle du véhicule

— L'année de mise en circulation

— Le kilométrage

— Le type de carburant

Il était impossible de récupérer la carrosserie avec la maille utilisée. Nous avons donc fixé des hypothèses lorsque nous fusionnerons cette base à notre base d'étude.

Cette donnée est utile pour calculer une décote sur les véhicules. La base SRA nous renseigne sur les caractéristiques techniques du véhicule, mais ne prend pas en compte la vétusté du véhicule. Ainsi, la base SRA ignore totalement si le véhicule à 2 ans ou 4 ans d'âge. De fait, récupérer des données externes sur le prix d'occasion permettra de compléter l'information sur le véhicule si nous nous plaçons dans une maille d'étude adéquate.

Néanmoins, lors de la jointure de cette donnée à notre base véhicule, nous avons remarqué qu'il nous manquait beaucoup de modèles de véhicules. Seulement 1 712 modèles sont représentés dans notre échantillon sur des années différentes et pour plusieurs types d'alimentation. Si nous ne possédons pas une base de prix d'occasions assez profonde, l'apport de cette variable n'est pas fiable. Nous avons décidé de ne pas réaliser de modèle prédictif à partir des informations sur le prix d'occasion des véhicules pour ne pas biaiser nos modèles. Ainsi, **les données relatives au *scraping* ne seront pas utilisées dans notre étude.**

4. Une attention particulière est apportée à la nature de la valeur manquante, par exemple, un véhicule électrique ne possède pas de moteur et certaines variables ne sont alors plus pertinentes.

8.2.3 Ouverture : Données collectable potentiellement utilisable

Plusieurs informations fournies par la SRA peuvent être sujettes à une révision ou à une fiabilisation des données à l'aide du *webscraping*.

La SRA calcule un panier moyen pour le prix des pièces automobiles et les pare-brises. Nous pourrions actualiser cette donnée à l'aide de site de concessionnaire ou créer un panier équivalent à celui de la SRA. En effet, la seule information fournie par la SRA est la revalorisation des prix par marque, mais cette maille est beaucoup trop grande pour notre modélisation. Un site de pièce automobile pourrait être utilisé, d'autant plus que nous possédons le type mine d'un véhicule.

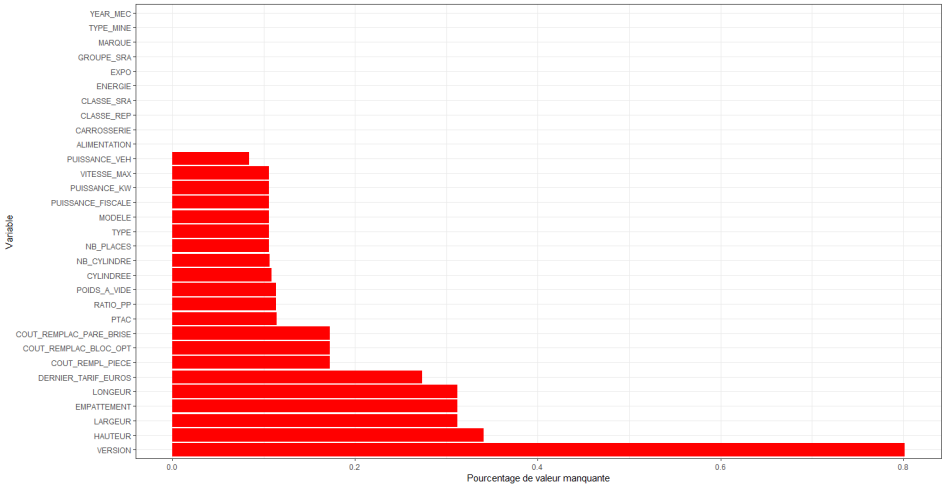

FIGURE 8.1 – Valeurs manquantes dans la base véhicule.

La base SRA comporte beaucoup de valeurs manquantes (Figure 8.1). Les causes de ces valeurs manquantes sont multiples : véhicule ancien, modèle non renseigné... Le problème majeur est que la classification SRA semble très sensible à la présence des données manquantes. Le retraçage de ces informations reste possible à l'aide du numéro d'immatriculation ou bien de la marque, du modèle ou encore de la version. Ce retraçage va permettre de compléter les informations véhicules, en apportant de nouvelles caractéristiques ou en complétant les valeurs manquantes. Par exemple, nous pourrions compléter la taille du véhicule, ou bien récupérer la fiche technique complète d'un véhicule.[5] Ainsi, cette étape de collecte de données permettrait d'affiner les détails techniques du véhicule, les variables liées à la carrosserie ou à l'énergie étant déjà très bien renseignées dans la base SRA.

La récupération des fiches techniques n'a pas été réalisée, car nous ne possédons pas la plaque d'immatriculation compte tenu de l'anonymisation de la base liée à la dernière norme RGPD et nos informations sur le modèle ne sont pas assez précises pour récupérer la fiche technique correspondante.

L'enrichissement de la base SRA serait donc une piste d'amélioration envisageable pour affiner le véhiculier.

5. La fiche technique contient des informations complémentaires, comme le temps nécessaire pour passer de 0 à 100km/h ou des informations sur le châssis, si le véhicule est encore en production etc ...

8.3 Création de la base finale

A partir des fichiers contrats qui ont été alimentés des données externes et de la base sinistre, nous allons construire deux bases d'études : une base spécifique au véhicule (nommé base véhicule) et une base individuelle comportant les détails de l'individu et sa sinistralité sur une période donnée.

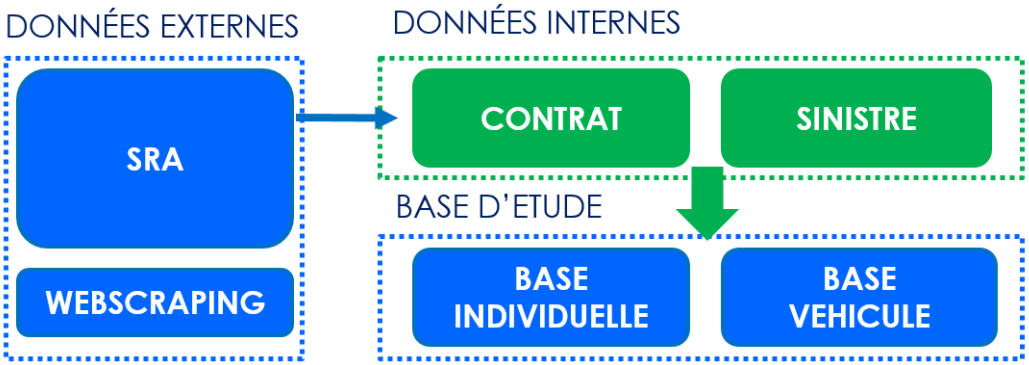

FIGURE 8.2 – Création des bases d'études

La base individuelle correspond à la réunion de la base contrat et la base sinistre. L'actuaire devant être garant de la qualité et de la pertinence des données, nous avons vérifié la qualité des données avant de réaliser notre méthodologie :

— Pertinence de la jointure des données externes.

— Détection des données aberrantes dans la base contrat.

— Assainissement de la base contrat (pour éviter des superpositions de contrats par exemple ou pour supprimer des incohérences dans la base de données).

— Création de nouvelles variables en amont des variables de *clustering* (étape de *feature engineering*).

Le logiciel ADDACTIS® Pricing a été utilisé pour réaliser l'importation de la base contrat et de la base sinistre. Les analyses de la section suivante et les retraitements mineurs seront effectués sur ce logiciel. Dès que ces étapes seront réalisées, nous aurons à notre disposition une base de données exploitable, avec le coût total et la fréquence pour chaque garantie et pour chaque période d'observation des assurés. Cette base va nous permettre de mener à bien la réalisation du véhiculier pour les garanties de notre choix.

8.4 Choix de modélisation pour la réalisation de la base finale

Afin de réaliser la méthodologie développée dans ce mémoire, nous avons choisi les garanties pour lesquelles il est souhaitable de modéliser un véhiculier. Le choix s'est tourné sur deux garanties :

— **La garantie vol** possède peu de volume dans notre base de données (7000 sinistres). L'approche non supervisé permet d'apporter une manière innovante d'isoler l'effet véhicule.

— **La garantie dommage** possède une plus forte volumétrie. Des approches non supervisée et supervisée pourront apporter de l'information aux méthodologies actuelles.

Ainsi, les analyses descriptives seront poussées seulement sur ces garanties pour la modélisation GLM de l'effet conducteur et géographique. Néanmoins, comme ce véhiculier va être testé sur un modèle GLM, il faut adapter les distributions des variables aux problématiques classiques : détermination d'une loi de fréquence et de coût et détection d'un seuil de sinistre grave.

Par la suite, seules les analyses descriptives de la garantie dommage seront présentées.

8.4.1 Détection et application du seuil de sinistres graves

Dans un premier temps, nous réalisons une analyse des coûts des sinistres sur le segment dommage afin de déterminer une loi de modélisation et un seuil de sinistres graves pour la modélisation du véhiculier (Figure 8.3).

Un test de Kolmogorov-Smirnov sur la distribution des sinistres dommage a été réalisé, sans les sinistres négatifs ou nuls. L'hypothèse H_0 est que la distribution des sinistres suit une loi Gamma (dont les paramètres sont estimés par maximum de vraisemblance). Suite à ce test, nous ne rejetons pas H_0 et nous choisissons d'utiliser la loi Gamma pour la modélisation du coût.

Les sinistres négatifs seront retirés de la modélisation en affectant un coût nul. Le quantile à 99% correspond à un montant de 10 000 €. Nous retirons les sinistres dépassant ce seuil pour la modélisation de notre véhiculier afin de rester sur une modélisation attritionnelle sur laquelle nous testerons les véhiculiers créés.

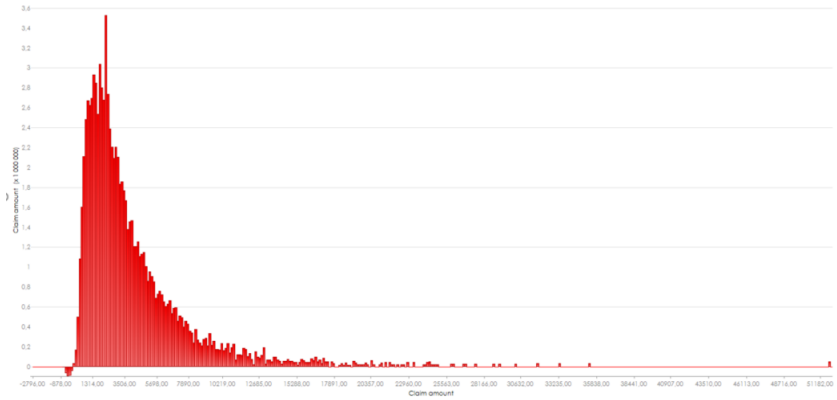

FIGURE 8.3 – Distribution des montants de sinistre dommage

Pour conforter notre choix de seuil, nous allons utiliser **la Théorie des Valeurs Extrêmes**. L'étude du comportement de la fonction des excès moyens (Figure 8.4) permet de choisir un seuil à partir duquel il faut appliquer une autre loi pour modéliser le coût. Nous choisissons un seuil à 12 500 € car au-delà de cette valeur, la FEM possède un comportement linéaire. Nous retrouvons ainsi un seuil assez proche que le niveau de quantile à 1% déterminé précédemment.

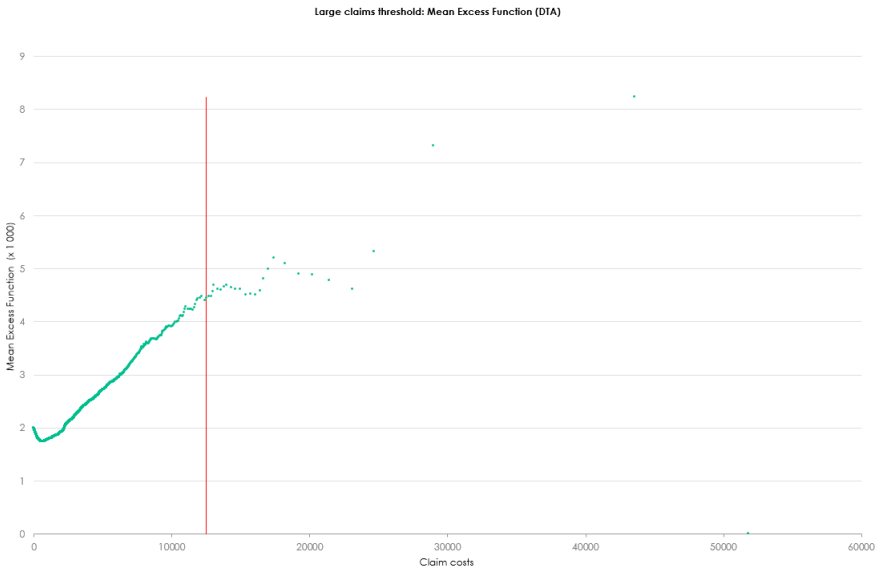

FIGURE 8.4 – Fonction d'excès moyen pour la garantie dommage

La même méthodologie est employée pour la garantie vol, les résultats de cette étude sont résumés dans le tableau suivant :

Garantie	Seuil	Type	Proportion des sinistres
Vol	20 000€	Extraction des sinistres	1%
Dommage	10 000€	Extraction des sinistres	1%

TABLE 8.2 – Hypothèse retenue pour les seuils de sinistre grave

8.4.2 Création de l'espace des véhicules

L'objectif de la partie non supervisée est d'étudier la relation entre caractéristiques techniques et sinistralité d'un véhicule. Cette analyse nécessite la création d'une base de travail relative aux véhicules.

Une première base des véhicules est créée en captant toutes les variations observables pour un véhicule avec le croisement des variables qualitatives. L'exemple suivant permet de mieux comprendre la méthodologie employée pour capter la variation des modalités de chaque véhicule.

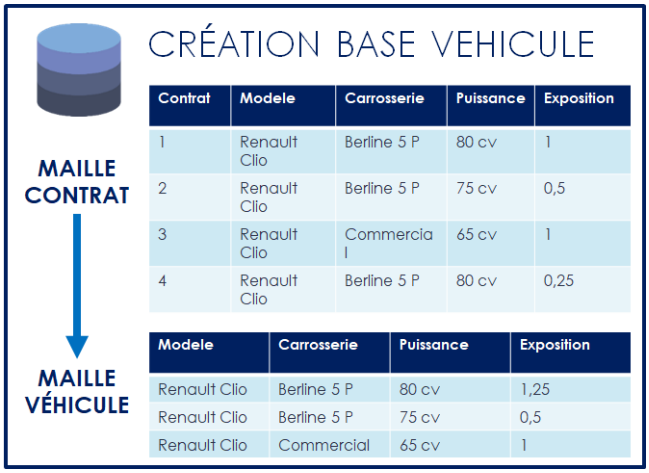

FIGURE 8.5 – Explication de la création de la base véhicule

Pour chaque modèle et chaque version, nous allons capter la variation des variables catégorielles (différence de carrosserie ou d'alimentation) et des variables numériques (différence de puissance, de prix de remplacement des pièces ...). Cela nous permettra d'avoir la maille la plus fine possible, ne possédant pas les codes identifiants SRA dans notre base de données.

C'est sur cette base que nous allons étudier nos véhicules issus du portefeuille de l'assureur. La maille véhicule pourra dépendre des méthodes utilisées. En effet, dans la partie non supervisée, nous devons regrouper les individus possédant des véhicules identiques afin de remplir les conditions liées à la complexité de nos algorithmes. Il faudra trouver un affinage optimal pour capter le plus d'information possible.

Il faut aussi tenir compte du fait que certains véhicules sont rentrés manuellement par les agents généraux ou les courtiers. Ainsi, il nous a été fourni une variable permettant de savoir si le véhicule a été saisi manuellement ou si ces informations proviennent de la base SRA. La base véhicule ne contiendra que des véhicules SRA. Pour les véhicules non classés, nous pourrons les assigner à un groupe grâce au représentant du groupe que nous déterminerons à l'aide d'un algorithme de *clustering*. Ainsi, chaque véhicule saisi manuellement pourra être affecté à un identifiant à la maille voulue.

8.4.3 Choix de la maille des véhicules

Deux mailles de véhicules vont être utilisées pour cette étude. Une première maille exploratoire correspondant à toutes les variations de toutes les variables véhicules et une autre maille captant les variations présentes entre certaines variables. Le tableau présentant ces deux mailles est présenté en annexe (Figure 11.5).

L'objectif de la nouvelle maille est de réduire le nombre de véhicules au sein de la base et d'avoir des représentants plus fiables et représentatifs au niveau de l'exposition. Des informations, comme l'âge du véhicule, ne seront pas intégrées dans l'analyse non supervisée, car elles seront détectées comme peu contributrices lors de l'exploration des données par AFDM. De plus, ces informations créent beaucoup de lignes différentes pour des véhicules similaires et réduisent la fiabilité de la maille.

8.5 Les analyses descriptives de la base d'étude

8.5.1 Distribution de la fréquence

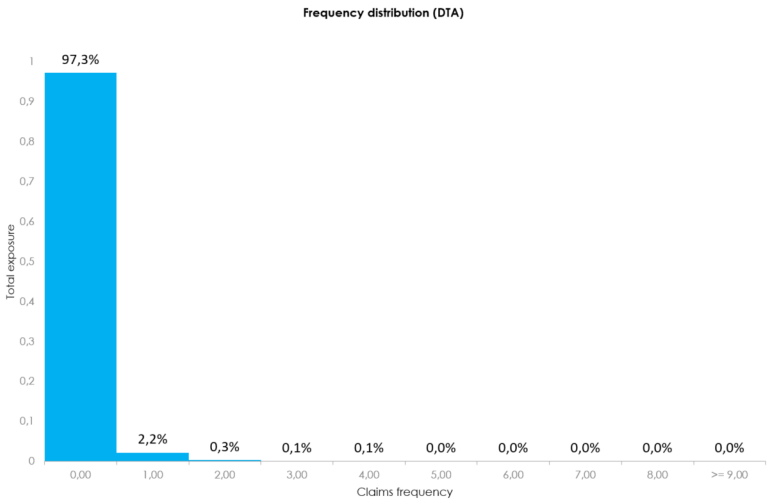

FIGURE 8.6 – Distribution de la fréquence de la garantie dommage

Dans notre portefeuille, 97.3% de nos assurés couverts par cette garantie n'ont pas déclaré de sinistres. 2.2% d'entre eux en ont déclaré un et très peu en ont déclaré plus de 2. Il reste ensuite à savoir si la loi modélisée est équidispersée ou surdispersée. Si l'espérance du nombre de sinistres est inférieure à la variance, alors il est préférable de modéliser ce nombre à l'aide de la loi binomiale négative.

Garantie	Espérance	Variance
Vol	0.136%	0.139%
Dommage	1.21%	1.23%

TABLE 8.3 – Espérance et variance de la distribution du nombre de sinistres pour la garantie vol et dommage.

Cette analyse nous dirige vers le choix d'une modélisation de la loi de fréquence par la loi de Poisson, qui est très utilisée en assurance pour la modélisation de loi discrète.

8.5.2 Analyse univariée pour la garantie dommage

L'étape suivante de notre analyse est de détecter les segments tarifaires où il est possible de capter une évolution de coût moyen ou de fréquence, pour ensuite intégrer ces segments dans une modélisation GLM.

L'extraction de l'effet véhicule est opérée seulement sur la garantie dommage. Nous présenterons (Figure 8.7 et Figure 8.8) quelques graphiques discriminant la fréquence associée à cette dernière. Des regroupements seront effectués lors de la modélisation GLM de la variable cible.

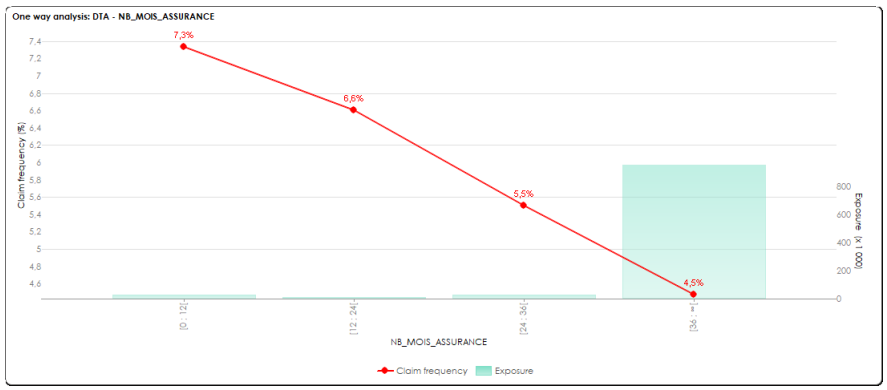

FIGURE 8.7 – Fréquence observée en fonction des antécédents d'assurance

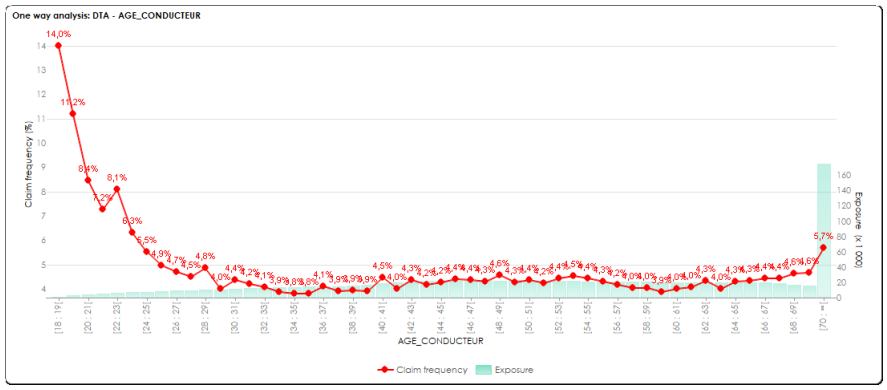

FIGURE 8.8 – Fréquence observée en fonction de l'âge du conducteur

8.5.3 Etude des corrélations entre les variables explicatives

Les modèles GLM sont très sensibles aux corrélations entre les variables. En effet, si deux colonnes sont très ressemblantes, la matrice de design est très difficile à inverser et l'algorithme pourrait ne pas converger.

Ainsi, nous avons effectué une analyse de corrélation à l'aide de la statistique de Cramer V. Les variables numériques ont été découpées en segment lors du mélange numérique / catégoriel.

Variable 1	Variable 2	Cramer V
Ancienneté Permis	Age conducteur	65%
Type de personne	CSP	51%

TABLE 8.4 – Couple de variables fortement corrélées (hors véhicule)

8.5.4 Statistiques descriptives pour les véhicules

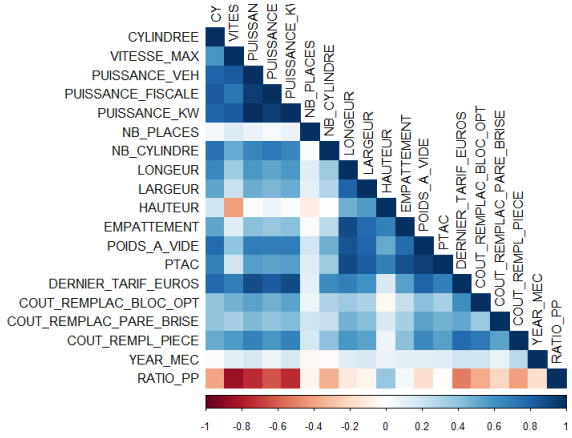

FIGURE 8.9 – Corrélation entre variables quantitatives de la base de véhicule (Rho de Pearson)

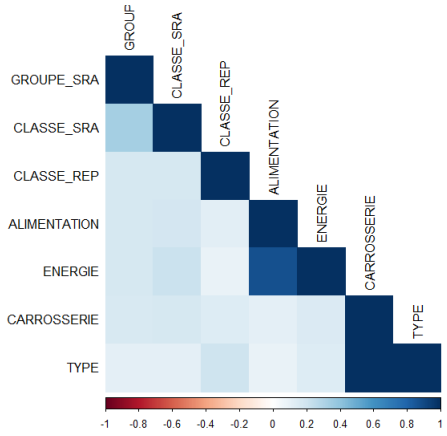

FIGURE 8.10 – Corrélation entre variables qualitatives de la base de véhicule (Cramer V)

Nous détectons aussi de fortes corrélations entre les variables véhicules (Figure 8.9). Par exemple, les variables relatives à la puissance sont dans des unités différentes :

— La puissance fiscale est en cheval fiscal.

— La puissance est en cheval.

— La puissance kw est en kilowatt.

Elles apportent donc la même information, nous ne garderons que la puissance en kw. Néanmoins, les variables relatives à la taille du véhicule sont corrélées, mais n'apportent pas la même information : elles seront conservées.

Les variables qualitatives possèdent des informations redondantes. Ainsi la carrosserie et le type sont fortement corrélés, tout comme l'alimentation et l'énergie (Figure 8.10).

En résumé

Les choix de modélisations ont été effectués pour créer des véhiculiers apportant une nouvelle informations. **La fréquence vol** a été choisie pour sa faible volumétrie sur la sinistralité et sera l'objet du chapitre suivant. **La fréquence dommage** a été choisie pour illustrer la création d'un véhiculier supervisé à partir de la modélisation d'un GLM sans variable véhicule.
Les analyses descriptives ont permis une première connaissance du jeu de données fourni par l'assureur partenaire et d'orienter nos choix sur la modélisation du risque véhicule pour le GLM, et de comprendre en profondeur la construction de la base SRA pour réaliser correctement le *clustering* de la base véhicule, et détecter d'éventuelles interactions entre les variables.

Chapitre 9

Construction des groupes de véhicules par approche non-supervisée.

L'objectif de ce chapitre est de détailler l'approche non supervisée [1] menée dans ce mémoire pour classer les véhicules. Cette approche commence par l'exploration de la base des véhicules dont la construction est détaillée dans une première partie. Ensuite, après la réalisation d'une étude descriptive des véhicules en portefeuille, une première projection de la base des véhicules par AFDM est effectuée. Elle permet d'obtenir une carte de véhicules représentant les véhicules dans un espace réduit. Ce nouvel espace résume la variabilité de la base initiale et isole différentes typologies de véhicules. Cette projection est utilisée afin de créer un premier véhiculier technique [2] à l'aide d'algorithme de *clustering* par partitionnement.

Une seconde approche pour la réalisation de ce véhiculier technique repose sur la génération d'une matrice de dissimilarité de nos véhicules. Compte tenu des performances de calcul, une maille adaptée à notre portefeuille est choisie.

L'exploration de cette base de données s'effectue à l'aide de deux méthodes :

- Réduction de l'espace initiale de la base, par un algorithme de réduction de dimension pour données mixtes : l'AFDM.
- Calcul d'une matrice de dissimilarité pour les individus à l'aide de la distance de Gower puis intégration de la matrice de dissimilarité dans un algorithme de réduction de dimension (SOM) ou de *clustering* (PAM).

9.1 Réduction de l'espace initial de la base par un algorithme de réduction de dimension pour données mixtes : l'AFDM

9.1.1 Projection de l'espace des véhicules par AFDM

La maille exploratoire est utilisée pour réaliser l'espace des véhicules afin d'effectuer les premières analyses sur nos véhicules. Celle-ci va nous permettre de réaliser une première exploration approfondie des véhicules, en plus de l'analyse de corrélation que nous avons réalisée précédemment. L'étude de l'AFDM est essentielle pour déterminer une maille d'étude optimale et a permis d'orienter notre choix vers la sélection de la maille d'étude numéro 4 (Figure 11.5).

1. L'objectif étant de ne pas intégrer la sinistralité dans un premier temps
2. Nous appellerons véhiculier technique un véhiculier qui capte les caractéristiques techniques

La projection AFDM ne doit pas contenir trop de variables qualitatives, au risque d'augmenter trop fortement l'espace d'étude et donc de ne plus pouvoir expliquer les variations au sein des différentes poches créées par les variables qualitatives. Ainsi, nous décidons de ne pas inclure la classification SRA dans notre projection. De plus, la variable alimentation et la variable relative au type d'énergie alimentant le véhicule sont fortement corrélées : nous ne garderons que la variable énergie.

Nous avons en notre possession 189 518 véhicules et 24 variables que nous allons explorer à l'aide de l'AFDM.

Dans un premier temps, nous réaliserons l'AFDM sur 10 dimensions et nous observerons la part de variance expliquée par chaque axe (Figure 9.1) :

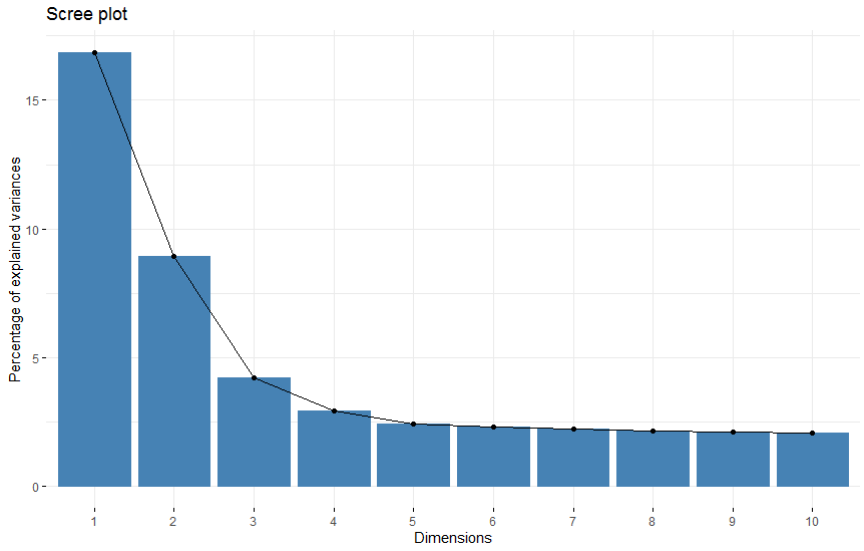

FIGURE 9.1 – Scree plot de l'AFDM sur la base véhicule (maille exploratoire)

A l'aide des 3 premières dimensions, nous arrivons à expliquer 30% de la variance de la base totale. En réalisant un affinage des variables qualitatives, un niveau de 40% est atteint et permet la réalisation du *clustering* directement sur ce nouvel espace, car nous avons réussi à résumer une quantité suffisante d'informations.

En utilisant le critère du coude, nous ne jugeons pas nécessaire d'utiliser les dimensions supérieures à 3, car celles-ci risqueraient d'expliquer du bruit ou des poches de véhicules très précis (véhicule électrique par exemple) qui ne représenteraient pas la globalité du portefeuille. Ces effets seront mieux captés par des algorithmes locaux, comme t-SNE ou les cartes de Kohonen. L'avantage de l'AFDM est l'interprétation des différentes composantes qui contribuent à chaque axe et possédant alors un réel intérêt dans la modélisation du risque véhicule. Nous les interpréterons dans la partie suivante pour nous approprier les axes de projections.

9.1.2 Interprétation des axes

Dans un premier temps, analysons la contribution des variables pour les axes d'intérêts, c'est-à-dire les 3 premiers. Par exemple (Figure 9.2) , le poids à vide et le dernier tarif en euros contribuent fortement au premier axe. La ligne rouge en pointillé désigne la contribution dans le cas où les variables auraient des contributions uniformes. Ainsi, pouvons conclure qu'une variable au-dessus de ce seuil contribue fortement à l'axe en question.

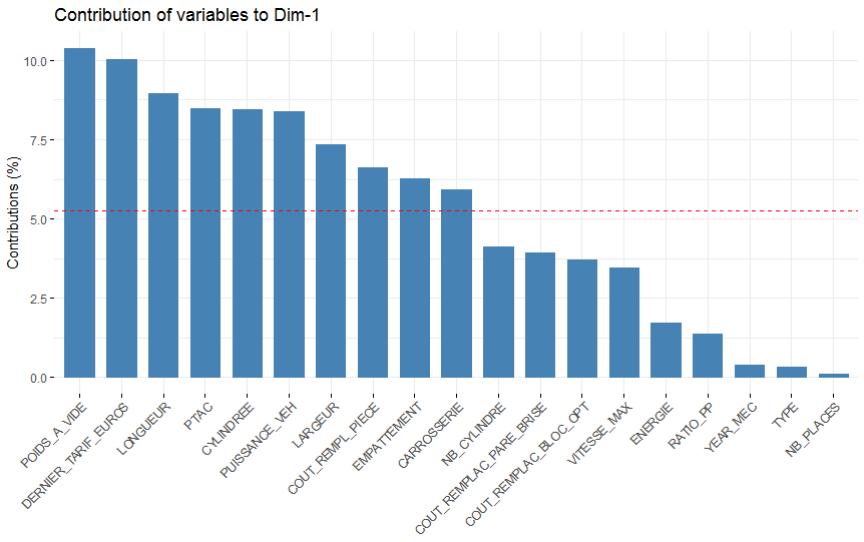

FIGURE 9.2 – Contribution des variables véhicules pour le premier axe d'inertie

Le tableau suivant récapitule les variables les plus contributrices aux trois premiers axes. Nous remarquons l'importance de la variable carrosserie, qui est souvent utilisée pour capter le risque véhicule lors de l'extraction du risque géographique. Cette variable est donc importante pour capter le risque véhicule.

Dimension 1	%	Dimension 2	%	Dimension 3	%
Poids à vide	10,4 %	Carrosserie	19,3 %	Carrosserie	37,3%
Dernier tarif	10 %	Type	16,4 %	Nombre de places	28,4%
Longueur	9%	Ratio Poids Puissance	13 %	Type	8,5%

TABLE 9.1 – Top 3 des variables les plus contributrices pour les trois premiers axes. Le pourcentage désigne la contribution sur la dimension concernée.

Pour les variables numériques, nous utiliserons le cercle unitaire dont les flèches montrent la contribution de la variable dans les deux dimensions considérées (Figure 9.3). Ainsi, nous remarquons que l'année de mise en circulation contribue faiblement aux deux premiers axes, car la SRA a une vision à neuve du véhicule. Celle-ci n'est pas captée par la base de données. Enfin, la puissance du véhicule, le prix, le ratio poids/puissance et les dimensions du véhicule ressortent dans notre AFDM.

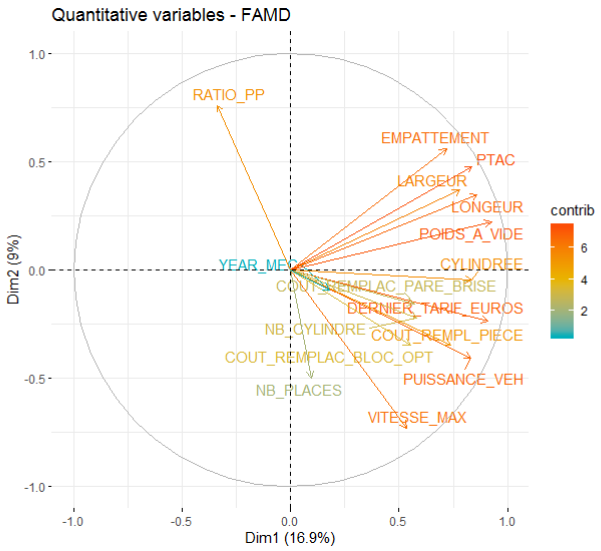

FIGURE 9.3 – Cercle unitaire pour les deux premiers axes, pour l'interprétation des variables numériques

Nous représentons les modalités qualitatives (Figure 9.4) sur les premières dimensions. Cette analyse nous permet de réaliser nos regroupements. Par exemple, nous regrouperons les modalités "Châssis cabine" et "Châssis double cabine" pour améliorer les performances de l'algorithme, car les coordonnées de ces deux modalités sont très proches dans notre espace. De plus, nous observons que les véhicules utilitaires contribuent fortement à notre analyse. Le *clustering* non supervisé permettra d'isoler ces véhicules, tout en gardant un modèle unique de véhiculier.

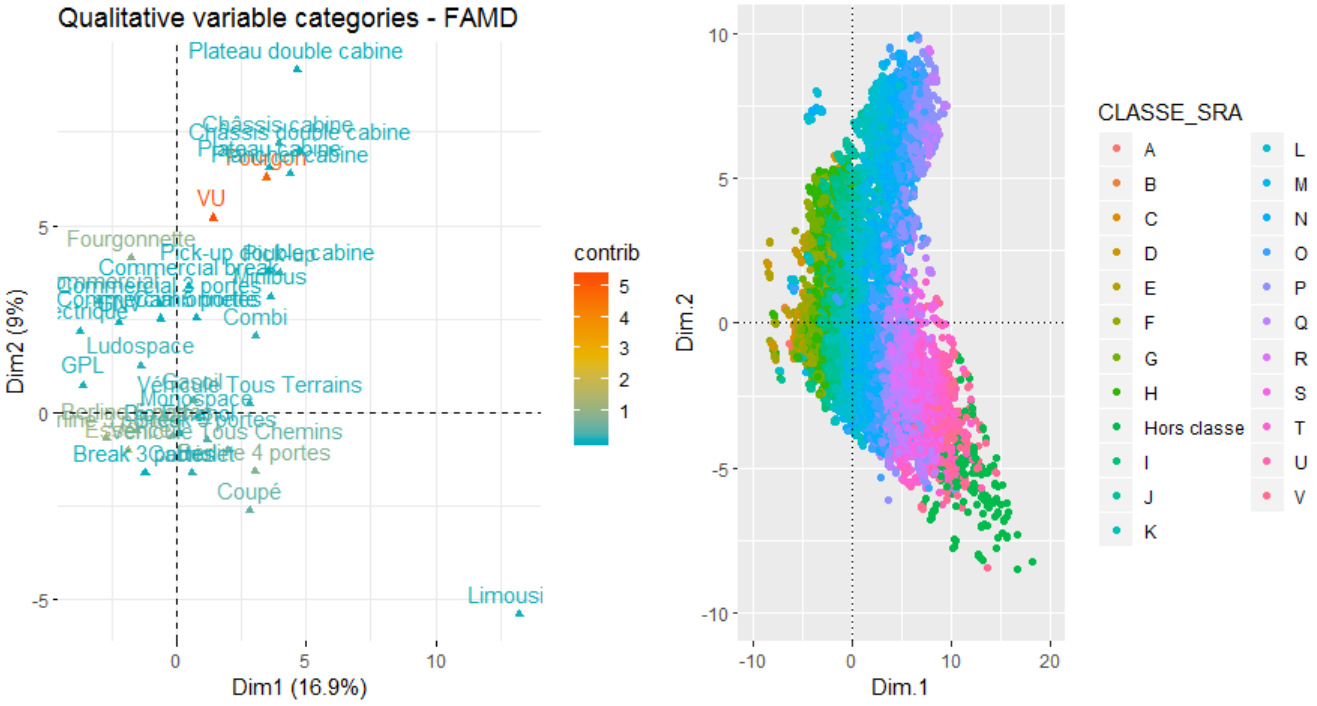

FIGURE 9.4 – Position des modalités qualitatives sur les deux premiers axes et comparaison avec la classe SRA

Nous avons isolé les typologies de véhicules suivantes à l'aide de cette analyse :

— Les véhicules de luxe, puissant, dans la partie inférieure droite de la carte.

— Les véhicules utilitaires, de dimension assez élevée, dans la partie supérieure droite.

— Les véhicules commerciales et les petits véhicules, dans la partie supérieure gauche de la carte.

9.1.3 Comparaison de la projection avec la classification SRA

Comme spécifié plus tôt, il a été décidé de ne pas intégrer la classification SRA dans la projection, pour ne pas rajouter des segments supplémentaires. La classification SRA résumant de l'information, elle ne semble pas utile à priori pour expliquer la variance de la base véhicule, car cette classification est construite à partir des variables que nous souhaitons explorer. Néanmoins, il est possible de projeter les modalités de la classification SRA sur le nouvel espace, pour voir si celle-ci ressort (Figure 9.4). La superposition de la classe SRA permet de voir l'importance portée par le premier axe, résumant l'information.

De plus, il est possible de superposer la sinistralité observée sur notre portefeuille, en recalculant la fréquence et le coût moyen associés à chaque groupe. Cette analyse permet de détecter des segments de tarification non pris en compte avec la classification actuelle. La Figure 9.5 et la Figure 9.6 montrent les différents niveaux de sinistralité sur un gradient de couleur (bleu pour faible et rouge pour élevé). La taille des étiquettes est proportionnelle à l'exposition associée à chaque groupe. Dans un cas extrême, une petite étiquette représente un groupe sous-représenté. Les coordonnées des points sont les médianes associées à chaque modalité dans le nouvel espace des véhicules.

Nous observons sur la Figure 9.5 l'impact de la variable classe SRA sur le coût moyen de la garantie vol. Un véhicule "Hors classe" étant un véhicule de luxe, son coût moyen est en moyenne plus cher qu'un véhicule classique. Sur la Figure 9.6, des poches plus atypiques sont présents : les véhicules possédant une valeur de véhiculier faible ou élevée sont sous-représentés en termes d'exposition. Les petits véhicules, classés dans le groupe 25/26 ressortent en fréquence dans notre analyse pour la fréquence vol, car ce sont des véhicules faciles à dérober. Néanmoins, les classes et groupes de véhicules extrêmes sont sous-représentés en termes d'exposition, d'autant plus que ces analyses ne tiennent pas encore compte des véhicules possédant des valeurs manquantes.

Ces projections permettent aussi de montrer que la classification SRA et le groupe SRA ne sont pas adaptés à des véhicules utilitaires, que nous avons identifiés sur la partie supérieure de la projection. La création de croisement Type/Classe SRA ou Type/Groupe SRA n'a pas permis de recréer ces segments, d'où l'intérêt de l'exploration réalisée.

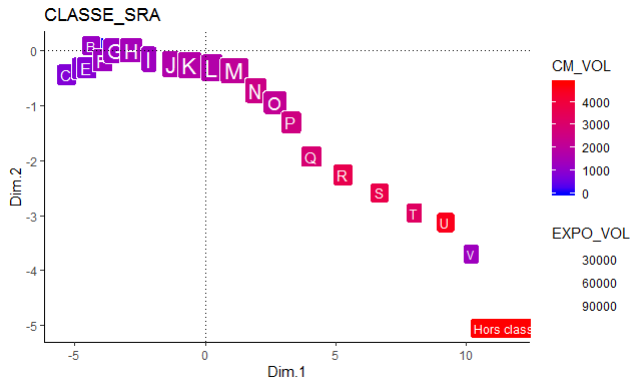

FIGURE 9.5 – Projection de la classe SRA sur les deux premiers axes

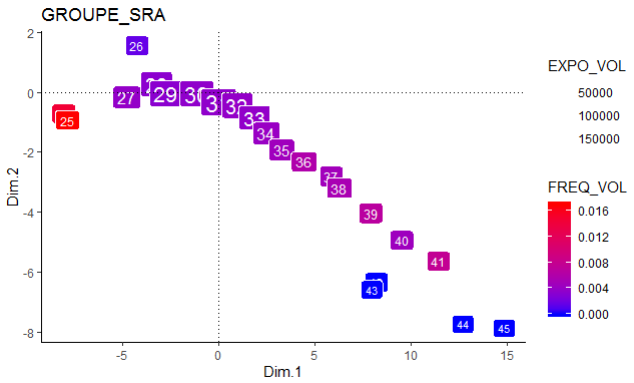

FIGURE 9.6 – Projection du groupe SRA sur les deux premiers axes

9.1.4 Création d'une première carte des véhicules et création des clusters

L'objectif est désormais de regrouper les variables catégorielles. La Figure 9.4 permet de regrouper des modalités relatives à la variable catégorielle carrosserie et énergie. Ainsi, à ce stade, nous possédons 34 429 véhicules et 18 variables explicatives que nous devons projeter à l'aide de l'AFDM. Les modalités qualitatives de la variable carrosserie ont été réduites de 36 à 18.

Nous obtenons une amélioration significative de notre graphique de silhouette de 10% sur les trois dimensions. Nous segmentons notre espace à l'aide d'un algorithme de partitionnement comme PAM pour réaliser un *clustering* non supervisé.

La maille véhicule étant la maille exploratoire, nous devons adapter l'algorithme PAM pour travailler sur des échantillons *bootstrap* car cette maille possède beaucoup de véhicules. L'algorithme CLARA (*CLustering for Large Applications*) est une version bootstrap de l'algorithme PAM (*Partitioning Around Medoids*)permettant de supporter l'application de PAM sur des jeux de données avec beaucoup d'individus. L'interprétation est similaire à PAM. Nous sélectionnerons le nombre de *clusters* à l'aide du critère de silhouette que nous calculerons pour un nombre de *clusters* allant de 2 à 50 clusters.

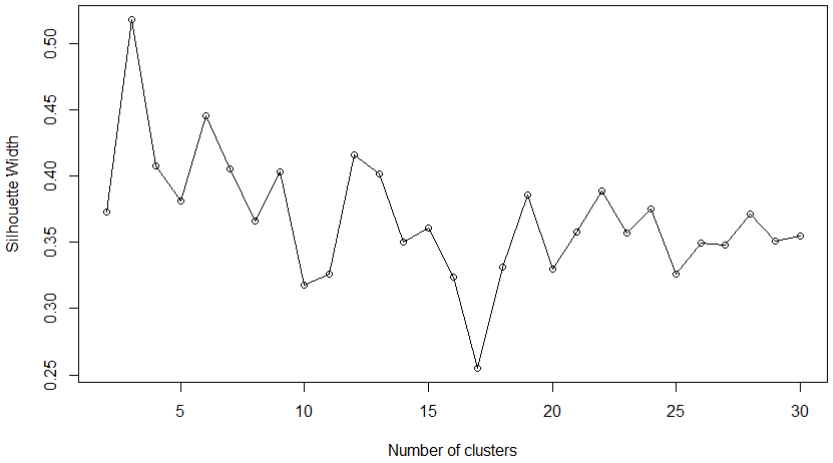

FIGURE 9.7 – Critère de silhouette pour un nombre k de clusters. (CLARA)

Notre objectif est de créer des groupes de véhicules homogènes en termes de caractéristiques techniques. Nous ne sommes pas dans une problématique de détection de profils atypiques. Ainsi, nous souhaitons obtenir un nombre de groupes supérieur à 5, juste avant le pic observé à k=6.

Au regard du critère de silhouette, nous retenons $k \in \{6, 12, 19\}$. Une fois le nombre de *clusters* k fixé, l'algorithme assignera les groupes à chaque véhicule. La vérification des résultats du *clustering* s'effectue grâce à un graphique [3] affichant l'indice de silhouette pour les individus (Figure 9.8). La ligne en pointillé rouge représente le niveau de silhouette moyen. Idéalement, chaque *cluster* doit avoir au moins plusieurs individus au-delà du niveau moyen. Ensuite, il faut éviter d'avoir des indices négatifs, signifiant que l'individu peut être classé dans le mauvais groupe.

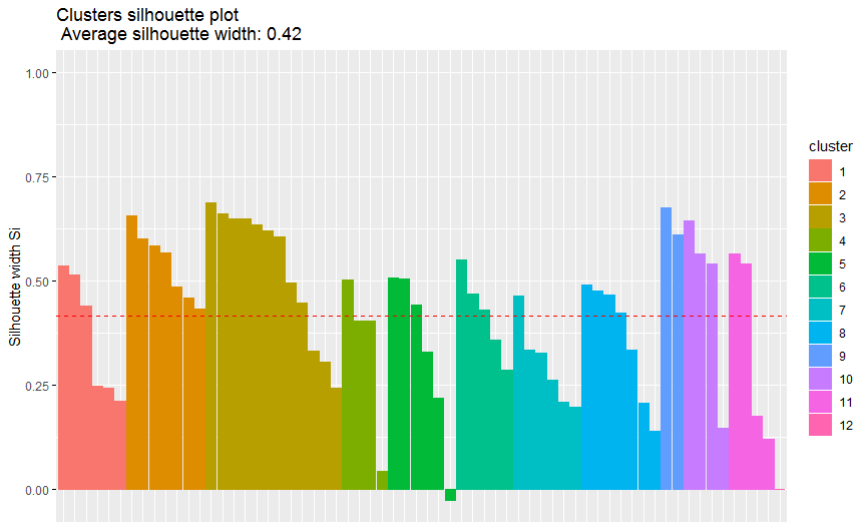

FIGURE 9.8 – Silhouette plot pour 12 clusters (CLARA)

Une projection des points sur les deux premières dimensions de l'AFDM permet de vi-

3. Ce type de représentation est appelé *silhouette plot*

sualiser le résultat de CLARA. L'interprétation des axes dans la partie précédente permet de donner du sens au *clustering* réalisé par l'algorithme. Dans le cas où k = 12, nous remarquons que le groupe 6 correspond aux véhicules puissants et chers et que le groupe 12 correspond aux véhicules utilitaires comme des camionnettes, des châssis nus, etc ... (Figure 9.9). Nous superposons ces clusters avec la sinistralité observée. Les groupes observés font ressortir des niveaux de sinistralité différents et permettent de valider la présence d'une corrélation entre les caractéristiques techniques et la sinistralité (Figure 9.10).

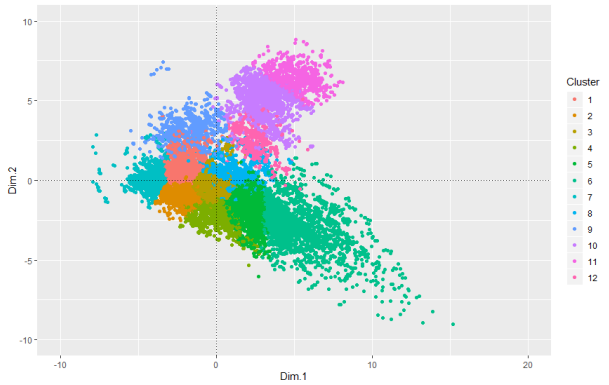

FIGURE 9.9 – Projection des clusters (k=12) sur les deux premiers axes (CLARA)

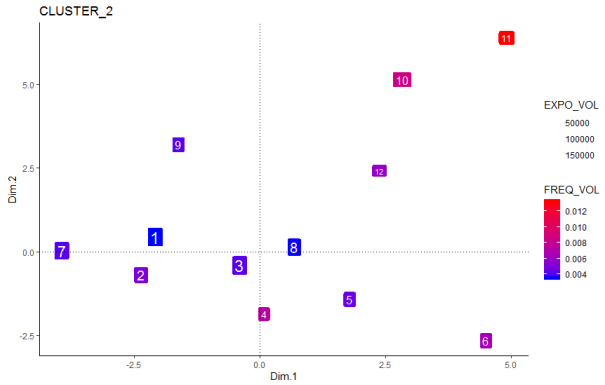

FIGURE 9.10 – Sinistralité des clusters relative à la fréquence vol (CLARA k=12)

Maintenant que nous avons ces coordonnées, nous devons projeter les véhicules qui présentent des valeurs manquantes. Nous pouvons compléter les valeurs manquantes, essentiellement numériques dans notre base, grâce à des forêts aléatoires. Cela permet d'éviter d'imputer la moyenne, sachant que l'AFDM est très sensible aux valeurs manquantes. De plus, nous souhaitons prédire des coordonnées dans le nouvel espace des véhicules et certains véhicules SRA sont plus ou moins complets. Une fois notre base complète, nous ajustons les modalités catégorielles et nous effectuons des regroupements (certaines modalités ne sont pas présentes dans notre base d'apprentissage de la AFDM) puis nous réalisons une prédiction des coordonnées dans le nouvel espace des véhicules. Nous assignerons ensuite à chaque observation un *cluster*.

Les algorithmes PAM et CLARA ne sont pas créés pour réaliser des prédictions. De fait, nous calculerons la distance d'un individu par rapport à chaque représentant des *clusters* et nous affecterons le *cluster* dont le représentant est le plus proche de l'individu. Nous obtenons alors une carte complète, avec tous les *clusters*.

En résumé

A partir d'une base contenant les véhicules présents et leurs caractéristiques associées, 3 *clusterings* ont été crées et les coordonnées des 3 premiers axes de la carte des véhicules seront conservées pour un lissage spatial.

Chaque véhicule de la maille appartient à un *cluster*. Nous sommes capables d'assigner à chaque individu le groupe auquel le véhicule appartient et nous verrons si le groupe ressort au sein d'une analyse de Type-III dans un GLM.

9.2 *Clustering* sur une matrice de dissimilarité par l'algorithme PAM

La matrice de dissimilarité est calculée à partir des 34 429 véhicules et de 18 variables explicatives caractérisant les véhicules.

9.2.1 Méthodologie mise en place

L'algorithme CLARA, contrairement à l'algorithme PAM, ne permet pas d'utiliser une matrice de dissimilarité en entrée. Ainsi, il faut reprendre le principe de l'algorithme CLARA et l'adapter.

Pour ce faire, nous prenons un échantillon de la matrice de dissimilarité composé des véhicules supérieurs à un certain seuil d'exposition à déterminer. Puis la réalisation d'un *silhouette plot* nous permet de de sélectionner le nombre de *clusters* nécessaire à notre modélisation.

Ensuite, nous générons les groupes sur la totalité de la matrice de dissimilarité et garder en mémoire le résultat de l'algorithme, dont la validation est effectuée grâce à l'algorithme t-SNE, en projetant notre matrice de dissimilarité et en vérifiant le caractère local du *clustering*.

Enfin, nous assignons à chacun des véhicules présentant des données manquantes un score de dissimilarité permettant de lui assigner le médoïde le plus proche.

9.2.2 Choix du seuil d'exposition

La garantie vol étant incluse dans la garantie dommage, le seuil d'exposition est choisi en fonction de l'exposition vol de chaque véhicule. Pour ce faire, nous regardons la taille de la base véhicule en filtrant les véhicules qui ont une exposition vol au-dessus d'un certain niveau (Figure 9.11). Nous choisissons alors un seuil d'exposition égale à 20 ans.

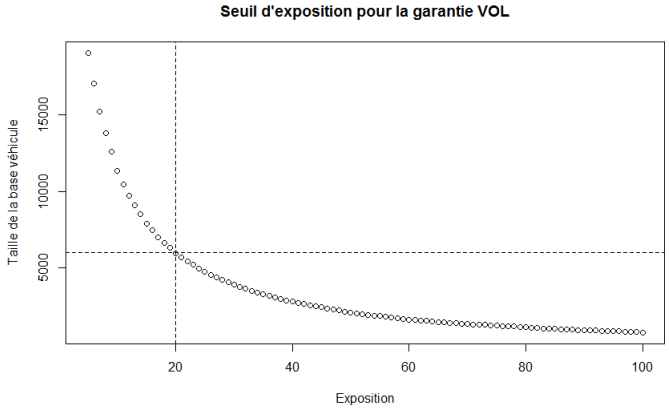

FIGURE 9.11 – Choix du seuil d'exposition pour la garantie vol

9.2.3 Choix du nombre de clusters k et analyse des silhouettes

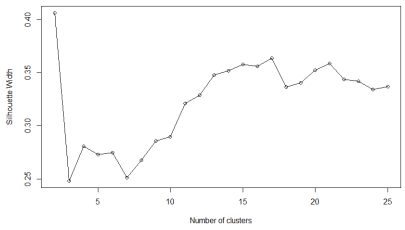

FIGURE 9.12 – Indice de silhouette moyen en fonction du nombre de clusters (PAM equipondéré)

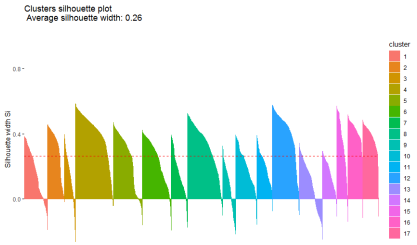

FIGURE 9.13 – Silhouette plot pour k=17 clusters (PAM equipondéré)

Nous décidons de prendre un nombre de clusters égal à k = 17 (Figure 9.12), valeur où l'indice de silhouette est le plus élevée au-dessus de k = 2. Nous exécutons ensuite une seconde fois l'algorithme PAM mais sur la matrice de dissimilarité complète, puis nous affichons le silhouette plot en fonction du *clustering* pour pouvoir le valider. Le nombre de classes étant assez élevé, il semble logique d'avoir quelques individus qui sont proches des voisins et des valeurs d'indice de silhouette négatif.

9.2.4 Visualisation du *clustering* par t-SNE

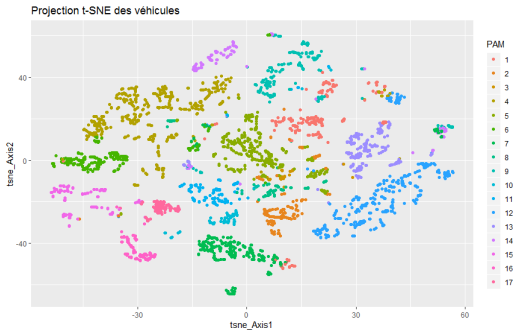

FIGURE 9.14 – Visualisation du *clustering* par t-SNE (PAM equipondéré)

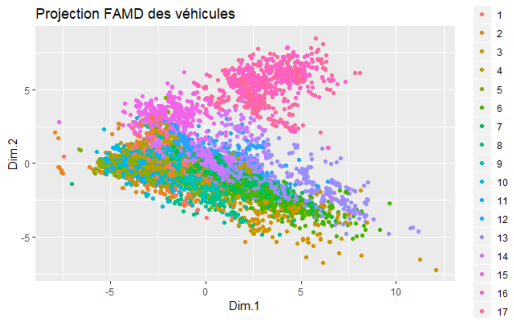

FIGURE 9.15 – Visualisation du *clustering* sur la carte des véhicules (PAM equipondéré)

L'algorithme t-SNE permet de réduire un espace de grande dimension à un espace de faible dimension (2 ou 3 dimensions). Nous observons des petits groupes parmi les clusters, justifiant le caractère local de ceux-ci. Sur la Figure 9.14 nous observons des groupes assez proches, même si certains individus semblent mal classés du fait des valeurs négatives prises dans la Figure 9.13. Néanmoins, le résultat est plutôt encourageant et nous intégrerons cette variable dans notre modélisation, quitte à effectuer des regroupements pour certaines classes.

Pour améliorer les résultats, les pistes suivantes sont envisagées :

— Réduire l'espace de travail en réalisant une première sélection de variables, quitte à guider cette sélection par rapport à la fréquence vol par exemple.

— Pondérer par certaines variables pour faire des groupes plus homogènes.

— Pondérer par la sinistralité observée.

9.2.5 Visualisation du *clustering* sur la carte des véhicules

Pour conforter notre approche précédente, nous projetons la classification sur la carte des véhicules (Figure 9.15). Cette projection permet de retrouver des structures détectées précédemment. Par exemple, les groupes 16 et 17 occupent la partie supérieure droite relative aux véhicules utilitaires.

De plus, la structure des groupes n'est pas aléatoire ce qui conforte sur la démarche effectuée. Néanmoins, elle perd en interprétabilité pour certaines classes.

En résumé

A ce stade, nous avons calculé à partir des variables explicatives de chaque véhicule un score de dissimilarité entre les véhicules. Ce score permet la construction d'une matrice de dissimilarité qui est l'entrée de l'algorithme PAM.
Les statistiques calculées permettent de retenir 17 groupes de véhicules, tout en retrouvant des structures identifiées par l'algorithme AFDM.

9.3 Ajustement de la matrice de dissimilarité par une pondération arbitraire

La distance de Gower accepte une pondération en fonction des variables présentent dans le calcul de la distance. Il est possible de donner plus d'importances à des variables quantitatives ou qualitatives.

Nous retombons alors sur une problématique supervisée, sans pour autant utiliser un modèle prédictif comme des arbres par exemple, nécessitant une variable réponse Y.

Nous nous sommes alors posé la question suivante : **Que se passe-t-il en pondérant par la sinistralité observée ?**

Une nouvelle matrice de dissimilarité est calculée en pondérant la fréquence vol sur notre base de données. Nous visualiserons le résultat du *clustering* à l'aide de t-SNE.

9.3.1 Introduction d'une pondération arbitraire

Pour pondérer notre matrice de dissimilarité, il est souhaitable de sélectionner des variables améliorant la segmentation. Pour ce faire, étant donné que l'on souhaite intégrer la sinistralité, nous allons déterminer le gain d'information [9] qu'une variable a sur la variable cible (la fréquence vol dans notre situation).

La librairie **FSelector** implémente la mesure de l'*information gain*, mesure utilisée dans l'algorithme des forêts aléatoires.

Nous obtenons la sélection de variables présentent sur la Figure 9.16. Les variables relatives à la taille et aux coûts de réparation pour différentes pièces du véhicule, sont les plus discriminantes pour expliquer la fréquence vol. Néanmoins, les variables relatives à la puissance ou au dernier prix en euros ne ressortent pas. Enfin, la variable carrosserie ne ressort pas, due à ses nombreuses modalités, mais nous décidons de la conserver dans notre pondération, car celle-ci a prouvé son importance dans l'approche par AFDM.

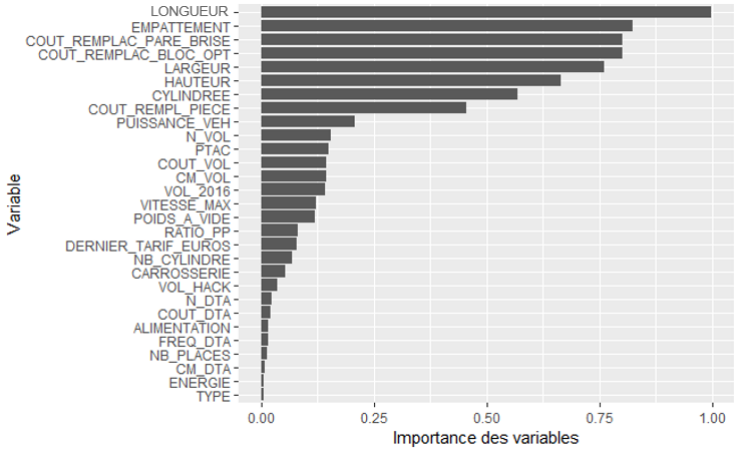

FIGURE 9.16 – Gain d'information pour la variable relative à la fréquence vol.

En supplément de cette sélection de variables, nous allons pondérer par rapport à la fréquence vol pour donner plus d'importance à ce critère. Nous avons décidé d'accorder 50 % du poids à cette variable, pour équilibrer l'aspect technique des véhicules et l'aspect historique.

9.3.2 Sélection du nombre de clusters et analyse des résultats

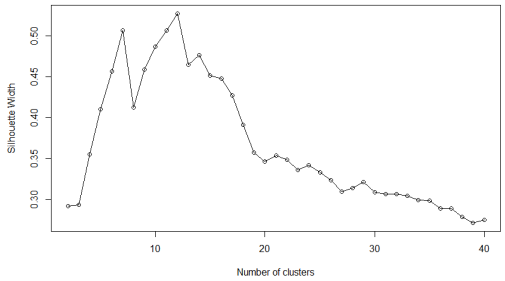

FIGURE 9.17 – Indice de silhouette moyen en fonction du nombre de clusters (PAM pondéré)

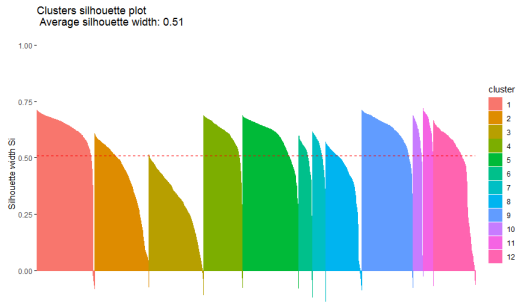

FIGURE 9.18 – Silhouette plot pour k=12 clusters (PAM pondéré)

La Figure 9.13 permet de choisir 12 clusters pour réaliser le *clustering* sur la matrice de dissimilarité. La Figure 9.18 permet de montrer l'amélioration du *clustering*, d'une part par l'augmentation globale de l'indice de silhouette moyen globale, mais aussi en observant que tous les clusters sont majoritairement au-dessus de ce niveau.

9.3.3 Visualisation des résultats par t-SNE

Enfin, la projection par t-SNE sur la Figure 9.19 permet de souligner une structure locale beaucoup plus dessinée que précédemment, d'où une amélioration signficative de notre *clustering*.

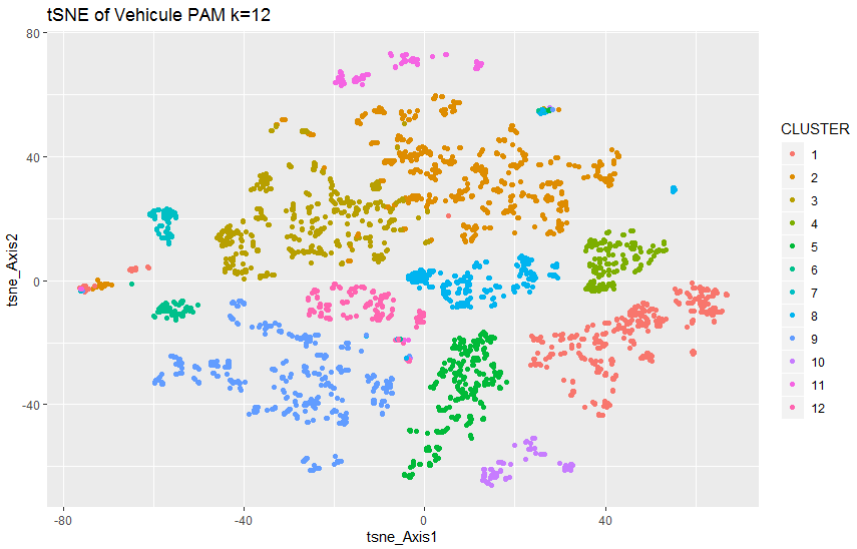

FIGURE 9.19 – Projection t-SNE de la matrice de dissimilarité (PAM pondéré)

La classification réalisée directement à partir de la matrice de dissimilarité permet de créer des groupes d'individus qui possèdent effectivement une structure locale. Nous avons décidé de challenger cette méthode en utilisant des cartes de Kohonen à des fins exploratoires et de classification.

9.4 *Clustering* sur une matrice de dissimilarité grâce aux cartes de Kohonen

9.4.1 Motivation sur l'usage des cartes de Kohonen

Dans un premier temps, il est possible d'appliquer les cartes de Kohonen sur une matrice de dissimilarité ce qui permet de rentrer dans notre problématique de données mixtes. Il faut préciser que la carte de Kohonen va nous donner une carte composée de neurones. Les prototypes ne sont pas à considérer comme nos clusters finaux et l'analyse par le *silhouette plot* n'est plus adaptée à ce type de problématique. Nous utilisons les outils statistiques présentés dans la partie théorique pour valider le résultat de l'algorithme.

L'avantage des cartes de Kohonen réside dans sa propriété relative à **la réduction de dimension**, tout en préservant **une structure locale** des véhicules grâce à la grille créée. De plus, il est très facile de visualiser et d'interpréter chaque cluster, contrairement à la méthode précédente qui recourt à des méthodes comme t-SNE si l'on souhaite identifier des structures. Ainsi, pour les variables numériques, il est possible d'afficher une *heatmap* en fonction du niveau de la variable pour chaque prototype. Pour les variables qualitatives, nous pouvons afficher à l'aide de diagramme en boite la proportion de véhicules possédant des caractéristiques spécifiques dans un prototype.

Après avoir construit la grille, nous regrouperons des prototypes à l'aide d'une CAH sur la classification établit par la carte de Kohonen. Ainsi, deux prototypes possédant des structures communes pourront être réunis.

9.4.2 Mise en pratique

Nous nous plaçons avec **la même matrice de distance pondérée calculée dans l'étape précédente**. Pour obtenir une grille fiable, nous décidons d'initialiser une grille de dimension 6x5, soit 30 prototypes.

L'optimisation de la taille de la grille n'a pas été envisagée dans notre méthodologie à l'aide de méthode type *grid-search* car le calibrage d'un SOM sur une matrice de dissimilarité de 34 000 véhicules est très conséquent en temps de calcul. De plus, les indicateurs calculés par le package **SOMBrero** ne sont pas adaptés à notre nombre de données. Néanmoins, une attention particulière a été apportée à la réalisation du *clustering* final et à l'interprétation des prototypes.

Nous avons décidé d'utiliser **une grille rectangulaire** munie d'**une fonction de voisinage gaussienne**. La distance utilisée est la distance de Gower et l'algorithme réalise au maximum 500 itérations.

Nous nous sommes posé les questions suivantes :

— Comment la grille est-elle répartie en termes de véhicules ?

— Les prototypes sont-ils éloignés les uns des autres ?

— Comment interpréter les différents prototypes ?

— Comment regrouper les prototypes entre eux pour obtenir un *clustering* final ?

9.4.3 Statistiques de validation de la grille

Les statistiques introduites dans la partie théorique ont été calculées sur la Table 9.2.

Statistiques	**Valeur**
Energie	1.56 %
Erreur topographique	0.05 %
ANOVA F-Score	882.32
Degrés de liberté	29
P-Valeur	0 (***)

TABLE 9.2 – Statistiques de validation pour la carte de Kohonen

L'énergie est la fonction objective à minimiser lors de la réalisation de la grille. Sa valeur est faible. L'erreur topographique est proche de 0, signifiant que chaque véhicule est relativement proche de son voisin le plus proche. Enfin, l'ANOVA permet de rejeter l'hypothèse H_0 où il n'y aurait aucune différence entre les groupes. Ainsi, nous sommes en possession d'une grille comportant des groupes de véhicules différents.

9.4.4 Analyse de la grille calculée par l'algorithme

La distance de Gower est une métrique de dissimilarité comprise entre 0 et 1. Ainsi l'échelle observée sur la Figure 9.21 est relative à cette mesure. Nous observons des individus assez éloignés du centre de la grille, alors qu'ils sont plus proches dans les extrêmes. Les contours de la grille possèdent plus de véhicules que le centre de la grille d'après la Figure 9.20. Chaque neurone constituant la grille possède assez d'individus, il ne nous semble alors pas judicieux d'augmenter la taille de l'espace topologique.

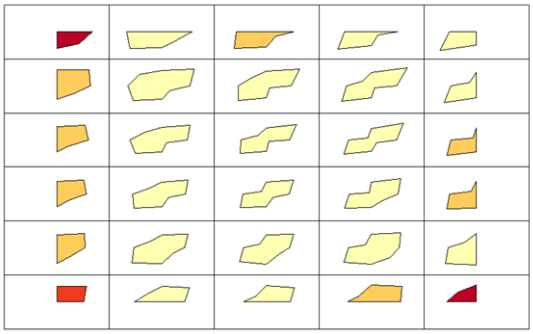

FIGURE 9.20 – Distance entre les prototypes

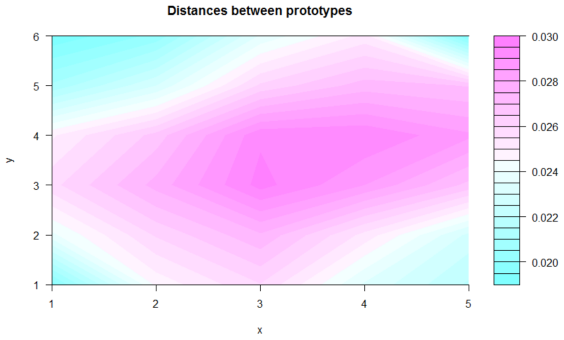

FIGURE 9.21 – Distance lissée entre les prototypes

L'avantage des cartes de Kohonen réside dans l'analyse très simple des prototypes à l'aide d'outil graphique. Ainsi, le cluster 12 possède des véhicules très couteux tandis que le cluster 1 possède des véhicules de moindre coût (Figure 9.22). Enfin, nous observons que les véhicules utilitaires se répartissent sur les extrémités de la grille (Figure 9.23).
Ces analyses vont nous permettre de comprendre et d'interpréter plus facilement les clusters réalisés et permettent à l'assureur d'expliquer de manière visuelle les véhicules présents dans un cluster.

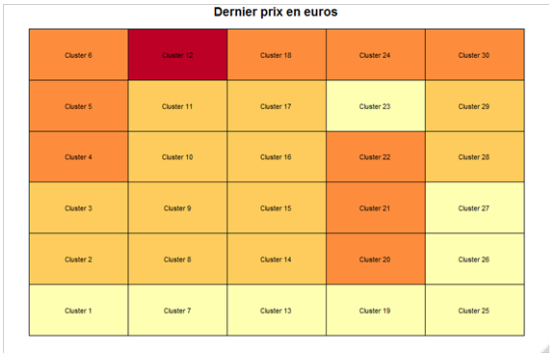

FIGURE 9.22 – Impact de la variable du prix à neuf du véhicule sur les prototypes

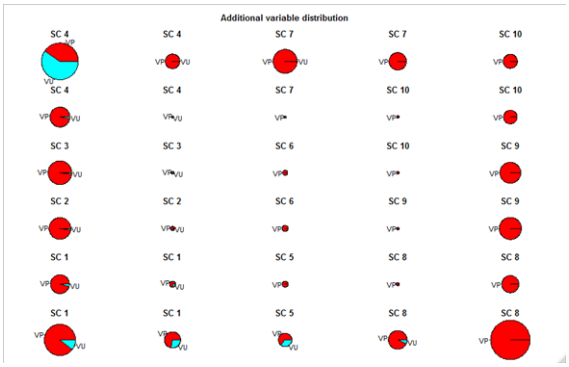

FIGURE 9.23 – Répartition de la variable sur le type de véhicule sur les prototypes

Le package **SOMBrero** permet le regroupement des prototypes en un nombre k de clusters. Nous avons décidé de regrouper les résultats dans 10 clusters. La fonction fait appel à la fonction hclust du package **stats** sous R. Nous gardons la méthode d'agglomération *complete* dont l'objectif est d'assigner des clusters similaires. Une autre option aurait été de choisir la méthode de Ward cherchant à minimiser l'inertie intraclasse.

Ainsi, nous obtenons les clusters finaux. La numérotation s'effectue d'en bas à gauche, jusqu'en haut à droite de la grille. Ainsi le cluster en haut à gauche de la grille est le cluster numéro 4 dans l'analyse suivante.

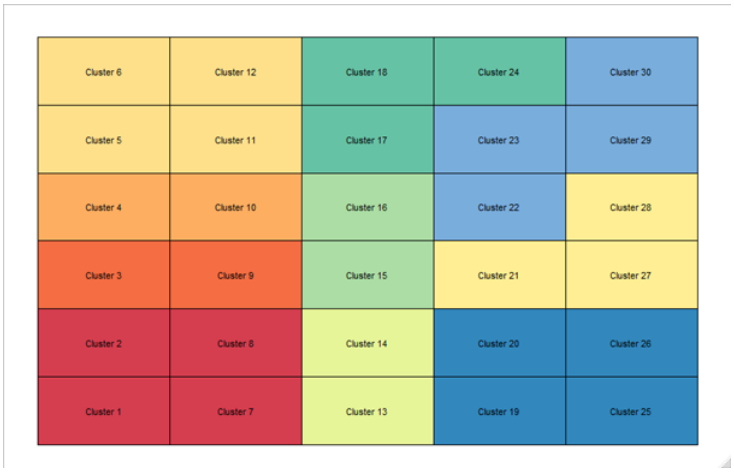

FIGURE 9.24 – Regroupement des prototypes à l'aide d'une CAH

En résumé

L'approche par pondération de la matrice de dissimilarité permet de prendre en compte la sinistralité antérieure.
Les groupes réalisés par PAM possèdent un caractère local plus prononcé avec la pondération que sans la pondération. Les groupes réalisés par Kohonen sont interprétables grâce aux outils graphiques à notre disposition et permettent de présenter une seconde approche de visualisation des données.

9.5 Rattachement des véhicules présentant des valeurs manquantes

Pour assigner un cluster, nous avons réalisé trois approches, dont les deux premières se sont révélées inefficaces pour les méthodes relatives aux matrices de dissimilarité :

— Affectation à un cluster grâce à la distance de Gower.

— Affectation à un cluster grâce aux coordonnées de la carte des véhicules.

— Calibrage d'un arbre de classification CART pour classer les véhicules possédant de l'information manquante.

Les deux premières méthodes brisaient la structure des clusters établie précédemment pour les méthodes issues d'une matrice de dissimilarité. En effet, les niveaux de sinistralité observés se retrouvaient différents et modifiaient l'ordre initial.

Le tableau suivant récapitule les méthodes employées

Algorithme	Méthode utilisé
AFDM	Coordonnées de la carte des véhicules
Matrice de dissimilarité équi-pondéré	CART
Matrice de dissimilarité pondéré	CART
Carte de Kohonen	CART

TABLE 9.3 – Récapitulatif des méthodes choisies pour rattacher les véhicules présentant des valeurs manquantes

Pour la méthodologie relative au CART, nous nous sommes placés dans le cadre de la maille véhicule avec les véhicules complets. Nous avons scindé la base en un échantillon d'apprentissage, composé de 80% des véhicules et un échantillon de test. Un *grid-search* a été effectué pour sélectionner les paramètres optimaux permettant de minimiser l'erreur de classification sur la base test. Ainsi, après élagage de l'arbre à un paramètre de complexité *cp* convenable, nous réalisons les prédictions sur la base des véhicules incomplets et proposer un cluster à ces véhicules, tout en proposant **une méthodologie permettant de limiter le surapprentissage**.

9.6 Analyse des résultats face à la sinistralité observée

Pour juger de la fiabilité de nos *clusterings*, nous avons décidé de proposer deux visualisations des *clusters* face à la sinistralité :

— La comparaison de la fréquence vol moyenne (en gris) pour chaque groupe comparée à la fréquence moyenne du vol pour un tirage aléatoire (en rouge). Cette analyse nous permet de détecter si nos groupes créés ont une influence sur la variable à expliquer.

— Un intervalle (par échantillonnage) de chaque cluster, en tirant aléatoirement 95% des véhicules de chaque cluster 1000 fois. L'intervalle est l'estimateur Monte-Carlo du minimum et du maximum des simulations de la fréquence vol obtenue précédemment. Cette visualisation permettra de vérifier le caractère de risque homogène souhaité lors d'un regroupement de variables.

Ces résultats seront présentés pour une méthodologie, mais les autres méthodes se trouvent dans la Figure 11.6 et suivante en annexe.

Méthode 3 : Carte de Kohonen sur une matrice de dissimilarité pondérée

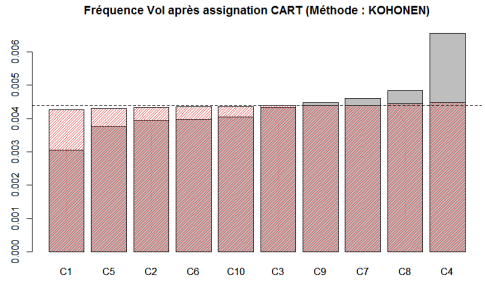

FIGURE 9.25 – Niveau de fréquence vol moyen pour chaque cluster

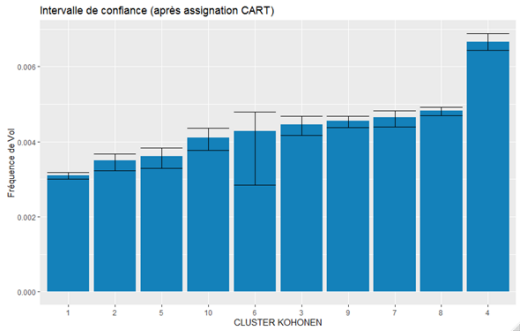

FIGURE 9.26 – Analyse des intervalles crées pour la fréquence vol de chaque cluster

Le *clustering* créé par CAH permet effectivement d'expliquer la fréquence vol observée sur le portefeuille (Figure 9.25). Néanmoins, certains clusters ne sont pas très fiables, comme l'atteste le cluster numéro 6 dont l'intervalle de confiance est plutôt large par rapport à ses voisins. Contrairement aux clusters 1 et 8 qui eux semblent très fiables (Figure 9.26). La volatilité observée sur le cluster 6 s'explique par sa présence au centre de la grille, zone très peu représentée par rapport aux autres clusters. L'intervalle crée est de fait plus large à cause du nombre faible de représentants.

9.7 Conclusions et apports de la méthodologie non-supervisée

Les étapes précédentes nous ont permis de réaliser un *clustering* à partir des informations véhicules. L'information sur les sinistres est rajoutée pour la réalisation de deux véhiculiers. Chaque véhicule est affecté à un cluster. Nous jugerons de la pertinence du regroupement face à la classification SRA dans la dernière partie de ce mémoire à l'aide d'un test de Type III pour la garantie vol.

Nous voyons que les clusters réalisés permettent d'expliquer une variabilité de la fréquence vol. La méthodologie permet de créer des groupes de risques homogènes, quitte à les regrouper pour améliorer les performances en vérifiant la volatilité des groupes obtenus face au critère de notre choix.

La première méthode par AFDM permet de voir que **les deux premiers axes** sont très proches de la classification SRA actuelle. Néanmoins, la classification précédente ne capte pas la présence des véhicules utilitaires dans notre portefeuille.

La seconde méthode par PAM permet de capter une seconde structure, relativement proche de l'AFDM. La pondération par la fréquence vol lors de la réalisation de la matrice de dissimilarité permet de fiabiliser le *clustering* et de créer des clusters plus stables.

Enfin, l'utilisation des cartes de Kohonen permet d'apporter une interprétation supplémentaire à nos clusters. Ainsi, nos groupes créés sont facilement interprétables par l'utilisateur.

Chapitre 10

Construction du véhiculier par approche supervisée

L'objectif de cette partie est de montrer l'adaptation de la théorie développée pour les zoniers pour un véhiculier. Ainsi, nous proposons de capter le risque véhicule résiduel pour un modèle de coût et de fréquence de **la garantie dommage**. Par la suite, seulement la méthodologie pour le modèle de fréquence est détaillée.

Dans un premier temps, nous réalisons une extraction de l'effet véhicule à l'aide d'un GLM, grâce aux variables explicatives identifiées dans les analyses univariées. L'estimateur résiduel de l'effet véhicule est défini par le résidu d'Anscombe :

$$r_i = \frac{3}{2} \frac{y_i^{2/3} - \hat{y_i}^{2/3}}{\hat{y_i}^{1/6}}$$

Cet estimateur est ensuite agrégé au niveau de la maille considérée. L'extraction du signal porté par le résidu demande un travail sur la maille des véhicules choisie et sur la pertinence de celle-ci. De fait, nous avons décidé d'utiliser la carte des véhicules réalisée dans la partie précédente pour lisser le signal véhicule grâce à une interpolation spatiale réalisée par Krigeage.

Une fois le lissage effectué, nous tenterons ensuite de modéliser le signal porté par les véhicules à l'aide d'un *Gradient Boosting Machine* (GBM), méthode de régression d'apprentissage statistique supervisée qui a fait ses preuves dans de nombreuses problématiques et a inspiré de nombreuses méthodes innovantes, dont l'XGBoost. [4]

Nous étudierons l'intégration du signal modélisé par deux méthodes. Pour la première, le résidu est directement intégré pour juger sur les performances de cette nouvelle variable dans la modélisation GLM sous ADDACTIS® Pricing. Pour la seconde, le résidu est segmenté à l'aide d'une CAH puis la segmentation réalisée est introduite dans le GLM.

Le processus développé pour la réalisation du véhiculier supervisé se retrouve sur le schéma 10.1 :

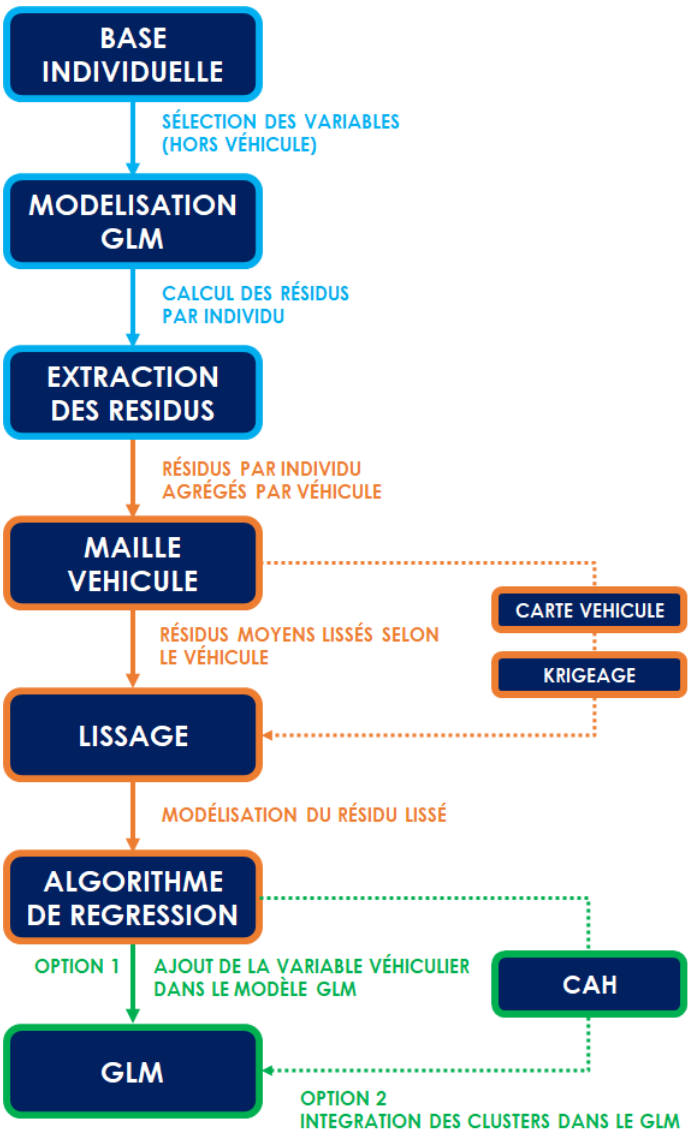

FIGURE 10.1 – Méthodologie employée pour la réalisation du véhiculier supervisé

10.1 Décomposition de la base de travail

Les méthodes relatives à la construction d'un zonier sont vulnérables aux problématiques de surapprentissage. Pour se prémunir de ce risque de modélisation, nous avons décidé de ne garder que les observations du 01/01/2012 au 31/12/2015. Ainsi, dans le cadre supervisé, le schéma suivant (Figure 10.2) est appliqué. L'année 2016 représente notre échantillon de test pour la méthodologie globale. Dans le cadre de la modélisation GLM, nous retiendrons aussi un échantillon d'apprentissage et de test pour vérifier de la fiabilité de la modélisation.

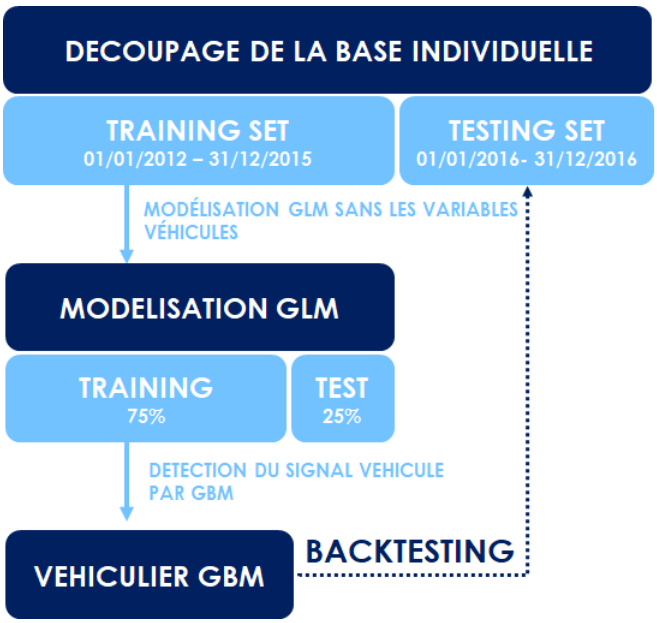

FIGURE 10.2 – Décomposition de la base de travail pour la réalisation du véhiculier supervisé

10.2 Extraction de l'effet véhicule par GLM

Pour capter les facteurs pouvant contribuer à la variation de la fréquence dommage au sein de notre portefeuille, nous avons réalisé des statistiques descriptives

Une première sélection de variables est réalisée par ***stepwise forward***. Cette méthode consiste à ajouter à chaque itération la variable la plus importante au sens du critère de Wald, en partant d'un modèle possédant aucune variable. La condition d'arrêt est de ne plus avoir de variable au-dessus du seuil de pertinence fixé par l'utilisateur.

Nous optimisons le modèle en réduisant le nombre de paramètres par des regroupements et en modélisant les variables quantitatives à l'aide de splines ou de polynômes. L'objectif est de trouver un compromis entre biais et variance :

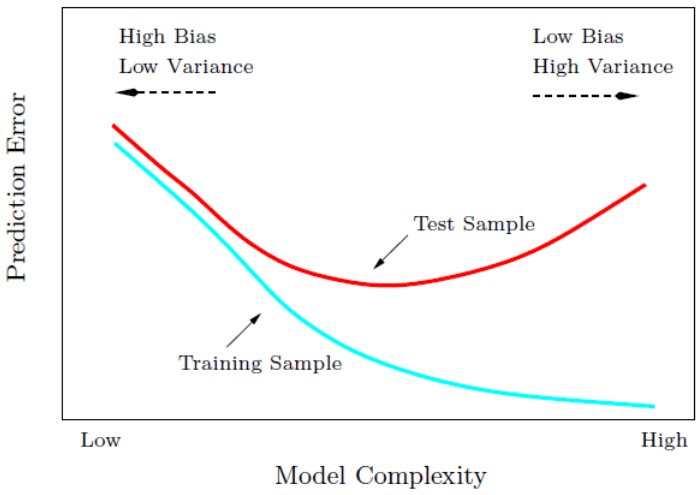

FIGURE 10.3 – Dilemme entre minimisation du biais et de la variance [16]

Le compromis se pose sur la complexité du modèle. Un grand nombre de paramètres p

permet un calibrage adapté de notre modèle à nos données, mais des coefficients β possédant une plus grande variance, se répercutant alors sur la variance du modèle globale. Mais un petit nombre de paramètres donne des coefficients plus précis, mais entraine une augmentation du biais de la modélisation.

Nous choisissons le modèle minimisant l'un de ces critères suivants. Nous utilisons les critères AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion) définis par :

$$AIC = -2l + 2p$$

$$BIC = -2l + p\ ln(n)$$

Où $l = ln(L_{modele})$ représente la log-vraisemblance du modèle, p le nombre de paramètres et n le nombre d'observations.

La Figure 10.4 résume les résultats obtenus. L'âge du véhicule est la première variable qui ressort par rapport à la statistique du Chi-2. Cette variable reflète un critère de comportement client, nous décidons alors de l'intégrer au sein de notre modélisation GLM. De plus, la base SRA ne prend pas en compte la vétusté du véhicule, nous décidons donc de ne pas l'intégrer dans la modélisation de notre véhiculier.

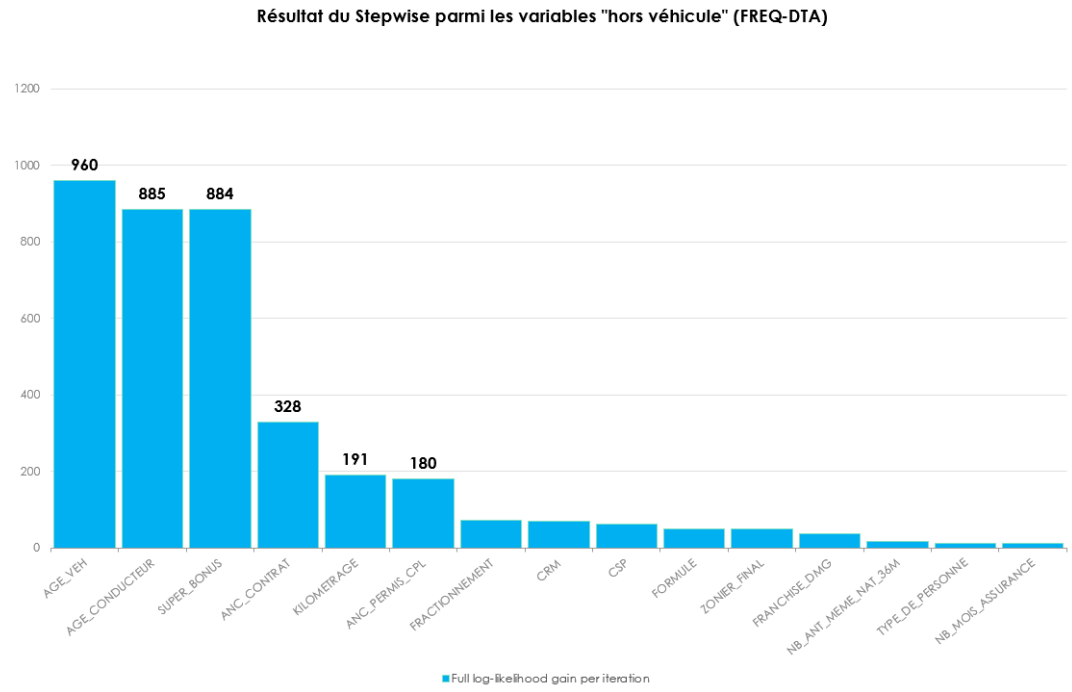

FIGURE 10.4 – Résultat du stepwise forward sur la fréquence dommage.

Une fois nos variables choisies, nous devons procéder à la simplification du modèle en cherchant à lisser les variables continues et en tentant de régulariser notre modèle. L'intégration de ces améliorations est suivie à l'aide des analyses de Type-I pour effectuer judicieusement nos regroupements et juger de la pertinence des coefficients estimés en vérifiant les p-values associées. De plus, il faut vérifier que les effets sont bien captés à l'aide de graphiques montrant les coefficients estimés et leurs intervalles de confiance, mais aussi en comparant la valeur observée sur la base d'apprentissage et la valeur testée sur l'échantillon test.

Finalement, nous retenons le modèle suivant :

	Modèle de fréquence
	Dommage (sans véhicule)
Loi de Y	Poisson
Fractionnement	**Annuel**, Trimestriel, Mensuel
Type de personne	Personne morale / **Individu**
Rachat de franchise	Oui, Non
Nombre de mois d'assurance	**< 36 mois**, 36 mois
Age du conducteur	Spline (numérique)
Age du véhicule	Polynôme (degré 2)
CRM	Spline (numérique)
CSP	Catégorielle
Bonus	[0,85 ; 0,95[, **[0,95,Inf[**
Option kilométrique	**NA** , 5000 km ou 10 000km
Usage du véhicule	Catégorielle
Nombre d'antécédents de sinistres (36 Mois)	Numérique
Ancienneté du contrat	Numérique
Zonier	Catégorielle
AIC	315 661
BIC	316 152
Déviance	248 831

FIGURE 10.5 – Modélisation GLM retenue pour la fréquence dommage

Nous représentons sur la Figure (10.6) la fréquence observée au regard du critère adopté par la SRA face à la fréquence estimée par notre modèle GLM sans variable véhicule. L'effet véhicule n'est pas capté dans les extrêmes, mais celui-ci s'explique à cause de la faible exposition associée. Dans la majorité du portefeuille, le risque véhicule est capté par d'autres variables du fait de la dépendance entre le conducteur et son véhicule. Nous observons alors que la classification SRA ne ressort pas en comparant les valeurs estimées et prédites au regard de ce critère, mais son intégration permet de corriger les véhicules faiblement exposés. **Quelle différence il y a t-il entre notre véhiculier et le véhiculier SRA ?**

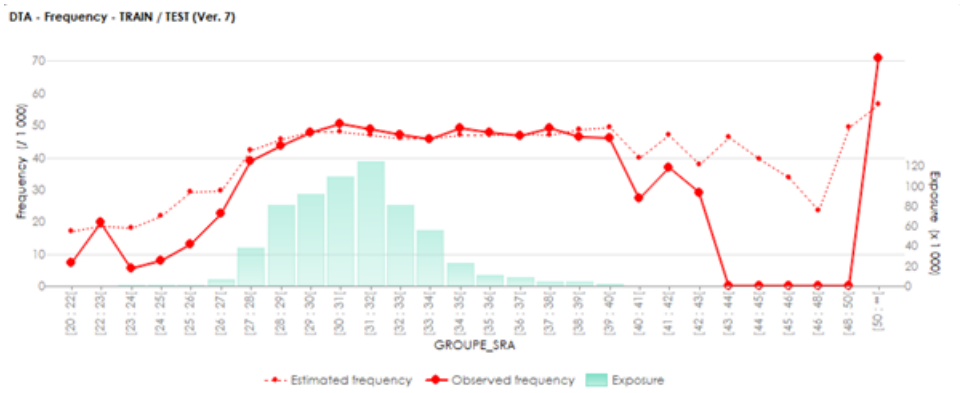

FIGURE 10.6 – Modélisation GLM retenue et comparaison avec le groupe SRA

En résumé

A partir d'une base individuelle comprenant les informations relatives à l'assuré, nous avons calibré un modèle GLM sans prendre en compte les variables relatives à son véhicule (hormis l'âge du véhicule).
L'information portée par le résidu du modèle doit posséder le signal des variables relatives aux véhicules. Nous avons choisi de modéliser **le résidu d'Anscombe** du modèle GLM dans la suite de la méthodologie. Le résidu sera préalablement agrégé à la maille véhicule.

10.3 Lissage du résidu modélisé par Krigeage

10.3.1 Méthodologie générale

Avec les résidus du modèle GLM et la carte des véhicules, nous fiabilisons le signal véhicule présent dans le résidu.

Nous devons trouver une maille sur laquelle agréger le résidu pour ensuite réaliser notre Krigeage. Cependant, ce modèle est très long à calibrer. Ainsi, nous avons cherché un compromis entre rapidité et qualité du lissage.

Pour ce faire, nous agrégeons dans un premier temps notre résidu d'Anscombe modélisé pour chaque variable qualitative. Nous fiabilisons ensuite le résidu en excluant les véhicules peu exposés dans notre base. Ainsi, nous obtenons une maille de 3000 véhicules types. Pour vérifier l'indépendance spatiale des véhicules éloignés, nous allons calculer le semi-variogramme. Nous calibrons ensuite le modèle de Krigeage sur cette base d'apprentissage, puis nous lissons le résidu sur la base de test pour enfin généraliser le processus sur la totalité de la base.

10.3.2 Vérification des hypothèses pour appliquer le Krigeage

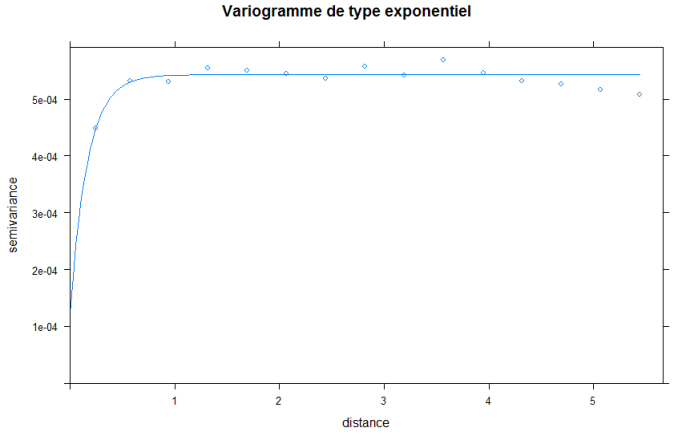

FIGURE 10.7 – Variogramme sur l'échantillon d'apprentissage du Krigeage

Nous observons **un effet pépite** important sur notre semi-variogramme, traduisant une erreur de modélisation de notre signal véhicule. En d'autres termes, le GLM n'a pas réussi à capter seulement le signal véhicule associé : pour deux véhicules relativement proches, celui-ci a aussi capturé du bruit, lié par exemple aux variables explicatives manquantes.

Nous décidons d'utiliser **une structure de noyau** de type *exponentiel généralisé* pour modéliser nos résidus en fonction de notre carte des véhicules.

10.3.3 Lissage des résidus de la base test

Dès que l'échantillon d'apprentissage et la structure de noyaux sont définis, nous exécutons le Krigeage sur la base d'apprentissage.

Comme évoqué précédemment, l'effet pépite est important. Ainsi nous devons vérifier que le résidu lissé est fidèle sur la base test. Nous observons que les résidus des véhicules de la base test ont effectivement été lissés. En effet, chaque point spatialement proche doit avoir une valeur de résidu proche. Sur la Figure 10.8, nous observons des sauts de couleurs discontinus alors que ces sauts ne sont plus présents après lissage (Figure 10.9).

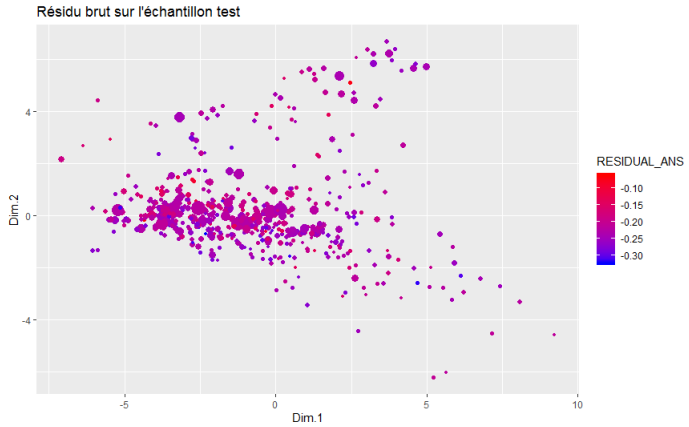

FIGURE 10.8 – Résidu brut sur la base test

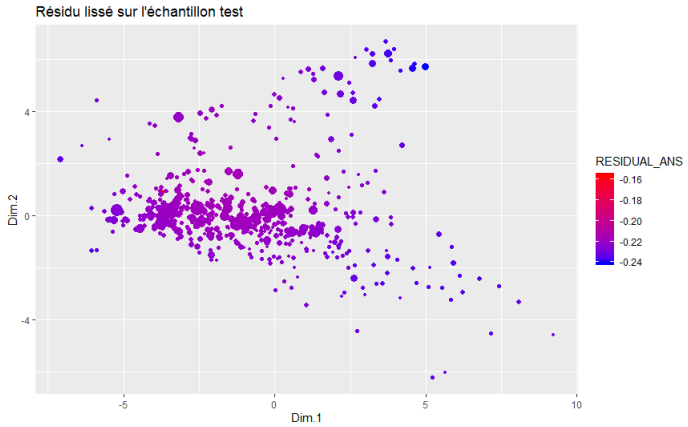

FIGURE 10.9 – Résidu lissé sur la base test

Nous observons effectivement un résidu lissé en fonction de la carte des véhicules permettant de vérifier le résultat du Krigeage.

10.3.4 Comparaison du résidu brut et du résidu lissé

Notre objectif est d'extraire le signal porté par les véhicules dans notre GLM. Nous avons utilisé le Krigeage pour capter l'information portée par les véhicules.

Pour vérifier notre approche, nous avons décidé de visualiser le résidu d'Anscombe du GLM et le résidu lissé (Figure 10.10). Nous observons une réduction de la variance pour la distribution du résidu lissé. Cela signifie que nous avons réussi à capter le signal véhicule de notre GLM dans ce nouveau résidu lissé. Ainsi, le GBM est plus enclin à capter les variations des variables véhicules grâce à cette étape de lissage.

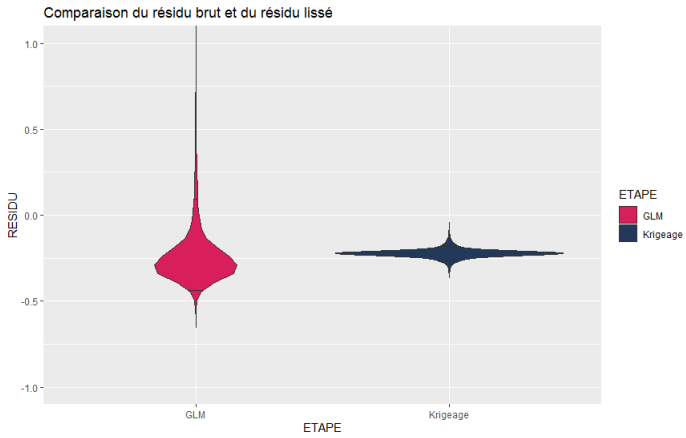

FIGURE 10.10 – Comparaison de la distribution des résidus finaux

En résumé

A ce stade, nous possédons un résidu d'Anscombe lissé pour chaque véhicule. Ce résidu comporte une forte part du signal véhicule. En effet, le résidu initial possède d'autres informations que nous n'avons pas captées dans notre modèle GLM.
Le résidu d'Anscombe lissé par Krigeage est la variable que nous modéliserons dans la suite de la méthodologie.

10.4 Modélisation du résidu lissé par GBM

10.4.1 Méthodologie mise en place

Pour rappel, le GBM appartient à la famille des algorithmes de régression utilisant le principe du boosting.

Lors de la modélisation par GBM, nous avons 4 paramètres à optimiser :

— **La profondeur** des arbres de régression

— **Le nombre d'arbres** à générer pour le boosting

— **Le paramètre d'apprentissage** (aussi appelé *learning rate*)

— **Le nombre minimum** d'observations dans un nœud

Le *learning rate* et le nombre d'arbres sont liés, communément appelé « rythme d'apprentissage ». Un paramètre d'apprentissage faible va nécessiter plus d'arbres pour obtenir un modèle robuste et stable, qui aura correctement appris les données. De fait, pour l'optimisation de nos paramètres, nous avons employé la méthodologie suivante :

1- Fixation du paramètre d'apprentissage à 5%

2- Détermination du nombre d'arbres nécessaire pour éviter le phénomène de surapprentissage.

3- Détermination des paramètres de l'arbre

Nous utilisons le $RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$ comme fonction de perte. C'est cette fonction que l'algorithme cherchera à minimiser à chaque itération. Nous prendrons les paramètres où la moyenne des RMSE sur les échantillons issus de la validation croisée est la plus faible. Enfin, pour juger de la qualité du modèle finale, nous utiliserons deux métriques : le R^2 sur la base d'apprentissage et le Q^2 sur la base de test. Le Q^2 est définit par : $Q^2 = 1 - \frac{MSE}{V(y_{test})}$

Cette métrique sert à vérifier la part du signal que nous arrivons à capter grâce à notre modélisation. Elle servira à accepter ou rejeter le modèle que nous retiendrons à l'aide des critères d'apprentissage précédent.

10.4.2 Optimisation des paramètres par *gridSearch*

Nous avons décidé d'étudier plusieurs combinaisons pour optimiser l'extraction de l'effet véhicule. Dans un premier temps, nous avons fixé le learning rate à 5%. Nous essayons ensuite plusieurs combinaisons.

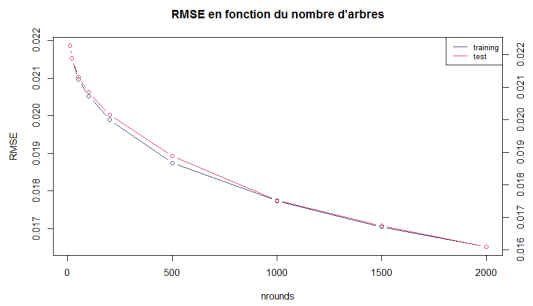

FIGURE 10.11 – Tuning du paramètre relatif au nombre d'arbres

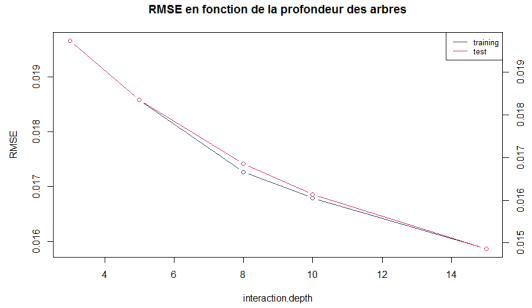

FIGURE 10.12 – Tuning du paramètre relatif à la profondeur de l'arbre

Le phénomène de surapprentissage est observable sur la Figure 10.11 et la Figure 10.12. A partir du moment où le RMSE de la base d'apprentissage se stabilise, nous devons simplifier le

modèle (au risque d'observer du surapprentissage si nous ne réalisons pas cette simplication). A l'aide de cette analyse, nous retenons le modèle présenté dans la section suivante, compromis entre minimisation de la fonction de coût et un modèle simplifié.

10.4.3 Présentation des résultats

R^2 (base d'apprentissage)	87%
Q^2 (base de test)	70%
Learning rate	5%
Nombre d'arbres	2000
Profondeur des arbres	15
Nombre minimum d'observations	1

TABLE 10.1 – Paramètres et résultats de la modélisation GBM

Nous arrivons à capter une forte part (70%) du signal véhicule à l'aide des variables présentes dans la base SRA et de la modélisation GBM sur la base test. Nous généralisons le modèle à toute la base de données pour obtenir un score pour le risque véhicule résiduel.

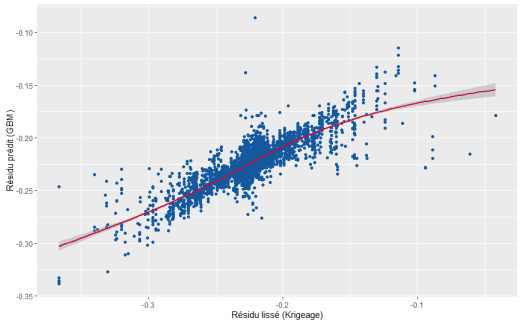

FIGURE 10.13 – Comparaison du résidu lissé avec le résidu modélisé par GBM

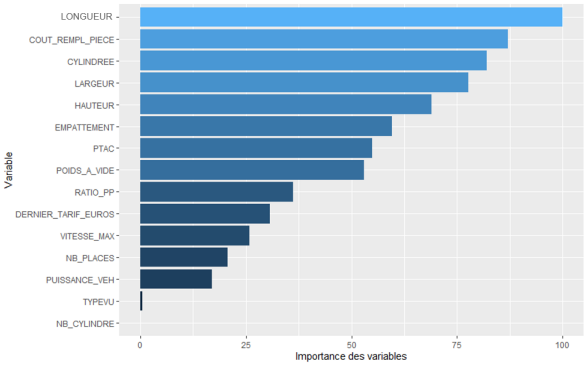

FIGURE 10.14 – Importance des variables pour la modélisation GBM du résidu.

Nous observons sur la Figure 10.13 un résidu fortement centré comme sur la Figure 10.10. L'intégration directe de ce résidu va permettre de corriger quelques individus. Néanmoins, afin d'améliorer les pouvoirs prédictifs du véhiculier crée, il faudrait segmenter le risque véhicule résiduel à l'aide de la fréquence observée pour des croisements que nous n'avons pas étudié, comme la marque et le modèle par exemple. Nous emprunterons des méthodologies utilisées dans la classification non supervisée pour créer des groupes de véhicules.

Parmi nos variables véhicules retenues pour la modélisation, lesquelles sont ressorties lors de la création des arbres ? Les variables relatives à la longueur du véhicule, au prix des pièces, et au cylindrée ressortent dans les trois premières variables contributrices. (Figure 10.14)

Comme présenté sur le schéma 10.1, nous allons dans un premier temps analyser l'impact de l'intégration du résidu modélisé dans notre GLM, puis dans un second temps, nous segmenterons ce résidu en fonction de la sinistralité observée pour créer des groupes de risques homogènes.

En résumé

Nous avons à notre disposition un modèle prédictif permettant d'approcher le risque véhicule résiduel observé sur notre portefeuille à partir des variables relatives au véhicule. Ce modèle se généralise sur l'ensemble de notre jeu de données.

10.5 Intégration du véhiculier supervisé dans le modèle de fréquence

L'intégration de la variable véhiculier dans le GLM (Figure 10.5) va permettre d'une part, d'interpréter le risque véhicule présent dans notre portefeuille et de mesurer l'impact de l'ajout de cette variable dans l'équation tarifaire.

10.5.1 Intégration de la variable SRA

Dans un premier temps, la variable SRA a un impact marginal sur notre portefeuille. En effet, le graphique présenté en Annexe 11.11 montre que l'impact du véhiculier actuel sur la fréquence dommage est marginal sur la zone fortement exposée. De fait, ce véhiculier ne pourra que capter l'effet véhicule sur les groupes faibles.

De plus, lors de la modélisation du GLM sans variable véhicule, nous avons vu que certaines variables conducteurs captent les effets de certaines variables véhicules. Cela signifie que le risque véhicule est aussi capté par d'autres variables présentes en portefeuille, comme le super bonus affecté aux conducteurs qui ont un CRM à 50% et qui ne déclarent pas de sinistres sur une période triannuelle. Ainsi, même si le groupe SRA est significatif dans les analyses de Type-III, la valeur du test de Chi-2 relative à cette variable reste très faible par rapport à d'autres variables de notre modèle.

La part du signal portée par la variable SRA est négligeable sur la majorité du portefeuille. Nous intégrons le véhiculier réalisé dans l'équation tarifaire pour voir l'impact de celui-ci sur notre modélisation. Nous avons décidé de regrouper le groupe SRA pour obtenir les coefficients β_{SRA} associés en rouge (Figure 10.15).
L'amplitude des bêtas varie de -47% à 10%, soit **une amplitude de 57%**. Les intervalles de confiance des coefficients pour les groupes supérieurs à 33 montrent que l'effet véhicule est proche du groupe de référence (ici 28).

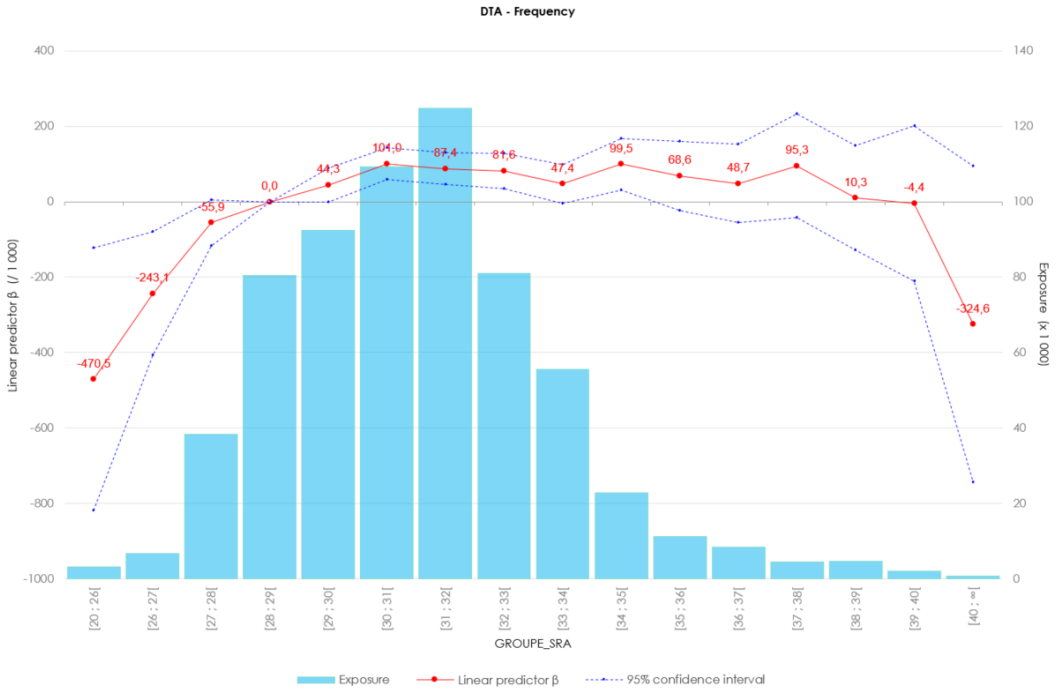

FIGURE 10.15 – Coefficients relatifs à la variable SRA sur le modèle GLM

10.5.2 Intégration de la variable supervisée

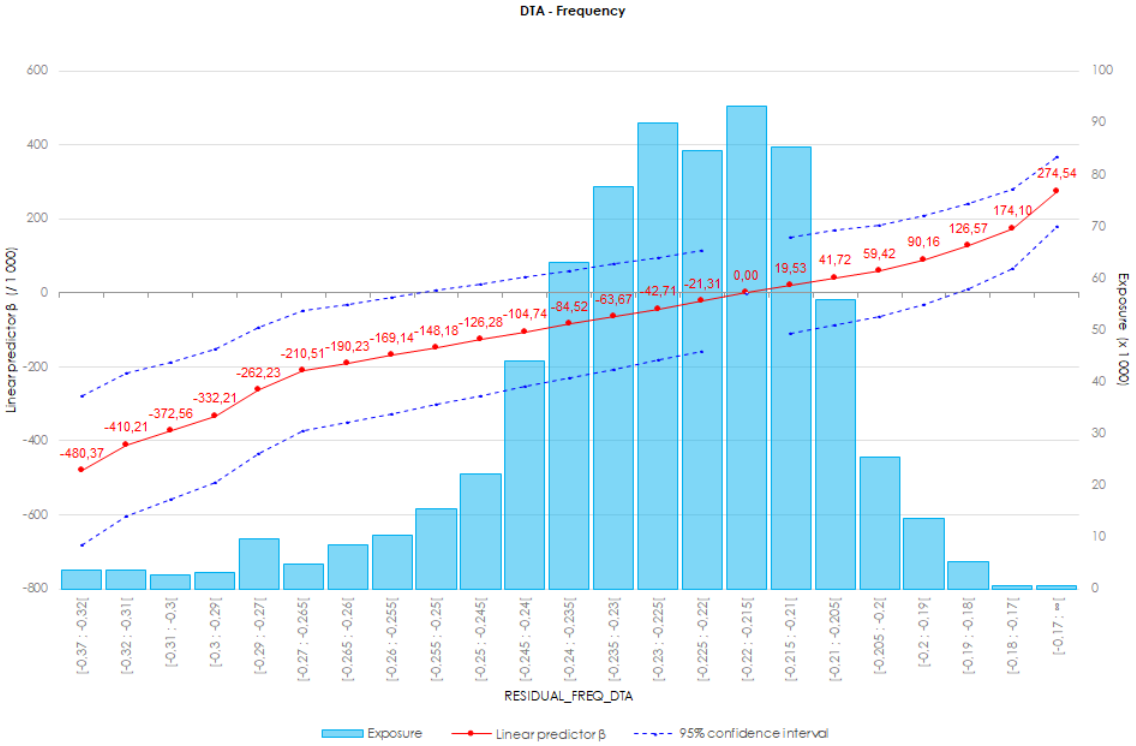

FIGURE 10.16 – Coefficients relatifs à la variable véhiculier sur le modèle GLM

Les coefficients β_{Veh} (Figure 10.16) sont croissants en fonction du résidu modélisé. Ainsi, cela signifie que plus la valeur du résidu est forte, plus la fréquence de l'individu considéré est

élevée. Cette relation est cohérente avec la structure des résidus d'Anscombe :

$$r_i = \frac{3}{2}\frac{y_i^{2/3} - \hat{y_i}^{2/3}}{\hat{y_i}^{1/6}}$$

Ainsi, l'étape du Krigeage a bien permis de capter un effet véhicule proportionnel avec la sinistralité observée. De plus, nous voyons une amplitude des β_{Veh} allant de -48% à +27%, soit **une amplitude de 75%**.

10.5.3 Comparaison du véhiculier supervisé et du véhiculier SRA

La relation linéaire n'est pas observée pour le groupe SRA à cause d'un plateau observé sur la majorité de la zone exposée. De plus, les coefficients observés sur la Figure 10.15 montrent que des groupes différents ont des sinistralités assez proches. Ainsi, nous retombons sur une problématique soulevée précédemment : le groupe SRA n'est pas adaptable pour toutes les garanties afin de modéliser la fréquence observée sur une garantie.

Même si le groupe SRA est interprétable grâce à son lien direct avec la puissance du véhicule, elle n'est pas aussi segmentante que la variable créée par notre modélisation. Cette nouvelle variable a l'avantage d'être interprétable [1] et de posséder une frontière moins stricte que des groupes : comment distinguer la différence entre le groupe 30 et 31 au regard du critère SRA ?

10.6 Segmentation du résidu lissé par CAH et intégration dans le modèle de fréquence

Nous avons à notre disposition pour chaque véhicule, un résidu moyen lissé obtenu par GBM et la fréquence moyenne de la garantie dommage observée pour chaque véhicule.

Nous voulons une segmentation du résidu lissé interprétable pour les équipes opérationnelles. De fait, nous utilisons le croisement Marque x Modele x Version pour agréger notre résidu lissé et la fréquence observée pour chaque croisement.

10.6.1 Mise en place de la CAH

Pour chaque croisement, nous calculons les indicateurs suivants sur lesquels nous allons réaliser la CAH :

— Le résidu moyen lissé

— La fréquence moyenne observée

Comme le croisement choisi permet de réduire le nombre de véhicules, nous réalisons une Classification Ascendante Hiérarchique (CAH) pour segmenter l'espace observé en fonction de la distance entre nos individus par rapport aux indicateurs calculés précédemment. Nous souhaitons créer plusieurs paquets : nous avons alors fixé le nombre de groupes à 25, quitte à regrouper par la suite les groupes dont l'exposition est faible ou les groupes possédant des fréquences similaires.

1. Le recours à un métamodèle est envisageable pour justifier de cette interprétabilité

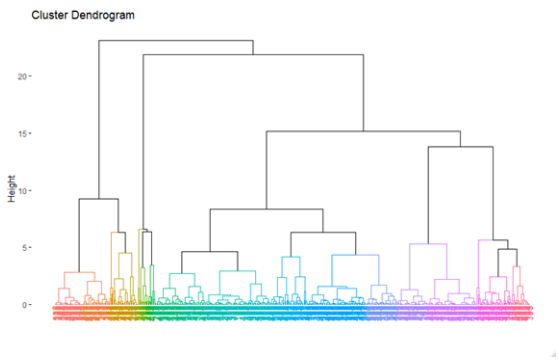

FIGURE 10.17 – Dendrogramme de la Classification Ascendante Hiérarchique

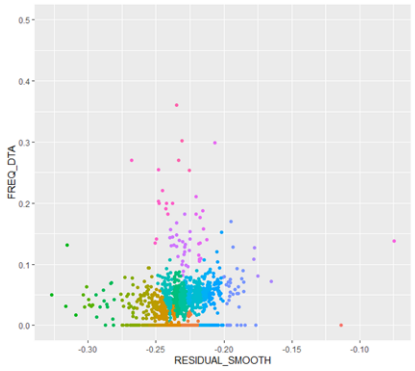

FIGURE 10.18 – Visualisation des points sur le regroupement Marque x Modele x Version

La Figure 10.17 nous présente le résultat de la classification. Celle-ci permet de repérer des structures de risque homogène, en croisant les caractéristiques techniques et la sinistralité observée. Il reste à analyser les segments obtenus pour voir si ceux-ci ne sont pas aberrants.

Pour vérifier le *clustering*, nous affichons les véhicules qui se trouvent dans un même groupe sur le graphique 10.18. En effet, chaque point représente un croisement, nous pouvons afficher les étiquettes relatives à chaque véhicule pour vérifier que deux modèles ne sont pas aberrants. Cette analyse a été réalisée, mais n'est pas présentée pour des raisons de lisibilité.

10.6.2 Comparaison avec l'introduction brute du résidu

Le *clustering* réalisé permet de détecter des segments non détectées avec le signal véhicule (Figure 10.19). En effet, le modèle LEXUS ou XEDOS 6 se situe là où la distribution du résidu est centrée. Cela est rassurant, car le résidu n'a pas capté ce signal. Ni la carte des véhicules ni le GBM ne possèdent l'information relative aux marques et aux modèles des véhicules. Pourtant, l'ajout de la fréquence observée va permettre d'améliorer le *clustering* et nous espérons une amélioration significative du GLM lors de l'intégration de la nouvelle variable créée.

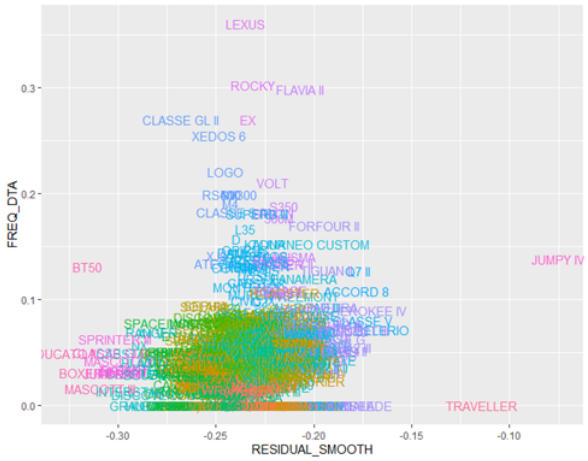

FIGURE 10.19 – Visualisation des modèles de véhicules en fonction de la sinistralité et du résidu

Néanmoins, certains véhicules sont peu représentés dans notre base ou bien ne possèdent aucune sinistralité. Ceux-ci risquent d'être avantagés dans notre modélisation.

Nous observons une amélioration marginale des scores AIC et BIC lors de l'intégration dans le GLM de la nouvelle variable véhiculier.

Statistiques	Brut	CAH
AIC	315 473	315 270
BIC	315 989	315 933
Déviance	248 640	248 412

TABLE 10.2 – Statistiques comparatives entre le modèle avec le résidu brut et le résidu segmenté.

10.7 Comparaison entre les trois véhiculiers sur la fréquence dommage

La modélisation des différents véhiculiers a été effectuée dans la partie précédente. Nous avons en notre possession un véhiculier non supervisé, et un véhiculier supervisé segmenté à l'aide d'une CAH, et un véhiculier SRA.

Afin de comparer de l'apport de ces véhiculiers sur la modélisation GLM, nous allons intégrer chacune des variables dans le modèle GLM sans variable véhicule, et analyser leurs apports d'une part sur la base d'apprentissage, et d'autre part sur la base test. (Figure 10.2)

10.7.1 Comparaison sur la calibration du modèle (base d'apprentissage)

	Comparaison des véhiculiers sur la calibration du modèle (Base d'apprentissage)			
	Sans véhicule	Véhiculier SRA	Véhiculier supervisé	Véhiculier Non-supervisé
Véhiculier	X	Groupe SRA	CAH résidu	AFDM / CLARA
AIC	315 661	**315 587**	**315 270**	315 487
BIC	316 152	**316 272**	**315 933**	316 272
Déviance	248 831	**248 723**	**248 412**	248 609

FIGURE 10.20 – Statistiques comparatives du modèle sans véhicule et des modèles incluant différents véhiculiers

Le véhiculier non-supervisé réalisé par AFDM a un pouvoir explicatif très proche du véhiculier SRA. Ce résultat s'explique par les typologies très proches des clusters avec la classification élaborée par la SRA. Le véhiculier supervisé améliore faiblement les statistiques de validation. Néanmoins, son interprétation est facile et la modélisation est plus lisse qu'une classification comme la SRA, affectant un véhicule à un groupe.

Cependant, le regroupement par CAH est problématique si nous n'avons pas en notre possession un historique pour le véhicule considéré, car il est alors impossible d'effectuer une classification pour un nouveau véhicule. Il faut alors avoir recours au véhiculier brut utilisant seulement la valeur du résidu prédit par GBM.

10.7.2 Pouvoir prédictif du véhiculier supervisé sur la base test

Pour rappel, la base test est l'année 2016 qui a été exclue du calibrage du modèle GLM. La stabilité dans le temps du véhiculier a été vérifiée en introduisant une interaction entre l'exercice et la variable véhiculier.

	Comparaison des véhiculiers sur la base test (Fréquence dommage)	
	Véhiculier SRA	Véhiculier supervisé
MSE	0.169321	0.169311
RMSE	0.411486	0.411474

FIGURE 10.21 – Comparaison du modèle SRA et du modèle véhiculier sur la base test

Les facteurs de la Figure 10.21 ont été retenus pour juger de la fiabilité du nouveau véhiculier réalisé. Notre modèle permet de minimiser les différents indicateurs (MSE / RMSE) mais dans de faibles proportions. Nous nous interrogeons sur la cause expliquant la faible baisse des indicateurs :

- **Les problématiques autour de la donnée** : nous n'avons pas pu capter tout l'effet conducteur et géographique. Par exemple, nous n'avons aucun recul sur la construction de la variable zonier et celle-ci n'a pas un pouvoir explicatif important par rapport à d'autres variables comme l'âge du véhicule.
- **La modélisation de notre GLM** : Nous essayons de modéliser un phénomène comportant beaucoup d'aléas.
- **Le choix de la garantie** : Certaines garanties peuvent être plus ou moins adaptées à la modélisation du risque véhicule. Sur notre portefeuille, la variabilité de l'effet véhicule avec le véhiculier SRA est marginale sur la fréquence dommage, contrairement à la fréquence bris de glace où l'impact du véhiculier SRA est plus conséquent.
- **Les interactions entre les variables** : Le véhicule est fortement lié à son utilisateur. Il est possible que les variables relatives aux conducteurs capturent une part d'information relative au véhicule, en complément de l'effet marginal capté par le véhiculier. En effet, le signal porté par les variables véhicules est déjà majoritairement expliqué par d'autres variables (conducteurs ou géographiques). De fait, cette information ne se retrouve pas dans le résidu modélisé.

En résumé

L'intégration du véhiculier supervisé dans le modèle de fréquence apporte une quantité d'informations supplémentaire par rapport à la classification SRA. L'amplitude des coefficients est plus forte, mais surtout, l'interprétation de la pente est plus facile avec la variable véhiculier que la variable SRA.

Le choix de la segmentation du résidu modélisé par GBM est un choix de modélisation. Ce choix perd en interprétation, car le résidu n'est alors plus continu, mais permet d'approcher plus fidèlement le risque véhicule.

Même si les améliorations sont minimes sur la base des statistiques calculées, il faut rappeler que ces indicateurs n'indiquent pas quel véhiculier choisir afin d'approcher correctement le risque véhicule.

Chapitre 11

Présentation des résultats et comparaison entre véhiculiers

Nous avons décidé de modéliser la prime pure dommage pour vérifier l'impact du véhiculier. En effet, l'impact est marginal sur la fréquence dommage. Nous pensons que la réalisation d'un véhiculier pour le coût est plus pertinente, étant donné que le coût de réparation d'un véhicule pour un sinistre responsable est fortement lié avec la typologique du véhicule. Nous appliquons le même procédé sur le coût dommage en adaptant le résidu d'Anscombe pour une distribution de loi Gamma. La méthodologie étant similaire, nous ne présentons pas les étapes réalisées précédemment sur la fréquence dommage. Les résultats de la modélisation finale pour le coût moyen seront présentés puis nous comparerons les modélisations des primes pures pour les deux modèles de véhiculiers.

Quant à la garantie vol, nous avons intégré le véhiculier réalisé dans la partie non supervisée dans un modèle de fréquence vol pour juger de la pertinence de l'approche non supervisée : permet-elle, comme la fréquence dommage, de répliquer seulement la classification SRA ou d'apporter plus d'informations ?

11.1 La garantie dommage

11.1.1 Résultat pour le modèle de coût

Le graphique avec les coefficients β pour le véhiculier SRA et le véhiculier supervisé se trouvent sur la Figure 11.12 et la Figure 11.13 en annexe . Nous voyons que la tendance observée est très similaire : plus la valeur du résidu augmente, plus les coefficients β relatifs à la valeur du résidu sont élevés.

Nous comparons la modélisation du coût sur la base test. Nous observons un faible écart entre nos métriques de validation sur l'échantillon test pour nos deux modèles, justifiant la robustesse de notre modélisation (Figure 11.1). Nous invoquons la même raison que le modèle de fréquence : tout comme le véhiculier SRA (Figure 11.10), le véhiculier crée capte une faible variation du coût moyen sur la zone fortement exposée. Ainsi, notre véhiculier permet de réduire le RMSE de la modélisation de 4 €. Néanmoins, il faut analyser ces différences sur la modélisation de la prime pure, certains individus se retrouvent pénalisés tandis que d'autres individus seront avantagés.

De plus, au niveau de l'intégration au sein du GLM, le nouveau véhiculier est lisse et facilement interprétable. En effet, celui-ci ne nécessite aucun avis d'expert pour comprendre la

classification réalisée à l'aide d'un boosting d'arbres. Nous possédons tous les outils nécessaires pour interpréter le modèle : analyse de l'effet marginal d'une variable, analyse des arbres générés ...

	Comparaison des véhiculiers sur la base test (Coût-moyen dommage)	
	Véhiculier SRA	Véhiculier supervisé
MSE	2 492 670	2 482 318
RMSE	1579	1575

FIGURE 11.1 – Comparaison du modèle SRA et du modèle véhiculier sur la base test

La modélisation du risque véhicule est compliquée du fait de la concentration des véhicules autour de la valeur moyenne du résidu. Comme nous avons pu le voir lors de la réalisation de la carte des véhicules, beaucoup de véhicules ont des caractéristiques très proches et il est impossible de capter ces variations au regard de la fréquence, ou bien du coût moyen sur notre portefeuille.

11.1.2 Résultat pour la prime pure

Nous avons décidé de modéliser la prime pure en multipliant les résultats des modèles de fréquence et de coût moyen. Nous avons ensuite lissé ce résidu en fonction de l'âge du véhicule et de l'âge du conducteur.

La prime lisée obtenue est comparée entre les deux véhiculiers (Figure 11.2). La différence des deux véhiculiers se retrouve majoritairement dans les extrêmes. En effet, notre véhiculier va pénaliser plus fortement les véhicules ayant une classe comprise entre A et I, alors qu'il va être plus avantageux pour les véhicules supérieurs à la classe M.

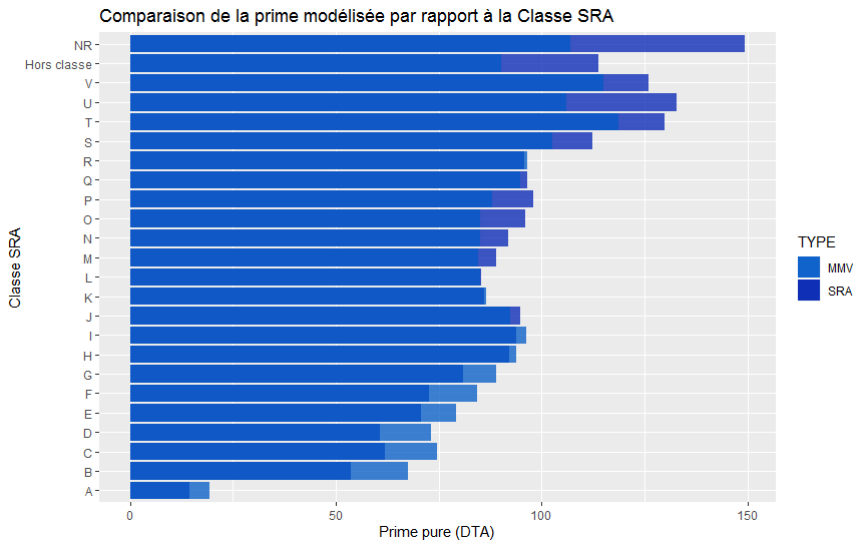

FIGURE 11.2 – Evolution de la prime pure dommage modélisée pour les deux véhiculiers en fonction de la classe SRA

Chaque véhiculier va se concentrer sur une partie précise de la population. Notre véhiculier a l'avantage de ne pas pénaliser les véhicules classés dans de hautes classes. Les tests statistiques

effectués plus tôt ne permettent pas de départager les deux véhiculiers. Chaque véhiculier apporte sa quantité d'informations et dépendra de l'avis de l'assureur.

11.2 La garantie vol

Pour rappel, nous avons réalisé dans la partie non supervisée plusieurs véhiculiers pour la garantie vol :

— Un premier véhiculier a été réalisé à l'aide de l'algorithme CLARA à partir de la carte des véhicules. Ce véhiculier a l'avantage de ne pas avoir en entrée les variables relatives à la sinistralité.

— Deux véhiculiers utilisant des techniques de *clustering* à partir d'une matrice de dissimilarité pondérée par la sinistralité observée. Le premier véhiculier a été réalisé à l'aide de l'algorithme PAM tandis que le second véhiculier a été réalisé à l'aide d'une carte de Kohonen.

Pour la réalisation du modèle GLM, nous avons utilisé la même méthodologie que pour la garantie dommage. :

— Réalisation d'un stepwise pour sélectionner les variables explicatives du modèle.

— Lissage des variables (zonier) pour simplifier le modèle.

— Validation du modèle (Analyse de Type I et de Type III).

11.2.1 Impact de la nouvelle variable sur un modèle de fréquence

Encore une fois, nous avons à notre disposition la classification SRA comme véhiculier d'usage. De fait, nous allons comparer le modèle établi dans l'étape précédente avec chaque véhiculier à notre disposition. Dans un premier temps, nous jugeons de la pertinence de l'ajout de la variable véhiculier dans le modèle en réalisant un test de Type-III.

L'ajout de cette variable est significatif et se positionne en seconde position, derrière la variable relative au zonier, selon la statistique du Chi-2, dans le cas du véhiculier réalisé par PAM. Ensuite, nous comparons l'impact de la nouvelle variable en la calibrant sur l'échantillon d'apprentissage.

	Comparaison des véhiculiers sur la calibration du modèle (Base d'apprentissage – Fréquence Vol)			
	Sans véhicule	Véhiculier SRA	Véhiculier PAM	Véhiculier KOHONEN
Véhiculier	X	Groupe SRA	**PAM**	KOHONEN
AIC	62 077	62 026	**61 937**	61 971
BIC	62 468	62 556	**62 429**	62 438
Déviance	53 371	53 297	**53 214**	53 252

FIGURE 11.3 – Comparaison des différentes métriques de validation du GLM pour les véhiculiers réalisés

Le véhiculier réalisé par PAM sort du lot, même si encore une fois, l'impact est marginal sur la base d'apprentissage. Ainsi, nous comparons les résultats sur l'échantillon test. Nous obtenons les scores suivants pour les statistiques de validation usuelles (MSE/RMSE) :

	Comparaison des véhiculiers sur la base test (Fréquence vol)		
	Véhiculier SRA	Véhiculier PAM (pondéré)	Véhiculier Kohonen (pondéré)
MSE	0.024023	0.023539	0.023540
RMSE	0.154995	0.153426	0.153427

FIGURE 11.4 – Comparaison des différentes statistiques de validation sur la base test pour la fréquence vol

L'impact du véhiculier est faible, du fait de la concentration des individus. Le véhiculier isole des structures que nous avons identifiées dans l'analyse non supervisée (mais ce n'est qu'une minorité du parc étudiée). L'effet véhicule n'est pas aussi important que l'effet géographique dans notre jeu de données pour la garantie vol. De plus, le véhiculier PAM et le véhiculier Kohonen ont un pouvoir prédictif similaire.

11.3 Apprentissage des méthodes employées dans le cadre de la réalisation d'un véhiculier

Afin de modéliser correctement l'effet propre au véhicule, il est essentiel de pouvoir subdiviser notre population. La SRA fournie un identifiant véhicule qui n'était pas présent dans notre base d'étude. Nous avons créé notre propre maille, tout en ignorant les facteurs relatifs à la marque et au modèle des véhicules lors du calibrage de nos modèles et en prêtant une attention particulière aux caractéristiques techniques (carrosserie, alimentation, puissance du véhicule, taille du véhicule ...).

L'utilisation de la marque ou du modèle nécessite d'une part la complétion des valeurs manquantes (beaucoup de marques peu exposées sont très mal renseignées sur la base SRA) et le traitement de celles-ci. Ainsi, dans le calibrage de notre modèle GBM, nous n'avions pas à notre disposition la totalité du parc automobile de notre portefeuille, mais seulement la partie correctement renseignée par la SRA. Un avis d'expert peut être utilisé pour les véhicules possédant de l'information manquante afin de les rattacher à des véhicules présents.[1] L'intégration de ces éléments peut venir en amont de la modélisation.

Outre ces mesures quant à l'identification des véhicules, nous sommes en capacité de réaliser une règle d'affectation des véhicules, comme établit par la SRA, soit en utilisant soit les modèles créés, soit en réalisant un métamodèle qui essaiera de répliquer le plus fidèlement possible la modélisation effectuée pour le véhiculier.

Pour résumer l'apport de ce mémoire, nous allons passer en revue les étapes clés de la modélisation, leurs apports dans la méthodologie adoptée, mais surtout leurs faiblesses, pour pouvoir remédier et apporter une ouverture aux problématiques rencontrées au cours de ce mémoire.

1. L'affectation à l'aide d'un CART ou la distance entre les véhicules n'est pas optimale pour garder la structure initiale des clusters.

11.3.1 L'apport de la méthodologie non-supervisée

La création de la carte des véhicules

La première notion clé abordée dans ce mémoire est la carte des véhicules utilisée d'une part pour explorer notre base de données et d'autre part afin de lisser le résidu modélisé à l'aide d'un modèle GLM.

Cette carte nous a permis de détecter des segments sous-tarifés qui ont été mis en lumière dans la partie supervisée, comme les véhicules possédant une classification SRA faible.

Dans un second temps, cette carte a permis de montrer que notre hypothèse relative à la bonne spécification du modèle GLM sans variable véhicule pour l'extraction de l'effet véhicule n'est pas vérifiée : l'effet pépite présent sur le semi-variogramme affirme que l'erreur de modélisation entre deux véhicules proches est forte. Ainsi, deux véhicules similaires peuvent avoir un résidu différent. L'étape de lissage permet de fiabiliser l'information pour que le GBM capte un effet véhicule à partir des variables véhicules à notre disposition.

Les causes pouvant expliquer ce phénomène La cause principale à l'aléa du phénomène modélisé et à l'absence d'informations nécessaires pour approcher fidèlement le risque observé (Sexe du conducteur, recul sur le modèle tarifaire en vigueur et sur le produit étudié ...).

Cette erreur de modélisation est d'autant plus discutable que les véhicules sont regroupés au centre de la carte. L'orthogonalité engendrée par la méthode AFDM incite à cette concentration. Ainsi, nous nous retrouvons avec des valeurs du résidu final très concentrées autour de la moyenne. Une pondération de l'AFDM permettrait d'amoindrir ce phénomène de concentration. La concentration des individus nous dirige vers deux conclusions plausibles :

— Les véhicules se ressemblent et empêchent le GLM de capter un signal propre à chaque zone de la carte.

— Le signal véhicule est faible. C'est ce que la variable SRA va nous indiquer, compte tenu de sa faible importance dans nos modèles actuels. Le GLM a capté très peu d'informations véhicule et capte plus du bruit.

Pouvoir explicatif de la classification SRA

Nous avons vu que la classification SRA ressort sur les deux premiers axes de l'AFDM. Cela signifie que la classification actuelle utilise beaucoup d'informations sur les véhicules. Ainsi, la variable SRA est fiable pour approcher le risque véhicule d'un point de vue technique.

De plus, nos véhiculiers modélisés permettent certes d'optimiser nos modèles, mais dans de moindres mesures. Ainsi, l'intérêt en matière de temps de paramétrage et d'opérationnalité semble très limité ici pour mettre en place un véhiculier ayant pour principal objectif de substituer la classification SRA.

11.3.2 L'apport de la méthodologie supervisée

Le Krigeage

Le Krigeage, en complément de l'utilisation de la carte des véhicules, a permis de fiabiliser l'information en amont du GLM.

Le Krigeage est une méthode d'interpolation spatiale nécessitant peu de données pour obtenir de bons résultats. En effet, le modèle minimise la variance et ne possède pas de biais lors de la calibration : peu de données peuvent donner un bon pouvoir de généralisation. Néanmoins, nous avons beaucoup de véhicules et nous ne sommes pas en capacité d'affirmer que la maille d'observation choisie est la plus optimale dans notre problématique de lissage du résidu.

Notre point de vue a néanmoins permis de fiabiliser l'information, malgré l'effet pépite présent lors de l'analyse du semi-variogramme.

Nous pouvons aussi nous questionner sur la pertinence de la méthode du Krigeage si nous avions mélangé les garanties avec des typologies différentes (exemple : dommage et bris de glace). Fort est à parier que l'effet pépite aurait été d'autant plus grand, le modèle GLM étant incapable de capter les variations entre deux typologies de sinistres totalement différents, sur la base des caractéristiques propres à l'assuré.

L'apport du véhiculier supervisé par rapport au véhiculier SRA

Le véhiculier SRA permet d'extraire une part d'information véhicule sur nos garanties. L'effet véhicule est très similaire sur la majorité de nos véhicules assurés (autour de la moyenne), le pouvoir explicatif du véhiculier SRA ne ressort pas dans les analyses de Type-III.

Le véhiculier supervisé a permis de pallier à cela en expliquant l'effet véhicule sur les véhicules présents dans les extrêmes, d'une manière plus poussée que le véhiculier SRA. Ainsi, nous obtenons une modélisation de la prime pure relativement différente sur les extrêmes.

Néanmoins, l'extraction de l'effet véhicule pourrait être meilleur, si nous changeons la maille d'observation pour éviter le phénomène de concentration des véhicules (ce qui semble assez compliqué étant donné la forte similarité des véhicules présents sur le marché). Un véhiculier pour le produit moto est plus sujet aux véhicules présents dans le portefeuille, qu'un véhiculier automobile qui va comporter beaucoup de véhicules très similaires. Ainsi, le véhiculier automobile établit par la SRA se généralise plus facilement sur un portefeuille automobile qu'un véhiculier moto, qui va réellement dépendre de chaque acteur.

L'analyse des résidus et l'intégration de ceux-ci permettent de corriger les effets qui n'ont pas été captés par le GLM. Si nous arrivons à apporter une part d'information supplémentaire que la SRA dans les extrêmes et une information similaire sur la globalité du portefeuille, cela signifie que notre méthodologie fonctionne pour extraire l'effet véhicule.

11.3.3 Véhiculier technique ou véhiculier commercial ?

Nous avons réalisé dans cette étude un véhiculier technique. Un véhiculier commercial s'utilise pour lisser le tarif en fonction d'un critère préalablement choisi, ou pour effectuer une opération d'optimisation tarifaire. La variable créée à l'aide de notre approche peut servir de facteur de lissage, la variable étant continue. Elle créera alors moins de sauts qu'un facteur qualitatif qui pourrait varier entre deux régions du risque.

De plus, la pondération de l'AFDM n'a pas été explorée dans ce mémoire, mais celle-ci pourrait permettre de réaliser un lissage complété avec un avis d'expert, ou un avis des équipes marketing. Plusieurs briques de ce mémoire peuvent être optimisées pour résoudre les problématiques que nous avons rencontrées au fur et à mesure et qui peuvent améliorer considérablement les performances de notre véhiculier.

11.4 Réflexion sur la mise en place opérationnelle

La mise en place opérationnelle d'un véhiculier est une question récurrente à laquelle un assureur est confronté. La transposition du GBM pour créer des règles simples d'affectation des véhicules est une tâche délicate. Le recours à un métamodèle est une solution envisageable afin de transformer un modèle complexe en un modèle plus simple tout en gardant une généralisation du phénomène observé. Le recours à des modèles dont l'impact est quantifiable, comme les GLMs ou les arbres de régressions sont des pistes concevables.

Les enjeux sont multiples d'un point de vue opérationnel et commercial :

— Les acteurs en charge de la partie opérationnelle doivent retranscrire les règles élaborées pour affecter au véhicule le score donné par le véhiculier.

— Le commercial doit justifier une différence de prix entre deux véhicules semblables. Des règles d'affectations simples permettent de gagner en interprétation tout en gardant un pouvoir de généralisation assez fort.

Nous pouvons voir que le véhiculier est un sujet riche sur de nombreux points. Celui-ci doit faire la jonction entre deux aspects : **hyper précision** et **simplicité**.

— L'hyper précision découle de la richesse des données et des variables explicatives du véhicule. L'intégration de ces variables multiplie les interactions et donc les règles d'affectation du véhicule.

— La simplicité renvoie à la partie commerciale et opérationnelle.

L'actuaire ne doit pas ignorer ces éléments afin de simplifier le travail des autres équipes avec qui la collaboration est primordiale.

Conclusion

Au cours de ce mémoire, nous avons développé deux méthodologies pour réaliser un véhiculier. La première repose sur une approche non supervisée où nous avons réalisé du *clustering* sur nos véhicules. La seconde s'appuie sur la modélisation de l'effet véhicule à partir du résidu d'une modélisation GLM en excluant les variables véhicules.

Les statistiques réalisées sur les groupes permettent de voir que la méthodologie adoptée dans le cadre non supervisé est pertinente pour isoler des groupes de véhicules homogènes au regard de la fréquence. La nouvelle méthodologie représente une alternative pour construire un premier véhiculier indépendant de la classification adoptée par la SRA, en se reposant essentiellement sur les caractéristiques techniques du véhicule. Elle accepte aussi une pondération arbitraire permettant de prendre en compte tout facteur externe.

L'extraction du risque véhicule à partir du résidu d'un modèle GLM est au cœur de la modélisation supervisée. Cette méthodologie permet de gagner sur deux terrains : la segmentation et l'interprétabilité. Pour le premier, la variable véhiculier possède une amplitude plus forte que la classification SRA, ce qui permet de mieux segmenter le portefeuille. Pour le second, l'intégration de cette nouvelle variable dans le GLM permet d'obtenir une pente linéaire. Enfin, le modèle GBM réalisé pour approcher le résidu issu du GLM est généralisable sur les nouveaux véhicules entrants et sur ceux présentant des valeurs manquantes.

Des pistes de réflexion sont envisageables sur les deux méthodologies. Pour la méthode non supervisée, celle-ci nécessite une maille de travail adaptée au nombre de véhicules pour ne pas être limitée par la complexité de l'algorithme ou le stockage de la matrice de dissimilarité. L'approche est donc envisageable sur un portefeuille deux roues, mais nécessite une adaptation comme dans notre étude, pour un portefeuille automobile. De plus, la concentration des individus sur la carte des véhicules reflète que les caractéristiques techniques des véhicules sont très proches. Enfin, la méthodologie supervisée s'appuie sur l'hypothèse fondamentale que le résidu contient de l'aléa et du signal véhicule. Or, le signal capté n'est pas composé seulement des variables explicatives du véhicule. Ainsi, l'étape de lissage est primordiale pour pouvoir capter précisément cette information.

Le véhiculier est soumis à de nombreuses contraintes opérationnelles. Le retraitement des données est nécessaire pour profiter pleinement de l'efficacité des algorithmes de Machine Learning. Le recours à la complétion de la base SRA est envisageable, mais nécessite une identification précise du véhicule, par le biais de sa plaque d'immatriculation ou de la version du véhicule. Pour résoudre ce problème, des modèles d'imputation des données manquantes adaptés à notre contexte de données mixtes existent, ce dernier point étant au coeur des sujets de recherche en statistique.

Liste des symboles

La liste suivante récapitule l'ensemble des abréviations utilisées dans ce mémoire.

AFDM	Analyse Factorielle de Données Mixtes (FAMD : Factorial Analysis of Mixed Data)
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAH	Classification Ascendante Hiérarchique
CART	Classification and Regression Trees (Arbres de classification et de régression)
CCFA	Comité des Constructeurs Français d'Automobiles
CLARA	CLustering for Large Applications
CRM	Coefficient de Réduction Majoration (ou Bonus Malus)
DTA	Dommage Tous Accidents (se réfère à la garantie dommage)
FFA	Fédération Française de l'Assurance
GBM	Gradient Tree Boosting
GLM	Modèle Linéaire Généralisé (GLM : General Linear Model)
LOM	Loi d'orientation des mobilités
PAM	Partitioning Around Medoids
RMSE	Root Mean Square Error (Racine carré des erreurs quadratiques moyen)
SOM	Self organizing maps (Carte de Kohonen)
SRA	Sécurité et Répération Automobile
t-SNE	t-distributed Stochastic Neighbor Embedding

Bibliographie

[1] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1) :32–46, 2001.

[2] Laura Bendhaiba. Création d'un package r pour des cartes auto-organisatrices. 2013.

[3] Arthur Charpentier. *Computational actuarial science with R*. CRC press, 2014.

[4] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[5] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[6] Yves Gratton. Le krigeage : la méthode optimale d'interpolation spatiale. *Les articles de l'Institut d'Analyse Géographique*, 1(4), 2002.

[7] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning : data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2) :83–85, 2005.

[8] Felix Hébert. Elaboration d'un zonier en assurance automobile à l'aide de données externes et d'algorithmes de data-science. *Mémoire ISUP*, 2016.

[9] José Hernández-Torruco, Juana Canul-Reich, Juan Frausto-Solís, and Juan José Méndez-Castillo. Feature selection for better identification of subtypes of guillain-barré syndrome. *Computational and mathematical methods in medicine*, 2014, 2014.

[10] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[11] Teuvo Kohonen. Exploration of very large databases by self-organizing maps. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 1, pages PL1–PL6. IEEE, 1997.

[12] Julie Lavenu. Les méthodes de machine learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ? *Mémoire ISFA*, 2016.

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov) :2579–2605, 2008.

[14] Xavier Milhaud. Notes de cours : Pratiques avancées de la tarification. *ISFA*, 2019.

[15] Jérôme Pagès. Analyse factorielle de donnees mixtes : principe et exemple d'application. *Montpellier SupAgro, http ://www. agro-montpellier. fr/sfds/CD/textes/pages1. pdf*, 2004.

[16] Gillian Z. Heller Piet de Jong. *Generalized Linear Models for Insurance Data*. International series on actuarial science. Cambridge University Press, 2008.

[17] Peter J Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65, 1987.

[18] Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5) :1308–1323, 2016.

[19] Rasa Sipulskyte. Development of a motor vehicle classification scheme for a new zealand based insurance company. *New Zealand Society of Actuaries Conference 2012*, 2012.

[20] Roman Timofeev. Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, 2004.

Table des figures

Annexe

Maille véhicule

Création de la maille véhicule

Variable	Maille 1 Exploration	Maille 2	Maille 3	Maille 4
Marque	x	x	x	X
Modele	x	x	x	X
Version	x	x	x	X
Type Mines	x		x	
Groupe SRA	x			
Classe SRA	x			
Classe Reparation	x			
Alimentation	x	x	x	X
Cylindrée	x	x	x	X
Energie	x	x	x	X
Carrosserie	x	x	x	X
Vitesse max	x	x	x	X
Puissance (cv)	x	x	x	X
Puissance fiscale	x			
Puissance (kw)	x			
Nb places	x	x	x	X
Type véhicule	x	x	x	X
Nombre de cylindres	x	x	x	X
Longueur	x	x	x	X
Largeur	x	x	x	X
Hauteur	x	x	x	X
Empattement	x	x	x	X
Poids a vide	x	x	x	X
PTAC	x	x	x	X
Dernier prix euros	x	x	x	X
Prix bloc optique	x	x	x	X
Prix pare brise	x	x	x	X
Prix remplacement piece	x	x	x	X
Ratio Poids/Puissance	x	x	x	X
Année mise en circulation	x	x		
Nombre de véhicules	189 518	141 765	122 319	34 429
dont EXPO_DTA > 100	676	705	1295	1503

FIGURE 11.5 – Différentes mailles de véhicules et nombre de véhicules associés

L'année de mise en circulation a créée beaucoup de véhicules et rend la base véhicule moins fiable par rapport à l'exposition. Nous avons décidé de mettre cette caractéristique de coté et d'inclure l'âge du véhicule directement dans le modèle GLM, pour capter un effet technique du véhicule. De plus, ce facteur ressortant très bien dans un GLM (aspect linéaire en lien avec la sinistralité), il serait dommage de perdre cet effet bénéfique dans un GLM.

Analyse des résultats pour le *clustering* non-supervisé

Méthode 1 : PAM équi-pondérée

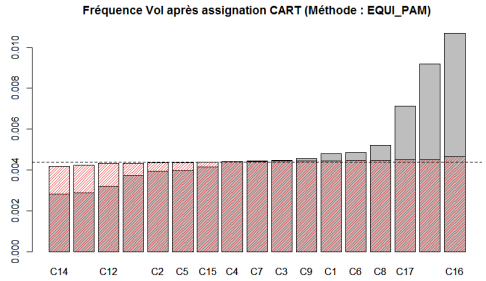

FIGURE 11.6 – Niveau de fréquence vol moyen pour chaque cluster (PAM equipondéré)

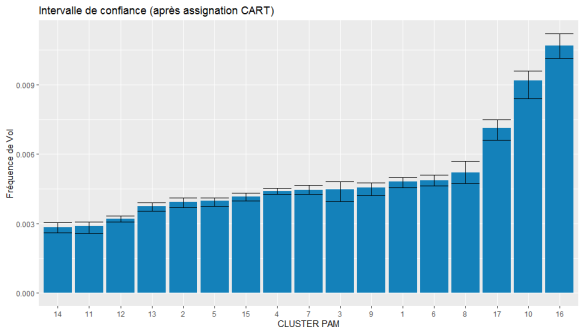

FIGURE 11.7 – Analyse des intervalles de confiance pour la fréquence vol de chaque cluster (PAM equipondéré)

Méthode 2 : PAM pondérée

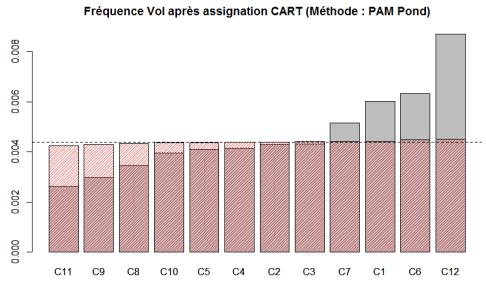

FIGURE 11.8 – Niveau de fréquence vol moyen pour chaque cluster

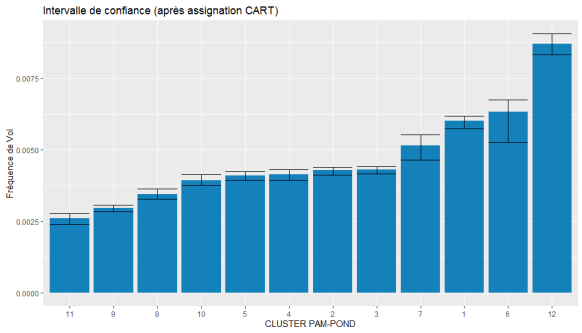

FIGURE 11.9 – Analyse des intervalles de confiance pour la fréquence vol de chaque cluster

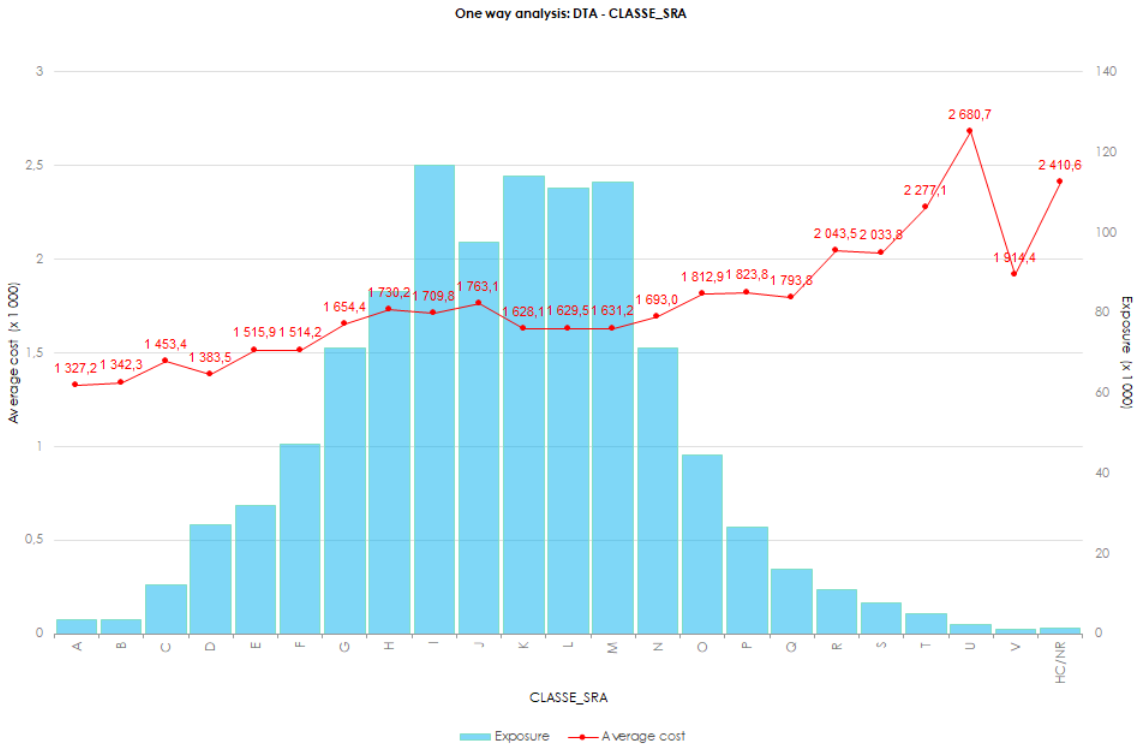

FIGURE 11.10 – Coût-moyen observé pour la garantie dommage en fonction de la classe SRA

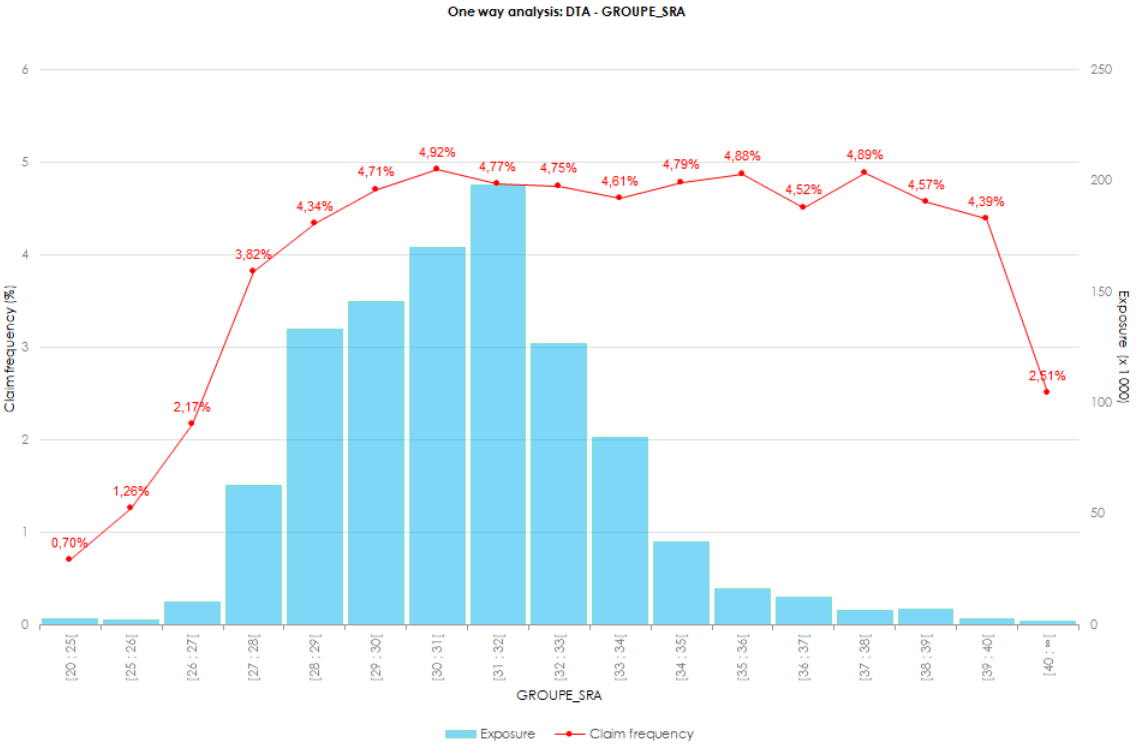

FIGURE 11.11 – Fréquence moyenne observée pour la garantie dommage en fonction du groupe SRA

Comparaison des bétas pour le véhiculier SRA et le véhiculier modélisé pour le coût-moyen dommage

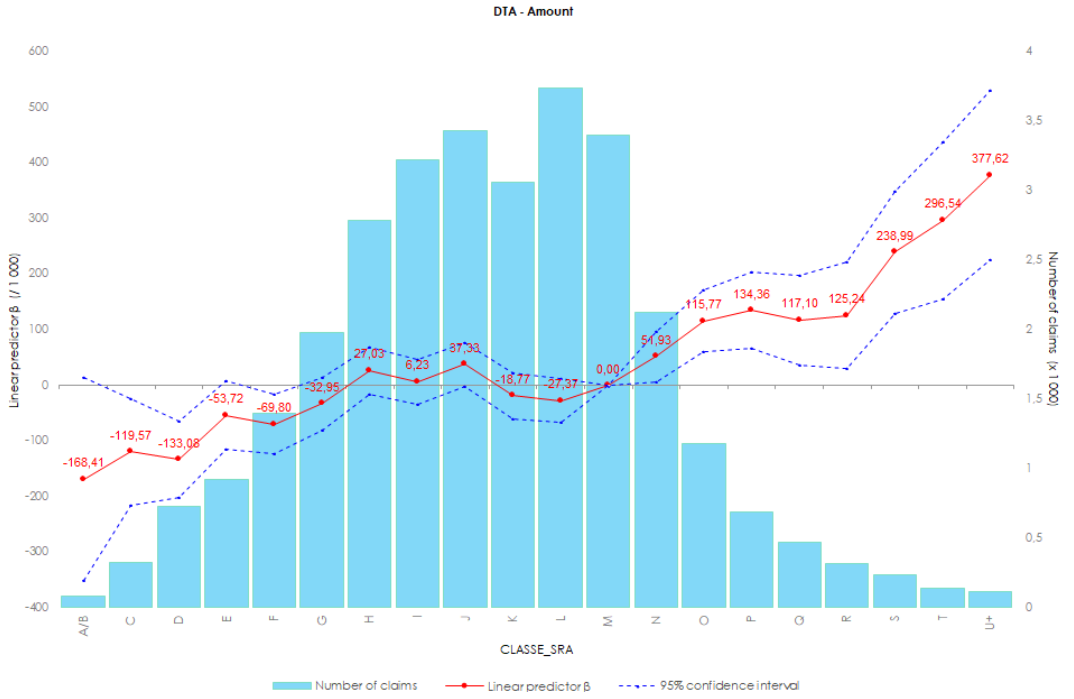

FIGURE 11.12 – Coefficients relatifs à la modélisation du coût-moyen pour la classe SRA

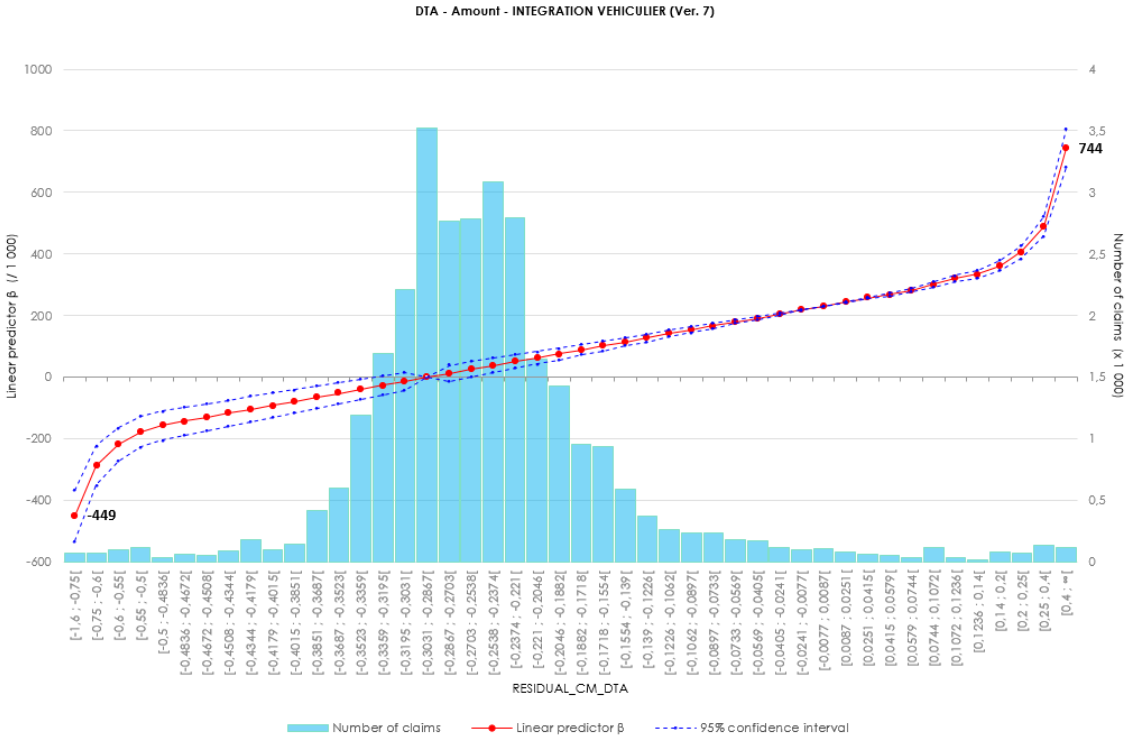

FIGURE 11.13 – Coefficients relatifs à la modélisation du coût-moyen pour notre résidu modélisé