Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires

Par : OLIVIA SECK
Titre : Construction d'un véhiculier à l'aide de méthodes de machine learning et mesure de l'impact tarifaire

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : AVIVA ASSURANCES
Nom : ANNE-SOPHIE VERSCHAVE

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : NOEMIE PEY

Signature :

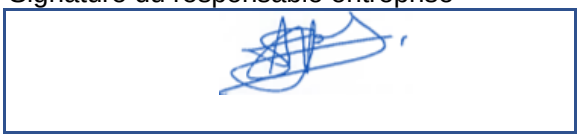
Invité :

Nom : _____

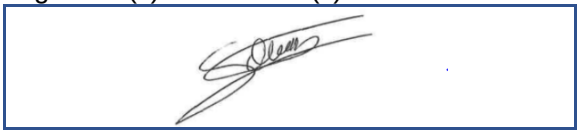
Signature :

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

Signature du responsable entreprise



Signature(s) du candidat(s)



Résumé

Afin de se démarquer dans un marché concurrentiel, les assureurs sont constamment à la recherche de méthodes innovantes leur permettant d'améliorer la compréhension et la segmentation de leurs risques. Avec l'arrivée des véhicules autonomes, l'importance de « l'aspect véhicule » va probablement croître dans la tarification au détriment de « l'aspect conducteur ». Se munir d'un véhiculier revient ainsi à disposer d'un levier de différenciation vis-à-vis de la concurrence et d'affiner le suivi du risque sous-jacent. Le véhiculier peut permettre de prendre en compte plus d'informations sur les véhicules que ce qui serait considéré dans un simple Modèle Linéaire Généralisé. La problématique sera de mesurer l'apport des données véhicules dans un modèle de tarification automobile par la création d'un véhiculier pour la garantie dommage à l'aide des méthodes de *machine learning*. La méthode de tarification de la garantie dommage décomposant la prime pure en modèle de sévérité / modèle de fréquence, des véhiculiers seront distinctivement réalisés sur ces deux composantes. Plusieurs méthodes de *machine learning*, comme le *Random Forest* et le *XGBoost*, seront mises en œuvre sur les résidus des Modèles Linéaires Généralisés hors variables véhicules. Afin d'appréhender les similarités entre chaque véhicule, une cartographie sera créée à l'aide des méthodes de représentations spatiales permettant ainsi l'identification des voisins par la triangulation de Delaunay. Une classification du score obtenu sera réalisée afin de créer les véhiculiers. Enfin, une fois introduits dans les modèles tarifaires, l'apport et l'impact des véhiculiers sur la tarification seront étudiés.

Mots clés : Tarification, Automobile, IARD, GLM, Véhiculier, *Random Forest*, *XGBoost*, ACP, AFDM, Triangulation de Delaunay, Classification, CAH.

Abstract

To stand out in a competitive market, insurers are constantly looking for innovative ways, which allow them to improve their understanding and classification of their risks. With the advent of autonomous vehicles, the importance of the « vehicle aspect » will probably increase at the expense of the « driver aspect ». The use of a set of more refined vehicles rates groups can be an interesting way to stay competitive as it might bring a better classification of the vehicles. Classifying the vehicles into rates group may allow to consider more vehicle information than what would be considered into a simple Generalized Linear Model. This is what this paper propose to study. Several machine learning methods such as Random Forest and XGBoost, will be implemented on the residuals of Generalized Linear Models excluding vehicle variables. In order to understand the similarities between each vehicle, a map will be created using spatial representation methods, allowing the identification of neighbors by Delaunay triangulation. A classification of the obtained score will be carried out, allowing the creation of vehicles classification. Finally, once introduced into the pricing models, the contribution and impact of the vehicles classification on pricing will be studied.

Key Words : Pricing, Automobile, General Insurance, GLM, vehicule classification, Random Forest, XGBoost, PCA, MDFA, Delaunay Triangulation, Classification, CAH.

Note de synthèse

L'automobile étant l'un des marchés les plus concurrentiels, l'amélioration constante de la segmentation du portefeuille est un enjeu majeur pour les assureurs. L'utilisation des méthodes de *machine learning* permet de répondre à ce besoin de sophistication tarifaire. Les modèles tarifaires doivent cependant pouvoir rester lisibles et facilement explicables pour des besoins d'implémentation, de communication et de suivi.

Jusqu'à présent, les assureurs étaient souvent contraints par le nombre limité de variables dans les modèles. En effet, les méthodes classiques de tarification contraignent le nombre de variables explicatives utilisables. Or, les dernières avancées automobiles (systèmes d'aide à la conduite, véhicules connectés etc.) mettent à la disposition des assureurs une multitude d'informations, représentant chacune des opportunités dans un environnement en constante évolution. Aussi, le véhiculier est un bon moyen de faire face à cette nouvelle réalité, puisqu'il permet de prendre en compte un grand nombre de variables en les regroupant dans un score.

L'objectif de ce mémoire est d'étudier la pertinence d'un véhiculier en mesurant l'apport de ce dernier au modèle de tarification dommage en termes de segmentation et prédiction (la garantie dommage est l'une des garanties pesant le plus en assurance automobile en France[1]). Pour ce faire, les modèles intégrant les véhiculiers sont comparés aux modèles initiaux contenant les variables véhicules libres. Ainsi, il est possible de juger de l'intérêt du remplacement des variables véhicules par un véhiculier.

Dans une tarification automobile, les Modèles Linéaires Généralisés (GLM), couramment utilisés, procurent une bonne lisibilité des impacts de chaque critère tarifaire sur la prime. Plusieurs approches de modélisation sont possibles : une modélisation en prime pure ou bien une modélisation séparant celle-ci en deux composantes, la fréquence et la sévérité. La deuxième approche est privilégiée dans ces travaux car l'effet capté par les variables du modèle joue différemment en fréquence et en sévérité. Le véhiculier s'appuie ainsi sur une modélisation de la sinistralité réalisée au préalable, ceux retenus ici étant les modèles en production. Le processus de tarification général est le suivant :

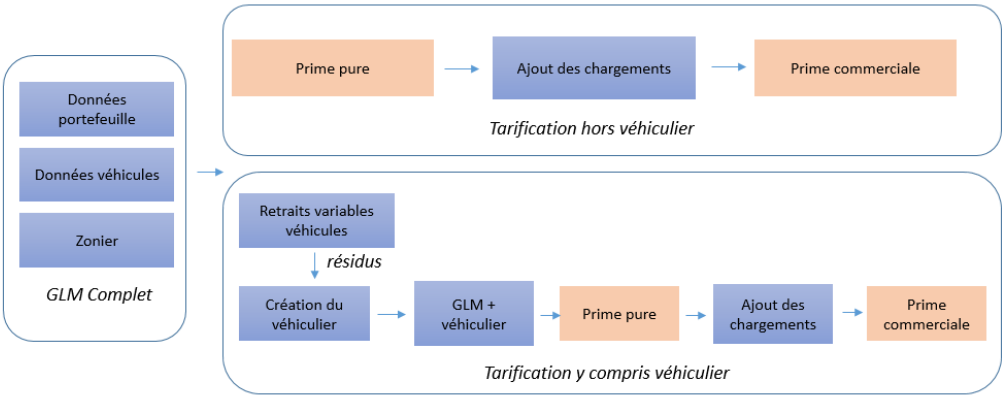


FIGURE 1 – Processus général de tarification

1. FFA (2020), *L'assurance française Données clés 2020*

Les variables tarifaires automobiles se résument en quatre catégories principales :

- les variables associées aux données « conducteur » ;
- les variables associées au risque (franchises, etc.) ;
- les variables associées aux données « géographiques » ;
- les variables associées aux données « véhicules » : les caractéristiques du moteur, les caractéristiques techniques, le genre, les dimensions du véhicule, les options, les aides à la conduite et autres.

La construction du véhiculier repose sur des données véhicules mises à disposition des assureurs par l'organisme Sécurité et Réparations Automobiles (SRA).

Parmi les différentes méthodes de construction de véhiculiers existantes, cette étude retient la méthode dite « résiduelle ». L'hypothèse principale repose sur des résidus s'expliquant par l'effet véhicule d'une part et le bruit d'autre part, les autres impacts étant captés par les autres variables présentes dans les modèles sévérité / fréquence.

Dans un premier temps, les résidus sont extraits des modèles, agrégés au niveau véhicule et expliqués par le biais de deux méthodes de *machine learning*. En parallèle, une cartographie des véhicules est réalisée à l'aide de méthodes multidimensionnelles. Elles permettent de représenter les véhicules dans un plan et d'en récupérer les coordonnées qui serviront de base à la triangulation de Delaunay. Un lissage des résidus prédits est ensuite réalisé afin de rendre ces derniers plus homogènes. Une fois ces étapes abouties, les groupes composant le véhiculier sont issus d'une classification de ces résidus prédits et lissés. Enfin, les résultats des modèles tarifaires intégrant les variables véhicules sont comparés à la combinaison des GLM hors variables véhicules et véhiculiers.

L'élaboration du véhiculier se décompose en plusieurs étapes résumées dans le schéma ci-après.

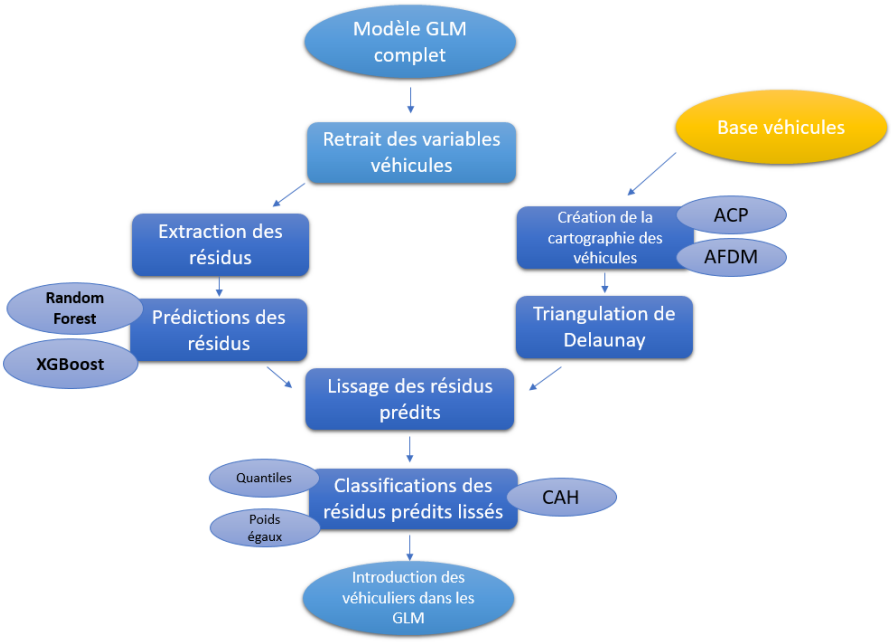


FIGURE 2 – Méthodologie de création du véhiculier

Explication des résidus par les variables véhicules

La première étape consiste à retirer les variables véhicules des modèles tarifaires fréquence et sévérité. Ainsi, l'effet « véhicule » peut être isolé à l'aide des résidus des modèles. Les résidus sont ensuite extraits et calculés au niveau véhicule. Chaque véhicule de la base a donc un résidu associé. Celui-ci est expliqué à l'aide des modèles non paramétriques *Random Forest* et *XGBoost*.

Ces algorithmes s'affranchissant des conditions de non corrélation entre les variables explicatives, une étape de sélection de variables en amont est nécessaire afin de limiter de potentiels effets de corrélation indésirables. La base de données contenant des variables quantitatives et qualitatives, l'analyse des corrélations est réalisée à l'aide du V de Cramer, pour évaluer le niveau de dépendance entre chacune des variables, quelles que soient leurs typologies. Ainsi, les variables les moins corrélées sont sélectionnées puis sont utilisées dans les modèles de *machine learning*.

Une étape d'optimisation des hyper-paramètres par le biais d'une validation croisée est mise en place, permettant ainsi de limiter le sur-apprentissage, et d'améliorer le pouvoir prédictif du *Random Forest* et du *XGBoost*. S'agissant d'une problématique de régression, les modèles ont été évalués à l'aide de l'erreur quadratique moyenne (RMSE).

Afin de limiter le sur-apprentissage, les modèles sont entraînés sur une base d'apprentissage appelée Train représentant 80% des données et une base de test 20%.

Pour les résidus fréquence, les résultats du *Random Forest* sont meilleurs sur la base d'apprentissage (Train) que ceux du *XGBoost*, du fait d'une erreur quadratique moyenne plus faible. Les résultats sur la base de test sont meilleurs sur le modèle *XGBoost*, mais cette amélioration n'est pas significative.

	Modèles	Critère	Train	Test
Fréquence	RF	RMSE	0,5201	0,5353
	XGBoost	RMSE	0,5226	0,5346

TABLE 1 – Évaluation des modèles sur les résidus prédits du modèle fréquence

Tout comme pour la fréquence, les résultats des modèles sur les résidus « sévérité » sont proches. Les résultats du *Random Forest* sont meilleurs que ceux du *XGBoost* sur la base d'apprentissage (Train) et la base de test.

	Modeles	Critère	Train	Test
Sévérité	RF	RMSE	31,0508	31,0107
	XGBoost	RMSE	31,0628	31,0375

TABLE 2 – Évaluation des modèles sur les résidus prédits du modèle de sévérité

Dans le but d'augmenter la compréhension des résultats obtenus, la contribution de chaque variable est calculée afin de mettre en exergue les différences entre les modèles *Random Forest* et *XGBoost*. Pour la fréquence, la variable la plus importante est la valeur du véhicule *(Random*

Forest) et la marque (*XGBoost)*. Pour le modèle de séverité du *Random Forest* et du *XGBoost*, il s'agit respectivement de la vitesse et des coûts de réparation.

Pour la fréquence et la sévérité, les écarts de résultats entre les deux algorithmes ne sont pas significatifs mais restent meilleurs pour le *Random Forest*. De plus, les variables contribuant le plus aux modèles sont plus cohérentes sur le modèle *Random Forest* d'un point de vue métier. Le *Random Forest* est donc privilégié dans les deux cas malgré des résultats moins précis sur la base de test que le *XGBoost* sur le modèle des résidus fréquence.

Cartographie des véhicules, identification des voisins et lissage

Une cartographie des véhicules est réalisée à l'aide de méthodes d'analyse multidimensionnelles. Celles-ci permettent de représenter les véhicules dans un plan afin d'en extraire les coordonnées. Les coordonnées servent ensuite de base à la triangulation de Delaunay. Celle-ci relie les véhicules entre eux en fonction de leur positionnement dans ce même plan et détermine les véhicules considérés comme les plus « proches ».

Deux méthodes sont comparées : l'analyse en composante principale (ACP) sur les variables quantitatives uniquement et l'analyse factorielle en données mixtes (AFDM) utilisable à la fois sur les variables qualitatives et quantitatives. L'ACP montre des résultats plus fiables puisque la part d'inertie expliquée sur les 2 premiers axes de l'ACP se porte à plus de 70% contre seulement 8% pour les 3 premiers axes de l'AFDM. Les coordonnées de l'ACP sont donc retenues pour la triangulation de Delaunay.

La triangulation de Delaunay identifie les véhicules voisins par le biais d'une triangulation d'un ensemble de points, reliant ainsi ces derniers de manière optimale en minimisant les angles de chaque triangle. L'une des propriétés de cette méthode est qu'elle est fermée, c'est-à-dire que tous les points sont reliés, engendrant potentiellement des liaisons considérées comme aberrantes d'un point de vue métier. Une optimisation des liens créés a été mise en œuvre pour réduire cet effet en supprimant les liaisons entre les points considérés comme les plus loin en termes de distance.

	Nombre de voisins minimum	Nombre de voisins moyen	Nombre de voisins maximum
Avant suppression des liens aberrants	2	15	29
Après suppression des liens aberrants	0	12	22

Table 3 – Nombre de voisins

Une table d'adjacence des véhicules est alors construite, permettant d'associer des voisins à chaque véhicule. Elle est utilisée pour réaliser un lissage sur les résidus précédemment prédits par le *Random Forest*.

Ce lissage consiste à estimer un résidu prédit lissé en pondérant les résultats obtenus par les modèles non paramétriques retenus (le *Random Forest*) et le résidu moyen de chaque voisin identifié par la triangulation de Delaunay. Le lissage uniformise ainsi les risques entre des véhicules considérés comme voisins.

Classification et mesure de l'impact tarifaire des véhiculiers

Une fois les résidus prédits et lissés, ils sont regroupés en différentes classes. Afin de construire les groupes optimaux de véhicules, plusieurs méthodes de classification non supervisées sont utilisées et comparées : les quantiles, les poids égaux et la classification ascendante hiérarchique (CAH). La valeur du véhicule et le poids sont les variables les plus discriminantes. Ces deux variables contenant respectivement 27 et 26 modalités, il a donc été décidé de créer au minimum 30 classes pour les véhiculiers. Les résultats de la CAH ne permettent pas d'atteindre ce nombre de classes minimum, aussi ces dernières ne seront pas reprises.

Il n'y a pas de méthode pour connaître le nombre de classes optimales pour la méthode des quantiles et la méthode des poids égaux. Des véhiculiers « quantiles » et « poids égaux » contenant un nombre de classes différents sont donc introduits un à un dans les modèles afin de tester leur pouvoir prédictif.

Les véhiculiers testés dans les modèles de fréquence et de sévérité sont les suivants :

Listes des modèles	Description
Modèle initial	Modèle référence contenant les variables véhicules
Modèle + véhiculier "Poids égaux" 50	Modèle sans les variables véhicule et avec le véhiculier de 50 classes créé avec la méthode des poids égaux
Modèle + véhiculier "Poids égaux" 40	Modèle sans les variables véhicule et avec le véhiculier de 40 classes créé avec la méthode des poids égaux
Modèle + véhiculier "Quantile" 30	Modèle sans les variables véhicule et avec le véhiculier de 30 classes créé avec la méthode des quantiles
Modèle + véhiculier "Quantile" 50	Modèle sans les variables véhicule et avec le véhiculier de 50 classes créé avec la méthode des quantiles
Modèle + véhiculier "Quantile" 80	Modèle sans les variables véhicule et avec le véhiculier de 80 classes créé avec la méthode des quantiles

TABLE 4 – Listing des modèles

Les indicateurs utilisés pour évaluer l'apport de ces véhiculiers dans cette étude sont le RMSE et le coefficient de Gini. Les résultats des véhiculiers dans les modèles sont les suivants :

Modèle	Indicateurs	Modèle initial avec variables véhicules	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Fréquence – Base d'apprentissage	RMSE	0,14062	0,14066	0,14068	0,14067	0,14066	0,14066
	Gini	0,24360	0,23440	0,23380	0,23370	0,23790	0,23480

Modèle	Indicateurs	Modele initial avec variables véhicules	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Fréquence – Base Test	RMSE	0,14055	0,14057	0,14058	0,14057	0,14058	0,14057
	Gini	0,24680	0,23780	0,23700	0,23650	0,24040	0,23790

TABLE 5 – Évaluation des modèles de fréquence

L'introduction des véhiculiers dans le modèle de tarification de la fréquence dommage ne contribue pas à améliorer les résultats. Le RMSE est plus élevé pour tous les modèles avec véhiculier, ce qui indique que la qualité d'ajustement des modèles est meilleure sur le modèle initial que sur les modèles contenant les véhiculiers. Les résultats des véhiculiers sont très proches et aucune méthode de classification ne se dégage.

Le coefficient de Gini mesure le niveau de segmentation du modèle. Celui-ci permet de confirmer que le modèle initial est également meilleur en termes de segmentation puisque le coefficient de Gini est plus élevé sur le modèle initial que sur les modèles contenant le véhiculier.

Ces résultats sont confirmés sur la base de test.

Les résultats sur le modèle de sévérité sont présentés ci-après :

Modèle	Indicateurs	Modèle initial	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Sévérité - Apprentissage	RMSE	264,484	270,803	269,768	265,832	264,281	264,35
	Gini	0,148	0,1498	0,1495	0,1427	0,1482	0,1483

Modèle	Indicateurs	Modèle initial	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Sévérité - Test	RMSE	261,569	264,512	263,869	262,393	260,536	260,556
	Gini	0,1565	0,156	0,1552	0,1493	0,1535	0,153

TABLE 6 – Évaluation des modèles de sévérité

Contrairement aux modèles fréquence, une méthodologie se détache, puisque les modèles contenant les véhiculiers construits à partir de la méthode des quantiles ont une qualité d'ajustement supérieure à ceux construits par la méthode des poids égaux. De plus, pour les véhiculiers « quantiles » avec 50 et 80 classes, les RMSE sont plus faibles que ceux du modèle initial. Le véhiculier contenant 30 classes a un RMSE plus élevé que le modèle initial mais reste meilleur que les véhiculiers construits avec la méthode des poids égaux. Ces résultats sont confirmés sur la base de test.

Le coefficient de Gini est meilleur sur tous les modèles contenant les véhiculiers, excepté pour le modèle à 30 classes et construit à l'aide des quantiles. Cependant, ces résultats ne sont pas confirmés sur la base de test.

En conclusion, les véhiculiers semblent pertinents sur les modèles de sévérité puisque la qualité d'ajustement des modèles est meilleure. L'effet véhicule semble donc prendre une part plus importante dans le modèle de sévérité que sur le modèle fréquence. Enfin le véhiculier permet d'améliorer en partie la segmentation puisque le coefficient de Gini s'améliore dans les modèles contenant les véhiculiers. Cependant, les résultats restent mitigés puisque ces résultats ne sont pas confirmés sur la base de test.

Pistes d'améliorations et conclusion

Malgré une légère amélioration sur le modèle de sévérité, les résultats des différents véhiculiers ne permettent pas de remplacer les variables véhicules des modèles initiaux. Cependant, cette étude met en exergue une amélioration potentielle des modèles GLM en place, mais des étapes d'optimisation sont nécessaires afin d'améliorer l'impact des véhiculiers dans les modèles tarifaires. Pour le modèle fréquence, d'autres méthodes peuvent être testées. En effet il existe une multitude d'algorithmes de *machine learning* et de méthodes de classification.

De plus, la triangulation de Delaunay est très dépendante de l'ensemble de points utilisé. Aussi, une optimisation des résultats de l'ACP et de l'AFDM par une meilleure gestion des données manquantes ou par l'ajout de nouvelles données, peuvent conduire à améliorer les résultats de la triangulation.

La période d'observation (2013 - 2017) peut aussi expliquer ces résultats mitigés. En effet, l'une des contraintes de cette étude fut de devoir utiliser les données du portefeuille 2013 à 2017. La sophistication des véhicules croissant avec le temps, ce sont les voitures récentes qui bénéficient des nouveautés en termes de sécurité. Celles-ci étant moins présentes sur les vieux véhicules, les données disponibles concernant ces options sont peu exploitables, du fait d'un grand nombre de valeurs manquantes. La mise à jour de cette étude sur des années plus récentes permettrait de tenir compte des derniers modèles de véhicules et donc de pouvoir augmenter le nombre de variables dans les modèles *Random Forest* et *XGBoost* et donc d'améliorer à termes la classification.

Enfin, la maintenance du véhiculier limite son utilisation. En effet, chaque semaine, de nouveaux véhicules sont disponibles sur le marché, obligeant les assureurs à une mise à jour régulière pouvant s'avérer contraignante et chronophage.

Synthesis Note

As the automobile is one of the most competitive markets, the constant improvement of portfolio segmentation is a major challenge for insurers. The use of machine learning methods can meet this need for pricing sophistication. However, pricing models must remain readable and easily explainable for implementation, communication and monitoring purposes.

Until now, insurers were often constrained by the limited number of variables in the models. Indeed, traditional pricing methods limit the number of explanatory variables that can be used. However, the latest automotive advances (driving assistance systems, connected vehicles, etc.) provide insurers with a multitude of information, each representing opportunities in a constantly changing environment. Also, the vehicle is a good way to deal with this new reality, since it allows you to take into account a large number of variables by grouping them into a score.

The objective of this paper is to verify the relevance of a vehicle classification by measuring its contribution to the damage pricing model in terms of segmentation and prediction (damage guarantee is one of the guarantees weighing the most in automobile insurance in France[1] . To do this, the models including the vehicles classification are compared to the initial models containing the variables vehicles. It is possible to judge the interest of replacing the vehicle variables by a vehicle classification . In automobile pricing, Generalized Linear Models (GLM), commonly used provide a good understanding of the impact of each pricing criterion on the premium. Several modeling approaches are possible : a pure premium model or a model separating into two components, frequency and severity. The second approach is favored in this works because the effect captured by the model variables is different in frequency and severity. The vehicle classification is based on a prior claims model. Those selected here being the models in production. The general pricing process is as follow :

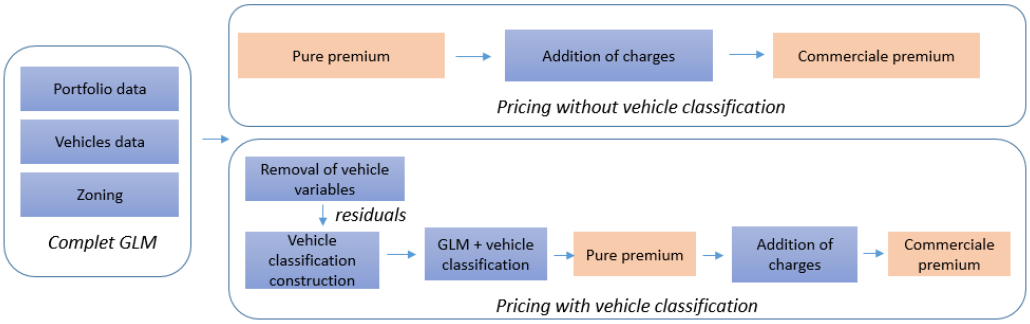


FIGURE 3 – Pricing process

The automobile fare variables can be summarized into four main categories :

- variables associated with « driver » data ;
- variables associated with risk variables (deductibles, etc.) ;
- variables associated with « geographic » data ;
- variables associated with « vehicle » data : engine characteristics, technical characteristics, type and vehicle dimensions, options, driving aids and others.

1. FFA (2020), French insurance Key data 2020

The construction of the vehicle classification is based on vehicle data provide to insurers by the Automobile Safety and Repairs (SRA) organization.

Among the various existing vehicle classification construction methods, this study retains the so-called « residual » method. The main hypothesis is that the residuals are explained by the vehicle effect on the one hand and noise on the other, the other impacts being captured by the other variables in the severity/frequency models.

First, the residuals are extracted from the models, aggregated at the vehicle level and explained through two machine learning methods. In parallel, a mapping of vehicles is performed using multidimensional methods. They allow to represent the vehicles in a plan, and to recover the coordinates which will used as a basis for the triangulation Delaunay. A smoothing of the predicted residues is then made in order to make them more homogeneous. Once these steps have been completed, the groups of the vehicle classification are derived from a classification of these predicted and smoothed residues.

Finally, the results of the pricing models with the vehicle variables are compared to GLMs without the vehicle variables but with the vehicles classification.

The development of the vehicle classification is summarized in the diagram below.

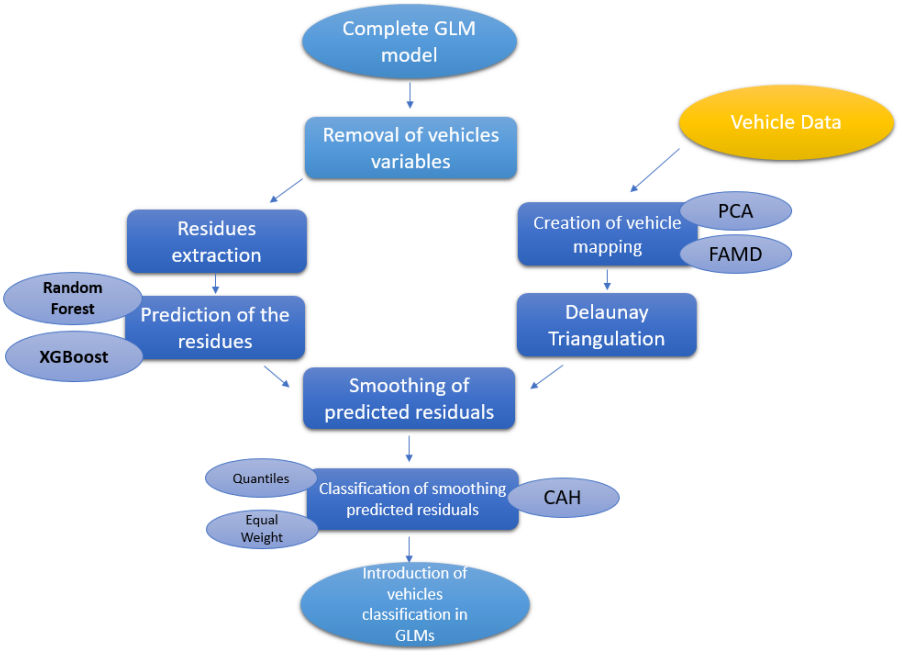


FIGURE 4 – Vehicle Classification pattern

Explanation of residues by vehicle variables

The first step is to remove the vehicle variables from the frequency models and severity. The vehicle effect can be isolated using the residuals of the models. The residues are then extracted and calculated at the vehicle level. Each vehicle in the base is associated to a residue. This residues is explained using the non-parametric models Random Forest and XGBoost.

These algorithms are free from the conditions of non-correlation between the explanatory variables an upstream variable selection step is necessary in order to limit potential undesirable

correlation effects. The database containing quantitatives and qualitatives variables the analysis of correlations is performed using Cramer's V, to evaluate the level of dependence between each of the variables, regardless of their typologies. Thus, the least correlated variables are selected and used in machine learning models.

A step of optimization of the hyper-parameters by means of a cross-validation is implemented to limit over-fitting, and to improve the predictive power of the Random Forest and XGBoost. As this is a regression problem, the models have been evaluated using the root mean square error (RMSE).

The models are trained on a training dataset called Train representing 80% of the data and a test dataset 20%.

For the frequency residues, the Random Forest results are better on the training dataset than those of the XGBoost due to a lower mean square error. The results on the test dataset are better on the XGBoost model, but this improvement is not significant.

	Models	Indicator	Train	Test
Frequency	RF	RMSE	0,5201	0,5353
	XGBoost	RMSE	0,5226	0,5346

Table 7 – Predicted residuals models evaluation's of the frequency model

As with frequency, the results of the models on the severity residuals are close. The results of the Random Forest are better than those of the XGBoost on both train and test set.

	Models	Indicator	Train	Test
Severity	RF	RMSE	31,0508	31,0107
	XGBoost	RMSE	31,0628	31,0375

Table 8 – Predicted residuals models evaluation's of the cost model

In order to increase the understanding of the results obtained, the contribution of each variable is calculated in order to highlight the differences between the Random Forest and XGBoost models. For the frequency, the most important variable is the value of the vehicle (Random Forest) and the brand (XGBoost). For the severity model of Random Forest and XGBoost, it is respectively the speed and the repair costs.

For frequency and severity, the differences in results between the two algorithms are not significant but remain better for Random Forest. Moreover, the variables that contribute the most to the models are more consistent with the Random Forest model from a business point of view. The Random Forest is therefore preferred in both cases despite less accurate results on the test basis than the XGBoost on the frequency residuals model.

Vehicle mapping, neighbors identification and smoothing

A mapping of the vehicles is done using multidimensional analysis methods. These methods allow the vehicles to be represented in a plane in order to extract their coordinates. The coordinates are then used as a basis for the Delaunay triangulation. This triangulation links the vehicles together according to their position in the same plane and determines which vehicles are considered to be the « closest ».

Two methods are compared : Principal Component Analysis (PCA) on quantitative variables only and Mixed Data Factor Analysis (MDFA) on both qualitative and quantitative variables. The PCA shows more reliable results, the share of inertia explained on the first two axes of the PCA is more than 70% against only 8% for the first three axes of the MDFA. The PCA coordinates will be retained for the Delaunay triangulation.

Delaunay triangulation identifies neighboring vehicles through a triangulation of a set of points, connecting them in an optimal way by minimizing the angles of each triangle. One of the properties of this method is that it is closed, i.e. all points are connected, potentially generating links considered as outliers from a business point of view. An optimization of the created links has been implemented to reduce this effect by removing the links between the points considered as the farthest in terms of distance.

	Minimum number of neighbors	Average number of neighbors	Maximum number of neighbors
Before removing outliers	2	15	29
After removing outliers	0	12	22

Table 9 – number of neighbors

A vehicle adjacency table is then built, allowing to associate neighbors to each vehicle. It is used to perform a smoothing on the residuals previously predicted by the Random Forest.

This smoothing consists in estimating a smoothed predicted residual by weighting the results obtained by the non-parametric models used (the Random Forest) and the average residual of each neighbor identified by the Delaunay triangulation. Smoothing standardizes the risks between vehicles considered as neighbors.

Classification and measurement of the pricing impact of vehicles classification

Once the residuals are predicted and smoothed, they are grouped into different classes. In order to construct the optimal groups of vehicles, several unsupervised classification methods are used and compared : quantiles, equal weights and hierarchical ascending classification (HAC). Vehicle value and weight are the most discriminating variables. Since these two variables contain 27 and 26 modalities respectively, it was decided to create a minimum of 30 classes for the vehicle owners. The results of the AHC do not allow this minimum number of classes to be reached, so these classes will not be included.

« Quantile » and « equal weight » vehicles classification are introduced into the models to test their predictive power. The metrics used to evaluate the contribution of these vehicles in this study are the RMSE and the Gini coefficient.

The class of vehicles tested in the frequency and severity models are :

List of models	Description
Initial model	Reference model containing vehicle variables
GLM model + vehicles class. « Equal Weight" 50	Model without the vehicle variables and with the 50 class of vehicle created with the equal weight method
GLM model + vehicles class. « Equal Weight" 40	Model without the vehicle variables and with the 40 class of vehicle created with the equal weight method
GLM model + vehicles class. « Quantiles » 30	Model without the vehicle variables and with the 30 class of vehicle created with the quantile method
GLM model + vehicles class. « Quantiles » 50	Model without the vehicle variables and with the 50 class of vehicle created with the quantile method
GLM model + vehicles class. « Quantiles » 80	Model without the vehicle variables and with the 80 class of vehicle created with the quantile method

TABLE 10 – List of models

The results of the vehicles in the models are as follows :

Model	Indicators	Initial model	GLM model + vehicles class. « Equal Weight" 50	GLM model + vehicles class. « Equal Weight" 40	GLM model + vehicles class. « Quantiles » 30	GLM model + vehicles class. « Quantiles » 50	GLM model + vehicles class. « Quantiles » 80
Frequency - Train	RMSE	0,14062	0,14066	0,14068	0,14067	0,14066	0,14066
	Gini	0,24360	0,23440	0,23380	0,23370	0,23790	0,23480

Model	Indicators	Initial model	GLM model + vehicles class. « Equal Weight" 50	GLM model + vehicles class. « Equal Weight" 40	GLM model + vehicles class. « Quantiles » 30	GLM model + vehicles class. « Quantiles » 50	GLM model + vehicles class. « Quantiles » 80
Frequency - Test	RMSE	0,14055	0,14057	0,14058	0,14057	0,14058	0,14057
	Gini	0,24680	0,23780	0,23700	0,23650	0,24040	0,23790

TABLE 11 – Vehicle classification evaluation for the damage frequency model

The introduction of the vehicles classification in the damage frequency pricing model does not improve the results. The RMSE is higher for all models with vehicles classification, indicating that the goodness of fit of the models is better on the initial model than on the models with vehicles classification. The results for the vehicle classification are very close and no single classification method really stands out.

The Gini coefficient measures the level of segmentation of the model. This confirms that the initial model is also better in terms of segmentation since the Gini coefficient is higher on the initial model than on the models containing the vehicle classification. These results are confirmed on the test dataset.

The results on the severity model are presented below :

Model	Indicators	Initial model	GLM model + vehicles class. « Equal Weight" 50	GLM model + vehicles class. « Equal Weight" 40	GLM model + vehicles class. « Quantiles » 30	GLM model + vehicles class. « Quantiles » 50	GLM model + vehicles class. « Quantiles » 80
Severity - Train	RMSE	264,484	270,803	269,768	265,832	264,281	264,35
	Gini	0,1480	0,1498	0,1495	0,1427	0,1482	0,1483

Model	Indicators	Initial model	GLM model + vehicles class. « Equal Weight" 50	GLM model + vehicles class. « Equal Weight" 40	GLM model + vehicles class. « Quantiles » 30	GLM model + vehicles class. « Quantiles » 50	GLM model + vehicles class. « Quantiles » 80
Severity - Test	RMSE	261,569	264,512	263,869	262,393	260,536	260,556
	Gini	0,1565	0,156	0,1552	0,1493	0,1535	0,153

TABLE 12 – Vehicle classification evaluation for the damage cost model

In contrast to the frequency models, one methodology stands out, since the models containing the vehicles classification constructed using the quantile method have a better goodness of fit than those constructed using the equal weight method. In addition, for the « quantile » vehicles classification with 50 and 80 groups, the RMSEs are lower than those of the original model. The vehicle classification with 30 classes has a higher RMSE than the initial model but is still better than the vehicles classification constructed with the equal weights method. These results are confirmed on the test dataset.

The Gini coefficient is better on all models containing the vehicles classification, except for the 30-class model and the model constructed using quantiles. However, these results are not confirmed on the test dataset.

In conclusion, the vehicles classifications seem to be relevant on the severity models since the goodness of fit of the models is better. The vehicle effect seems to be more important in the severity model than in the frequency model. Finally, the vehicle classification effect allows for a partial improvement of the segmentation since the Gini coefficient improves in the models containing vehicles. However, the results remain mixed since these results are not confirmed on the test dataset.

Improvement and conclusion

Despite a slight improvement on the severity model, the results of the different vehicles classifications do not allow to replace the vehicle variables of the initial models. However, this study highlights a potential improvement of the existing GLM models, but optimization steps are needed to improve the impact of the vehicles classification in the pricing models. For the frequency model, other methods can be tested. Indeed, there is a multitude of machine learning algorithms and classification methods.

Moreover, the Delaunay triangulation is very dependent on the set of points used. Therefore, optimizing the results of the PCA and MDFA by better management of missing data or by adding

new data can lead to improved triangulation results.

The observation period (2013 - 2017) and the high age of the vehicle in the portfolio may also explain these mixed results. Indeed, one of the constraints of this study was the need to use data from the 2013 to 2017 portfolio. As vehicles become more sophisticated over time, newer cars benefit from new safety features. As these are less present on older vehicles, the available data on these options is not very usable, due to a large number of missing values. Updating this study to more recent years would make it possible to take into account the latest vehicle models and thus to increase the number of variables in the Random Forest and XGBoost models and thus improve the classification in the long run.

Finally, the maintenance of the vehicle limits its use. Indeed, every week, new vehicles are available on the market, forcing insurers to update the system regularly, which can be very time-consuming and restrictive.

Remerciements

Je tiens tout d'abord à remercier AVIVA ASSURANCES, Anne-Sophie VERSCHAVE, et Solange HAMEL grâce à qui j'ai pu suivre la formation du Certificat d'Expertise Actuarielle dispensée par l'Institut du Risk Management.

Je remercie ma tutrice Noémie PEY pour son expertise sur le produit et sur la tarification automobile, ses précieux conseils et sa relecture attentive.

Je tiens également à remercier Baptiste ANDRIEU, Corinne LAFARGE ainsi qu'Audrey LANTUEJOUL pour leur soutien permanent tout au long de mon cursus et lors de la rédaction de ce mémoire.

Enfin, je tiens à remercier mes proches et mon entourage qui ont su m'apporter leur confiance et leur soutien.

Table des matières

Introduction

Le secteur de l'assurance automobile est un marché ultra concurrentiel obligeant les assureurs à se renouveler régulièrement. L'utilisation de méthodes innovantes permet aux assureurs de répondre à ce besoin constant d'amélioration de leurs modèles tarifaires.

La sophistication des véhicules et l'arrivée de nouvelles options et aides à la conduite augmentent l'impact tarifaire de l'effet véhicule dans les modèles de tarification. Un véhiculier est donc un potentiel moyen d'affiner la segmentation des véhicules puisqu'il permet de prendre en compte dans la tarification plus de caractéristiques véhicules, en les regroupant dans un score.

La problématique sera de mesurer l'apport des données véhicules dans un modèle de tarification automobile par la création d'un véhiculier pour la garantie dommage à l'aide des méthodes de *machine learning*.

Ainsi, une première partie posera le contexte et les objectifs de l'étude, en présentant tout d'abord la place de l'assurance automobile en France et chez Aviva France. Ensuite, des notions essentielles telles que l'utilité de la segmentation en tarification automobile ainsi que la notion de véhiculier et la place qu'il occupe dans le modèle de tarification seront explicitées. Cette partie rappellera enfin les éléments nécessaires à la compréhension des modèles tarifaires, ainsi que des généralités sur les Modèles Linéaires Généralisés.

Dans un deuxième temps, les données utilisées et les différents retraitements réalisés sur celles-ci seront détaillées. Ensuite, les différentes étapes de création du véhiculier seront explicitées, à savoir : la prédiction des résidus à l'aide des modèles de *machine learning* et la méthodologie de lissage des résidus prédits.

Enfin, la dernière partie sera consacrée à la classification de l'effet véhicule prédit lissé à l'aide de plusieurs méthodes de classification. Puis les véhiculiers seront introduits dans les modèles initiaux afin de permettre :

- une évaluation complète de chacun d'entre eux ;
- de mesurer l'apport de l'effet véhicule et l'apport de chaque méthode, notamment l'apport des méthodes de *machine learning* dans la création d'un véhiculier ;
- une comparaison entre les résultats des modèles avec et sans véhiculier.

Chapitre 1

Contexte, objectifs et fondements théoriques

Cette première partie décrira en quelques chiffres la place de l'assurance automobile dans le marché assurantiel français. Le contexte et les objectifs de ce mémoire seront explicités et certaines notions essentielles pour la suite de l'étude seront introduites. Cette partie rappellera également les grands principes de la garantie dommage, des métriques utilisées dans les modèles de tarification, et des modèles de Régression Linéaire Généralisés qui sont utilisés. Enfin, elle permettra de définir les différentes notions, référentiels et périmètres servant de base à l'étude.

1.1 Généralités

Aviva est le premier groupe d'assurance au Royaume-Uni et est l'un des leaders dans le secteur de l'assurance avec plus de 33 millions de clients dans le monde et 501 milliards de Livres d'actifs gérés par le groupe en 2019. Aviva France est une filiale du groupe Aviva et le deuxième contributeur au résultat opérationnel du groupe. Aviva France a elle-même plusieurs filiales et partenaires : Aviva Direct, Eurofil by Aviva, AFER, Union financière de France (UFF). Avec plus de 3 millions de clients en France et un réseau de plus de 1000 agents généraux, de 1000 courtiers partenaires et de son canal direct, Aviva France propose une large gamme de produits d'assurances en assurance vie, en épargne et en assurance dommage. Fin 2021, Aviva France devient Abeille Assurance.

L'assurance dommage ou assurance I.A.R.D. (Incendie, Accidents et Risques Divers) a pour objectif de couvrir les dommages aux biens, en indemnisant les assurés lors de la survenance d'un événement incertain. Le montant des cotisations de l'assurance de biens et responsabilités en France en 2020 s'élève à $60,1Md€$ soit $29,8\%$ du montant total de l'assurance en France. Les principales branches de l'assurance dommage sont : l'assurance habitation, l'assurance professionnelle, l'assurance emprunteur et l'assurance automobile. C'est sur cette dernière que porte cette étude.

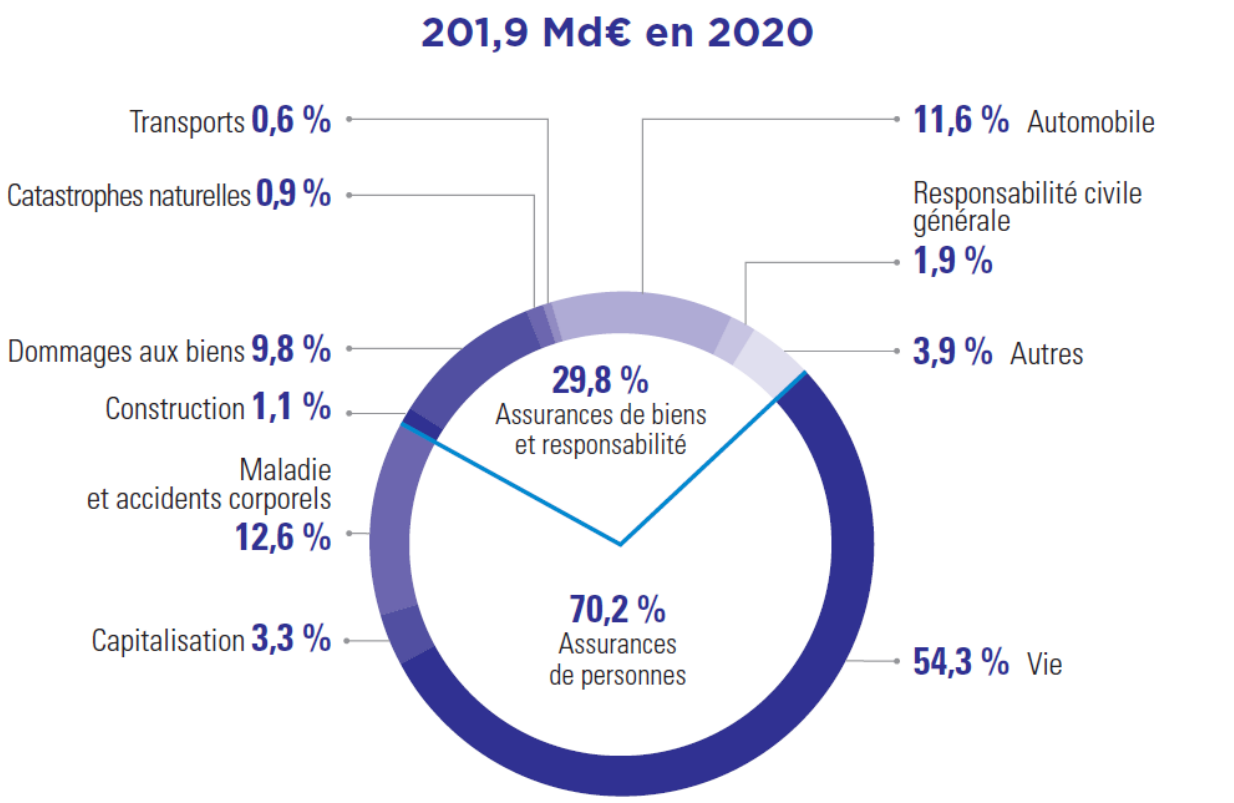


FIGURE 1.1 – Répartition des cotisations de l'assurance dommage en France[1]

1 FFA (2020), *L'assurance française Données clés 2020*

L'assurance dommage repose sur des éléments fondamentaux comme :

- L'aléa moral : qui est l'incitation qu'a l'assuré à prendre plus de risques sachant qu'il n'aura pas à assumer la prise en charge de ses risques ;
- L'anti-sélection : les assurés ayant une meilleure connaissance du risque qu'ils représentent, l'assureur doit segmenter ses risques pour faire face à cette asymétrie d'information ;
- La mutualisation des risques : les primes versées par tous les assurés servent à payer les sinistres de certains d'entre eux ;
- Le principe de non enrichissement : le montant d'indemnisation des sinistres ne doit pas dépasser la valeur des biens assurés.

1.2 L'assurance automobile

1.2.1 Généralités

L'assurance automobile est le secteur le plus concurrentiel de l'assurance dommage car il est considéré comme un produit d'appel. Elle s'adresse aussi bien aux particuliers qu'aux professionnels. En 2019, le chiffre d'affaires de l'assurance automobile est de $22,8 Md€$ soit 39% des cotisations en assurance dommages aux biens et de responsabilité civile.

Ces formules sont composées de différentes garanties. La liste suivante présente les principales :

- Les dommages causés aux tiers :
  - Responsabilité civile : seule garantie obligatoire, elle couvre les dommages matériels ou corporels causés aux tiers par le véhicule assuré.
- Les dommages subis par le véhicule :
  - Dommage : couvre les dommages matériels causés au véhicule assuré, à ses équipements et à ses accessoires. La garantie dommage se décompose en deux sous garanties : le dommage d'accident par collision et le dommage tous accidents. Le dommage d'accident par collision couvre les dommages matériels accidentels subis par le véhicule assuré uniquement lors d'une collision avec : un autre véhicule, un animal ou un piéton. Le dommage tous accidents couvre tous les dommages matériels accidentels subis par le véhicule assuré du fait d'un choc, de son versement, de son immersion ou du déplacement accidentel du chargement.
  - Vol : protège contre le vol ou contre la tentative de vol à l'encontre de son véhicule. Cette garantie peut inclure ou non les accessoires du véhicule.
  - Incendie : couvre contre un incendie, la chute de la foudre, une explosion, des dommages de natures électriques, tempête ou la grêle.
  - Bris de glace : couvre le coût des réparations ou du remplacement du pare-brise, de la lunette arrière, des glaces latérales, et autres éléments prévus aux garanties du contrat.
  - Protection du conducteur : protège le conducteur en cas d'accident quelle que soit sa responsabilité. Elle permet une indemnisation en cas de blessures corporelles comme par exemple les préjudices physiques, les dépenses de santé, ou encore l'éventuelle perte de gains futurs.

  Ces garanties sont facultatives. L'assuré doit les souscrire s'il souhaite être protégé pour ces risques. Il existe également d'autres garanties facultatives comme par exemple l'assistance ou encore la protection juridique.

1.2.2 La segmentation en assurance automobile

La forte concurrence en assurance automobile pousse les assureurs à être ultra compétitifs tant en termes de prix qu'en termes de garanties. Certes l'assurance automobile repose sur le principe de mutualisation, cependant il est important de trouver le juste équilibre entre mutualisation et

segmentation. L'objectif est de mutualiser le risque au sein d'une classe de homogènes. Ainsi, l'assureur s'assure que chaque assuré paie une prime qui reflète son niveau de risque et évite ainsi l'anti-sélection.

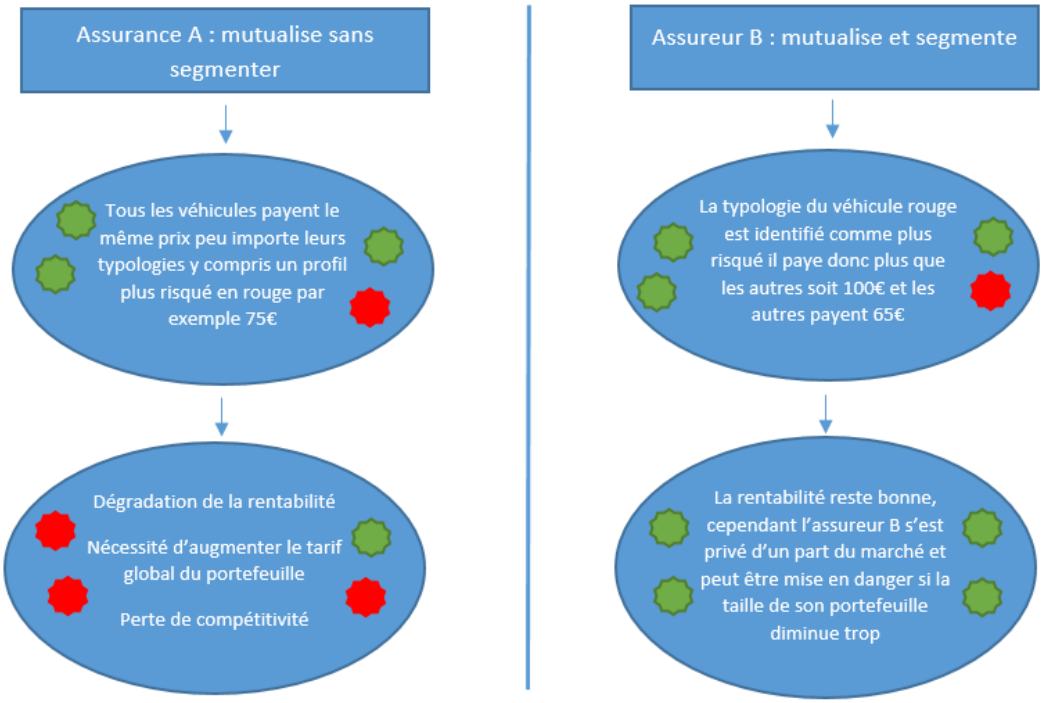


FIGURE 1.2 – Principe de segmentation

Une mauvaise segmentation entraîne une fuite des profils moins risqués, qui paieront moins chers chez un autre assureur pour des garanties équivalentes. En revanche, les profils plus risqués seront attirés par des prix plus attractifs qu'à la concurrence, du fait d'une moins bonne segmentation et d'une mutualisation plus forte.

En revanche, une segmentation trop fine entraîne un modèle trop complexe et une mauvaise estimation de la prime pour chaque classe de risques. En effet, la tarification repose sur la loi des grands nombres, si la segmentation est trop fine, il y a trop peu d'observations dans chaque classe de risques pour avoir une estimation fiable de la prime dans ces mêmes groupes. Il est donc essentiel de segmenter tout en gardant des classes de risques de tailles suffisamment conséquentes, afin de pouvoir mutualiser le risque au sein de ces classes. Il faut également noter que les méthodes de *machine learning*, peuvent accroître ce phénomène d'ultra-segmentation, si elles sont mal utilisées, .

Pour élaborer un tarif automobile l'assureur utilise un groupe restreint de données qui lui permettra d'estimer au mieux la prime pour chaque nouvel assuré. En tarification automobile ces données sont généralement divisées en 4 grandes catégories. Il existe toujours une part non expliquée, c'est le bruit.

FIGURE 1.3 – Décomposition de la Prime Pure

1.2.3 Qu'est-ce qu'un véhiculier ?

Le véhiculier est un regroupement des véhicules en classes de risques homogènes en fonction des variables véhicules, comme le poids, la valeur à l'origine du véhicule ou d'autres critères. Il permet une segmentation des véhicules suivants leurs caractéristiques propres. Pour être plus fin, un véhiculer peut différer selon les garanties.

Avec la sophistication constante des véhicules sur le marché, il est essentiel d'améliorer la segmentation de cette garantie restant l'une des plus souscrites. En effet, depuis plusieurs années, une baisse de la fréquence des sinistres est constatée, celle-ci est engendrée par une amélioration des systèmes de protections des véhicules. En revanche, cela entraîne également une hausse du coût moyen d'un sinistre, engendrée entre autre, par des coûts de réparations de plus en plus élevés.

L'objectif de ce mémoire est donc la création d'un véhiculier pour la garantie dommage pour ensuite l'intégrer au sein des différents modèles dommage. Le but de la création du véhiculier est de mieux mesurer l'impact de l'effet véhicule en dommage et de prendre en compte de nouvelles variables susceptibles d'être prédictives. De plus, dans un souci de simplification, les modèles sont souvent contraints en termes de nombre de variables. L'usage d'un véhiculier permet donc de palier cette contrainte, en regroupant toutes les informations véhicules en un seul score prédéfini. Enfin, l'augmentation constante du nombre de données oblige les assureurs à tester de nouvelles méthodes comme les méthodes de *machine learning* afin de mieux segmenter le risque.

1.3 La Prime Pure

La tarification d'une garantie a pour but d'estimer l'espérance de la charge sinistre associée. Elle est définie comme la prime pure que doit payer chaque profil de risque du portefeuille. La prime pure dépend du nombre de sinistres espérés et de l'espérance de la charge sinistre. Elle correspond le plus souvent à une approche fréquence - coût, et permet de modéliser la charge annuelle de sinistres attendue, qui sera notée S.

Les sinistres peuvent se décomposer en deux types, les sinistres dits « attritionnels » et les sinistres dits « graves ». Le nombre de sinistres attritionnels suit généralement une loi de Poisson de paramètre λ, mais peut également suivre une loi Binomiale Négative. Soit j, un sinistre quelconque de l'assuré i, soit K_i une variable aléatoire qui prend ses valeurs dans $\mathbb{N}$ et qui représente le nombre de sinistres de l'assuré i. Soit N_s, le nombre de total de sinistres pour tous les assurés, il correspond à :

$$N_s = \sum_{i=1}^{n} K_i$$

où n représente le nombre de polices ou l'exposition au risque.

Soit $X_{i,j}$ le coût aléatoire de chacun des K_i sinistres potentiels de l'assuré i. Soit S_i, la charge sinistre annuelle qui peut s'écrire

$$S_i = \sum_{j=1}^{n} \sum_{i=1}^{N_s} X_{i,j}$$

Sous l'hypothèse de l'indépendance entre la fréquence et la sévérité et pour chacune des classes de risque, la prime pure peut s'écrire de la manière suivante :

$$\mathbb{E}(S|X) = \mathbb{E}[(N_s)|X] \times \mathbb{E}[C|X]$$

1.4 Le Modèle Linéaire Généralisé

1.4.1 La famille de loi exponentielle

Le Modèle Linéaire Généralisé, appelé plus communément GLM, est le modèle le plus utilisé en tarification et présente un avantage majeur, celui d'être plus facilement interprétable. En effet, il permet d'identifier à une transformation près l'impact de chacune des variables explicatives sur la variable à expliquer Y. La théorie du GLM repose sur le même principe qu'une régression simple mais la variable à expliquer est une fonction de l'espérance de Y. Cette fonction sera appelée fonction de lien. La variable Y, fait en revanche nécessairement partie d'une famille de loi exponentielle.

Dans le cadre exponentiel, la densité de la loi de probabilité de Y s'écrira de la manière suivante :

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

avec y $\in S$, où S est un sous ensemble de $\mathbb{N}$ ou de $\mathbb{R}$, avec θ et ϕ appelé respectivement, paramètre naturel et paramètre de dispersion.

Le nombre de sinistres suit généralement une loi de Poisson de paramètre λ où S = $\mathbb{N}$, et la sévérité suit une loi Gamma de paramètre μ et ν où S =$\mathbb{R}^+$. Les densités s'écrivent donc respectivement :

$$f_\lambda(y) = exp(yln(\lambda) - \lambda + c(y))$$

et

$$f_{\theta,\phi}(y) = exp\left(\frac{-\frac{\mu}{\nu^2}y + ln\frac{\mu}{\nu^2}}{\frac{1}{\nu^2}}\right) + c(y,\nu)$$

Remarque : le nombre de sinistres peut également suivre une loi Binomiale négative, et le montant de sinistre une loi Log-Normale.

Cette étude se focalisera uniquement sur la création du véhiculier. En effet, le modèle GLM utilisé est déjà construit, et optimisé. Le modèle GLM en place ne sera donc pas détaillé.

1.4.2 Composante du modèle et fonction de lien canonique

Le Modèle Linéaire Généralisé dépend principalement de 3 éléments :

- Les variables à expliquer Y_i avec i = 1,...,n, ($i \in \mathbb{N}$) sont des variables aléatoires indépendantes mais non identiquement distribuées et dont la loi de densité appartient à la famille exponentielle ;
- Les variables explicatives X_i avec i = 1,...,n ($i \in \mathbb{N}$) et les coefficients issus de la régression associées aux variables explicatives $\beta = (\beta_0, \beta_1, ..., \beta_p)$ avec $p \in \mathbb{N}$;
- La fonction de lien g, qui s'exprime telle que :

$$g(\mu) = g(\mathbb{E}[Y|(x_1, ..., x_n)]) \tag{1.1}$$

$$= \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{1.2}$$

avec g qui est une fonction de lien strictement monotone et dérivable et avec p $\in \mathbb{N}$.

La fonction de lien établit la relation entre le prédicteur linéaire et l'espérance de la variable à expliquer. La fonction de lien varie en fonction de la loi utilisée pour expliquer Y.

Lien	Fonction de lien
Identité	$g(\mu) = \mu$
Log	$g(\mu) = \ln(\mu)$
Logit	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$

TABLE 1.1 – Exemple de fonctions de lien

La fréquence et la sévérité des sinistres sont les variables à expliquer dans cette étude. Les variables explicatives sont les variables « assurés », les variables liées au risque, les variables géographiques et les variables véhicules. Les variables géographiques sont regroupées dans le zonier et les variables véhicules dans le modèle initial seront quant à elles regroupées dans le véhiculier.

1.4.3 Estimation des paramètres et maximum de vraisemblance

La méthode du maximum de vraisemblance permet d'estimer les paramètres du modèle linéaire généralisé $\beta = (\beta_0, \beta_1, ..., \beta_p)$ et ϕ. L'objectif est de maximiser la fonction de log-vraisemblance. La vraisemblance peut être vue comme la probabilité que l'événement se réalise avec l'échantillon observé, elle sera notée L et s'exprimera pour $i = (1, ..., n)$ de la manière suivante :

$$L_{\theta,\phi}(y) = \prod_{i=1}^{n} f_{\theta,\phi}(y_i)$$

La log-vraisemblance s'exprimera donc comme suit :

$$L_{\theta,\phi}(y) = ln(\prod_{i=1}^{n} f_{\theta,\phi}(y_i)) \tag{1.3}$$

$$= \sum_{i=1}^{n} ln(f_{\theta,\phi}(y_i)) \tag{1.4}$$

$$= \sum_{i=1}^{n} l_{\theta,\phi}(y_i) \tag{1.5}$$

avec

$$l_{\theta,\phi}(y_i) = \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\right) + c(y_i, \phi)$$

Il existe alors une valeur pour le paramètre θ qui maximise la log-vraisemblance L. Cette valeur est déterminée grâce à la résolution des équations suivantes :

$$\frac{\partial ln(L)}{\partial \theta_i/\omega} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i} \tag{1.6}$$

$$= 0 \tag{1.7}$$

avec ω qui permet de pondérer par les observations.

1.4.4 Résidus

Les résidus du modèle correspondent à l'écart entre la valeur prédite par le modèle $\hat{\mu_i}$ et la valeur observée pour chaque individu y_i. Ils peuvent être considérés comme un écart de mesure. Ils se caractériseront pour chaque individu de la manière suivante :

$$\epsilon_i = y_i - \hat{\mu_i}$$

L'inconvénient de ces résidus est qu'ils ne sont pas hétéroscédastiques, c'est-à-dire que leur variance n'est pas constante. De plus, ils ne sont pas multiplicatifs. Ces résidus peuvent être positifs ou négatifs. Si le résidu est positif, le modèle aura tendance à sous-estimer la fréquence ou respectivement la sévérité. A l'inverse, un résidu négatif signifie que le modèle surestime la fréquence ou la sévérité.

Il existe plusieurs autres manières d'exprimer les résidus comme :

- Les résidus de Pearson : ce sont des résidus standardisés qui permettent de mesurer la contribution de chacune des observations à la significativité du test de cette statistique. Ils sont notés :
$$\epsilon_p = \frac{y_i - \hat{\mu_i}}{s_i}$$
avec s_i l'écart type des $\hat{\mu_i}$;

- Les résidus de déviance permettent de mesurer la contribution de chaque observation à la déviance du modèle par rapport au modèle parfait. Ces résidus existent aussi sous une forme standardisée et s'expriment de la manière suivante :
$$\epsilon_{D,i} = \pm\sqrt{d_i}(y_i - \hat{\mu_i})$$
où $D = \sum_{i=1}^{n} d_i$;

- Les résidus Anscombes permettent par une opération de transformation d'obtenir des résidus suivant une loi normale. De plus, ils sont hétéroscédastiques. Comme vu précédemment, la fréquence suit une loi de poisson et la sévérité une loi Gamma. Les résidus Anscombes pour une loi de poisson et pour une loi Gamma s'expriment respectivement de la manière suivante pour chaque individu i :
$$\epsilon_{a,i}^{f} = \frac{\frac{3}{2}(y_i^{\frac{3}{2}} - \hat{\mu}_i^{\frac{3}{2}})}{\hat{\mu}_i^{\frac{1}{6}}}$$

et

$$\epsilon_{a,i}^{cm} = \frac{3(y_i^{\frac{1}{3}} - \hat{\mu}_i^{\frac{1}{3}})}{\hat{\mu}_i^{\frac{1}{3}}}$$

.

Les résidus Anscombes seront utilisés pour construire les véhiculiers fréquence et sévérité.

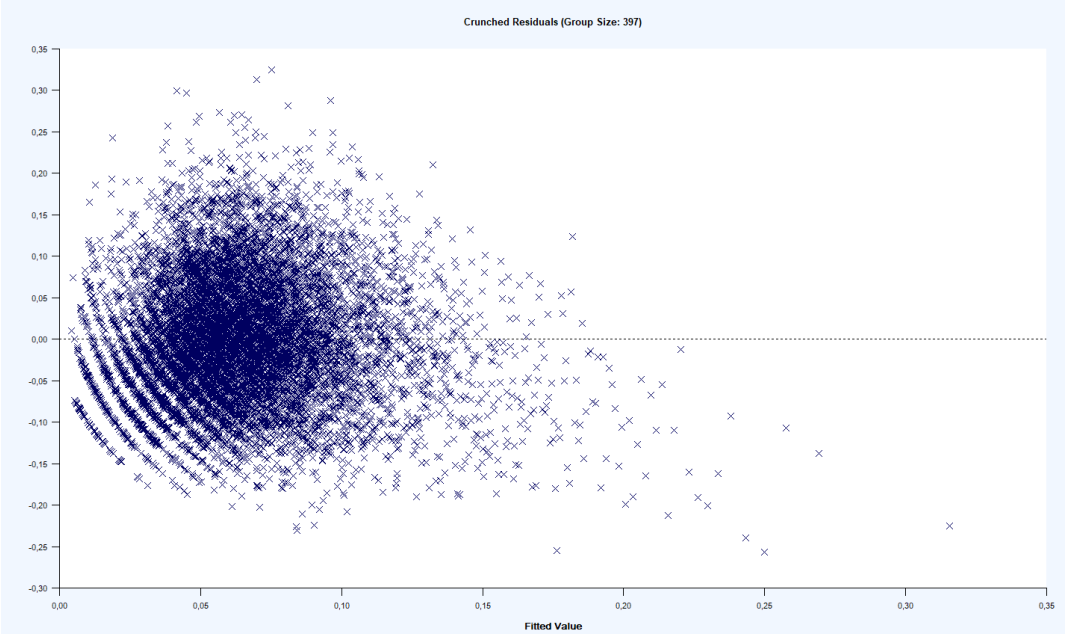


FIGURE 1.4 – Représentation des résidus du modèle fréquence

Ci-dessus les résidus du modèle fréquence complet, avec en abscisse la valeur prédite et en ordonnée le résidu de chaque observation. Il est donc possible de constater que les résidus sont bien centrés autour de zéro.

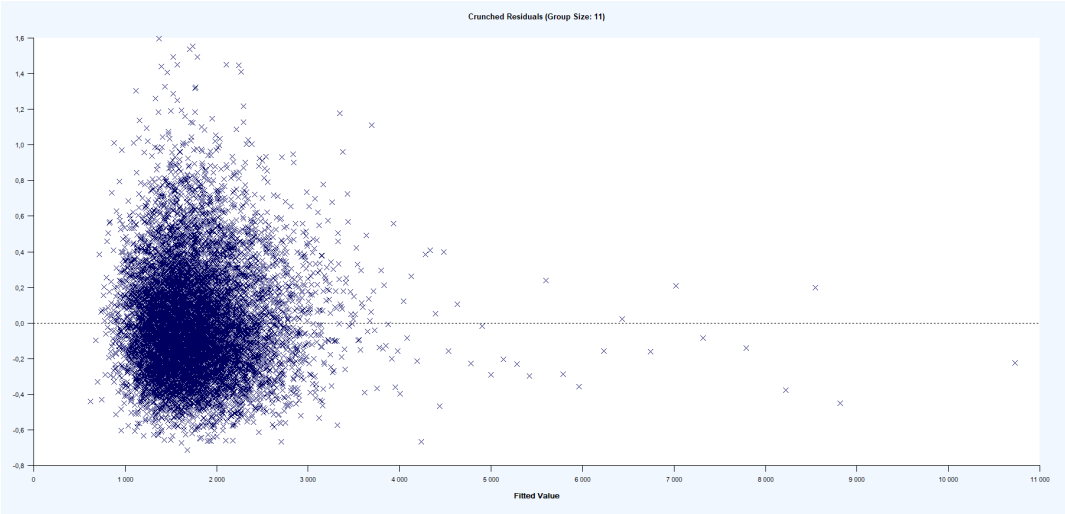


FIGURE 1.5 – Représentation des résidus du modèle CM

Concernant les résidus du modèle sévérité, ils sont bien centrés autour de zéro, en revanche contrairement au modèle fréquence, certains résidus sont élevés, mais reste très peu nombreux.

1.4.5 Extraction de l'effet véhicule

Comme décrit précédemment le risque dommage peut se décomposer en deux parties : la part du risque non liée au véhicule (l'information assurés, l'information liée à la zone géographique, et les données risque), et la part de risque liée au véhicule. C'est sur la part du risque véhicule que sera concentré le véhiculier. De plus, le modèle de prime pure se décomposant en deux modèles, deux véhiculiers distincts seront construits. Un véhiculier pour le modèle fréquence et un pour la sévérité.

Le modèle dommage existant contient déjà des variables véhicules. Pour la création du véhiculier, les variables véhicules sont retirées des modèles. Ainsi les autres coefficients se réajustent et captent une partie de l'effet précédemment capté par les variables véhicules. La part véhicule contient un effet véhicule et un bruit et peut s'exprimer de la manière suivant :

Part véhicule = GLM complet - GLM hors variables véhicules

Les coefficients du modèle hors variables véhicules sont utilisés pour extraire les résidus. La fréquence et la sévérité hors véhicules estimés pour chaque profil de risques s'expriment respectivement de la manière suivante :

$$\hat{\mu}_i^{'f} = e^{\beta_0^{'}} \times e^{\beta_1^{'} X_j} \times ... \times e^{\beta_k^{'} X_p}$$

et

$$\hat{\mu}_i^{'cm} = e^{\beta_0^{''}} \times e^{\beta_1^{''} X_j} \times ... \times e^{\beta_k^{''} X_p}$$

Les résidus Anscombes sont calculés pour chaque individu, puis agrégés au niveau véhicule. Ce sont ces résidus qui seront prédits et expliqués par les variables véhicules à l'aide des méthodes de *machine learning* dans le chapitre suivant.

Chapitre 2

Classification des résidus et lissage

Cette partie présentera les données utilisées et les principaux retraitements réalisés sur celle-ci. Elle détaillera ensuite la méthodologie utilisée pour prédire les résidus extraits précédemment, ainsi que les résultats obtenus par les différents modèles. Cette partie décrira les méthodes employées pour créer la cartographie des véhicules. Une fois les véhicules représentés dans un plan, la triangulation de Delaunay sera détaillée. Elle permettra d'identifier les véhicules considérés comme voisins. Enfin, le lissage des résidus prédits sera expliqué.

2.1 Base de données et retraitements

2.1.1 La base de tarification

Des contraintes internes imposent à cette étude l'utilisation de données peu récentes : la période d'observation est 2013-2017. Le véhiculier sera construit sur les données véhicules, seules ces dernières seront détaillées. En revanche, il est important pour le reste de ce mémoire d'expliquer comment est construite la base de tarification.

Le point de départ de la base de modélisation est la base image. Celle-ci permettant d'avoir une image des contrats pour chacun de leur mouvement : avenant, résiliation, ou autres. Pour chacune des images, la date de début et de fin de l'image sont utilisées pour calculer l'exposition au risque de chaque contrat. L'exemple ci-après illustre ce calcul pour un contrat qui aurait été souscrit le 01/05/2015 et qui aurait fait un premier avenant le 25/11/2015 et aurait été résilié le 07/03/2016.

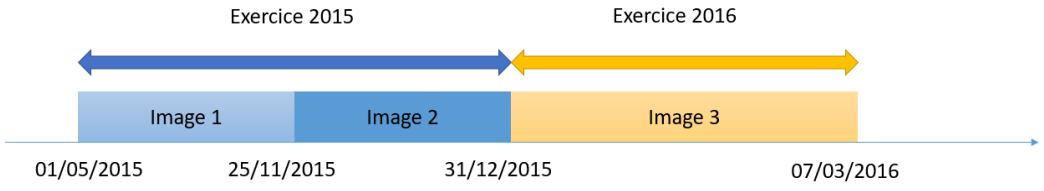


FIGURE 2.1 – Illustration de la construction de la base image

A chaque image sera associée, les informations du contrat (données de l'assuré, données

risques, données géographiques, données véhicules) et les sinistres en fonction de leur date de survenance. L'exposition, le nombre de sinistres et le montant de sinistres pour chacune des images de chaque contrat peuvent ainsi être calculés.

2.1.2 La base véhicule

La base de données utilisée est la base mise à disposition des assureurs par un organisme professionnel : SRA ou Sécurité et Réparation Automobile. Cette base de données répertorie les véhicules terrestres à moteur destinés au marché de l'automobile française. La quasi-totalité des véhicules y sont présents à l'exception de quelques marques ou de véhicules sortis en édition limitée. Chaque véhicule est identifié par un identifiant et contient un peu plus de 125 000 véhicules au moment de cette étude.

La base recense les principales caractéristiques de chaque véhicule. Elles peuvent être regroupées en plusieurs catégories :

- Caractéristiques moteurs :
  - La puissance DIN
  - La vitesse en km/h
  - La position du moteur
  - Disposition des cylindres
- Caractéristiques techniques :
  - Alimentation
  - Énergie : équivalent au carburant Gasoil, Essence, Bioéthanol...
  - Transmissions : 4 roues motrices, traction, propulsion
  - Boite de vitesse : manuelle, automatiques, semi-automatique
  - Nombre de rapports
  - Cylindrée
  - Le taux d'émission de CO2 du véhicule
- Genre et dimension du véhicule :
  - Carrosserie du véhicule : Break, Berline, Camionnette, VTC... La distinction est faite entre les véhicules 3 et 5 portes ce qui permet d'avoir une segmentation plus fine
  - Largeur du véhicule
  - Longueur du véhicule
  - Empattement : la distance entre l'axe des roues avant et l'axe des roues arrière
  - Poids à vide (en Kg)
  - Classe de prix : indice basé uniquement sur la valeur du véhicule (les modalités vont de "A" à "ZA")
  - Le groupe : représente la dangerosité du véhicule, les modalités sont comprises entre 20 et 50. Cette variable peut être interprétée comme un croisement du régime et de la puissance du véhicule.
  - La classe de réparation : définie le coût de réparation des pièces
- Options et Autres :
  - Types de freins

– Contrôle dynamique de stabilité
– Système de localisation
– Système localisation des obstacles

Cette liste est non exhaustive.

Les variables en option, sont souvent présentes uniquement sur les véhicules récents et sont donc souvent très peu remplies. C'est pourquoi une analyse des données est essentielle avant de commencer les modèles.

2.1.3 Retraitement des données

Une analyse des variables et de leurs modalités est nécessaire afin de préparer au mieux les données. La première étape consiste à repérer les variables inexploitables, du fait du trop grand nombre de valeurs manquantes.

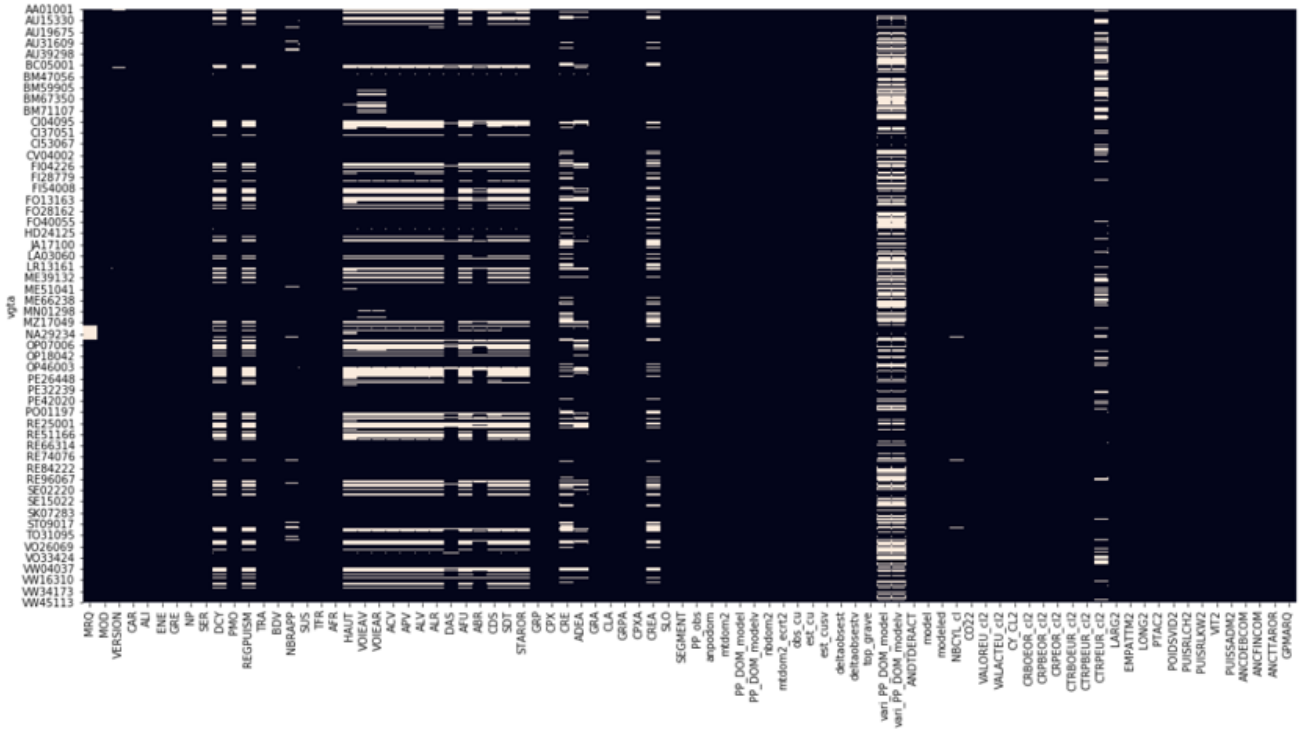


FIGURE 2.2 – SRA - Analyse des valeurs manquantes

La figure ci-dessus, est une vision partielle de la base de données véhicules, avec en abscisse les variables constitutives de la base et en ordonnée les individus, ici les véhicules. Les lignes blanches correspondent aux lignes où l'information est manquante. Les lignes noires correspondent aux lignes où l'information est renseignée. Ce graphique permet d'avoir une rapide connaissance du taux des valeurs manquantes de la base de données pour chaque variable.

Il permet également de repérer les variables potentiellement corrélées. En effet, les variables ayant des valeurs manquantes sur les mêmes lignes sont souvent dépendantes les unes des autres. Par exemple, la variable : ACV (Airbag conducteur), APV (Airbag passager avant), ALV (Airbags

latéraux avant), ALR (Airbags latéraux arrière).

Cependant, il ne s'agit pas forcément de valeurs réellement manquantes. L'information n'est pas remplie, mais il peut s'agir de données non disponibles au moment de la création de la variable. Plus précisément, certains véhicules sont des vieux véhicules, et au moment de leur commercialisation certaines options n'étaient pas disponibles (comme les airbags latéraux par exemple). Cependant la base étant mise à jour continuellement avec les nouveaux véhicules et les nouvelles options, ces informations ne sont disponibles que sur les véhicules les plus récents. Les variables avec plus de 45% de données manquantes sont supprimées de la base.

Les valeurs manquantes des variables catégorielles restantes sont remplacées par la modalité "NR" pour "Non Renseignée". Les valeurs manquantes des variables continues restantes sont traitées de la manière suivante :

– Valeur à l'origine et la vitesse : elles sont remplacées par la médiane des valeurs à l'origine et de la vitesse de chaque classe de prix.

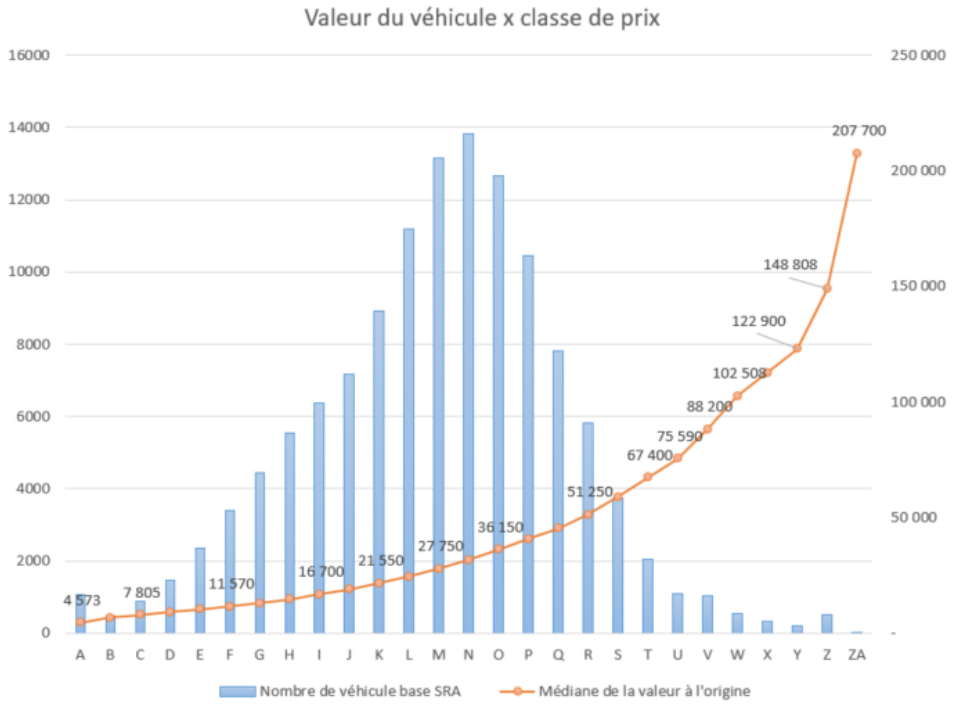


FIGURE 2.3 – SRA - Médiane de la valeur à l'origine du véhicule en fonction des classes de prix

– La puissance et le poids à vide : les modalités manquantes de la puissance et du poids à vide sont remplacées par la médiane des puissances ou du poids à vide de chaque groupe. Le groupe est un classement reposant principalement sur le rapport poids puissance.

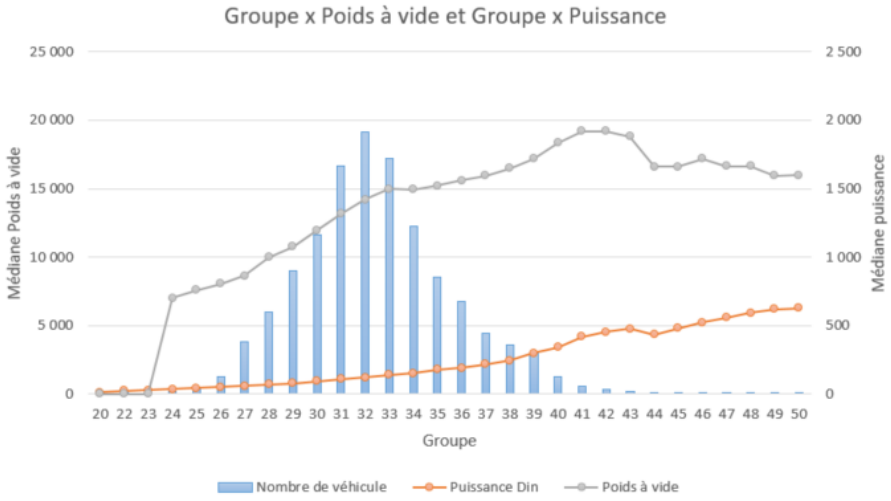


FIGURE 2.4 – SRA - Médiane du poids à vide et de la puissance en fonction du groupe

– Nombre de places et le nombre de rapport : les modalités manquantes sont remplacées par la modalité la plus représentée.

– Les autres variables numériques à savoir la longueur, la largeur, l'empattement, la cylindrée et le CO2 sont discrétisées en groupes de tailles égales. Ainsi, chaque groupe contient un nombre significatif d'individus.

Finalement, les variables sélectionnées sont :

Variables	Typologie de variables	Libellées variables
VIT	Variables quantitatives	Vitesse maximum (en Km/h)
NP		Nombre de places
NBRAPP		Nombre de rapports
CY_cl		CY transformé en numérique
LARG3		Largeur
EMPATTM3		Empattement
LONG3		Longueur
POIDSVID		Poids à vide (en Kg)
PUISRLCH		Puissance DIN (en CV DIN)
VALOREU_cl3		Valeur à l'origine
MRQ	Variables qualitatives	Code Marque
CAR		Carrosserie
ALI		Alimentation
ENE		Energie
GRE		Genre
SER		Série limitée
NBCYL		Nombre de cylindres
DCY		Disposition des cylindres
PMO		Position du moteur
TRA		Transmission
BDV		Boîte de vitesses
SUS		Suspension
TFR		Type de freins
AFR		Assistance de freinage
ACV		Airbag conducteur
APV		Airbag passager avant
ALV		Airbags latéraux avant
ALR		Airbags latéraux arrière
DAS		Direction assistée
groupe		Groupe
CPX		Classe de prix à l'origine
groupa		Groupe APSAD
AFU		Assistance de freinage d'urgence
ABR		Antiblocage de roues
CDS		Contrôle dynamique de stabilité
CRE		Classe de réparationà l'origine
ADEA		Antidémarrage actuel
SLO		Système de localisation
SEGMENT		Segment
CO22		CO2 discrétisé
CRBOEOR		Coût à l'origine de remplacement d'un bloc optique discrétisé
CRPBEOR		Coût à l'origine de remplacement d'un pare brise discrétisé
CRPEOR		Coût à l'origine de remplacement des pièces discrétisé
HAUT		Hauteur (en mm) discrétisée
VOIEAV		Voie avant (en mm) discrétisée

TABLE 2.1 – SRA - Variables SRA finales retraitées

2.2 Classification à l'aide des méthodes de *machine learning*

2.2.1 Sélection des variables et traitement des données

Les résidus sont désormais calculés au niveau véhicule. Il s'agit maintenant de les prédire et de les expliquer à l'aide de méthodes de *machine learning* et des données véhicules disponibles suite aux différents retraitements de la base de données. Deux méthodes dites de *machine learning* ont été utilisées pour prédire les résidus. Les forêts aléatoires ou *Random Forest*, et le *XGBoost*. Afin de s'assurer qu'il n'y ait pas d'effet de sur-apprentissage, la base de données est divisée en base d'apprentissage et en base de test. Les modèles seront calibrés sur la base d'apprentissage qui contiendra 80% des observations et seront validés sur la base de test contenant 20% des observations.

De plus, ces deux algorithmes ne permettent pas de limiter les effets de corrélation entre les variables et peuvent être utilisés avec toutes les variables aussi corrélées soient-elles. Ainsi un travail préalable est nécessaire afin de limiter les effets de corrélation et d'exploiter le potentiel complet de ces algorithmes.

Plusieurs méthodes existent pour mesurer le lien entre les variables, le plus souvent elles ne sont utilisables que sur une seule typologie de données : quantitatives ou bien qualitatives. Or, la base contient les deux typologies de variables. Le V de Cramer a l'avantage de gérer les deux. Il permet de calculer les corrélations entre chaque variable, deux à deux, quelle que soit leur typologie. Il donne une valeur entre zéro et un qui détermine le niveau de dépendance entre les variables.

Plus la valeur est proche de un, plus la relation est considérée comme forte, et inversement, plus cette valeur est proche de zéro, plus la relation entre ces variables est considérée comme faible. Cette valeur correspond donc à l'intensité de liaison entre deux variables.

Le V de Cramer est basé sur la statistique du χ^2 et s'exprime de la manière suivante :

$$V \quad de \quad Cramer = \sqrt{\frac{\chi^2}{\chi^2_{max}}} \tag{2.1}$$

$$= \sqrt{\frac{\chi^2}{n \times (min(l,c) - 1)}} \tag{2.2}$$

avec n le nombre de colonnes et l le nombre de lignes, avec n $\in \mathbb{N}$ et l $\in \mathbb{N}$.

La première étape consiste à fixer un seuil d'acceptabilité, ce seuil sera le niveau à partir duquel, une liaison entre deux variables sera supposée trop forte. Le nombre de variables étant faible, le seuil d'acceptabilité a été augmenté de manière à avoir suffisamment de variables pour les modèles.

Niveaux de corrélation	Intensité de la liaison
Non corrélées	]0 ; 0,2]
Peu corrélées	]0,2 ; 0,3]
Moyennement corrélées	]0,3 ; 0,5]
Fortement corrélées	> 0,5

TABLE 2.2 – Seuil de corrélation - V de Cramer

Le seuil d'acceptabilité choisi est 0,5. Toutes les variables pour lesquelles l'intensité de liaison dépasse 0,5 ne sont pas utilisées dans les modèles *Random Forest* et *XGBoost* car elles sont considérées comme ayant un lien trop important entre elles.

La base contient des variables comme « la classe de prix » du véhicule qui correspond à une classification des véhicules en fonction de leur valeur à l'origine, ou encore « le groupe » qui est un classement basé sur le poids et la puissance du véhicule. Or, la base de données contient également ces variables en format continu, à savoir la valeur à l'origine du véhicule, la puissance et le poids à vide du véhicule. Il s'agit de faire un choix entre les variables continues et les variables classifiées. Aussi, afin d'être le plus fin possible, les variables continues sont conservées au détriment des variables catégorielles. Les variables finales conservées pour la prédiction des résidus sont donc les suivantes :

Variables conservées	
Puissance du véhicule	Marque
Valeur du véhicule	Alimentation
Poids à vide (en kg)	Classe de réparation
Vitesse	Types de freins
Carrosserie	Emission de CO2
Cylindrée	Largeur et Longueur

TABLE 2.3 – Variables conservées pour les modèles

Le logiciel de tarification utilisé est le logiciel Emblem, qui limite le nombre de modalités maximum utilisables à 350. L'une des premières améliorations notable de l'utilisation des méthodes de *machine learning* dans cette étude, est donc la possibilité d'utiliser plusieurs variables continues comme la puissance, la valeur, le poids, la cylindrée et la vitesse du véhicule qui n'étaient pas utilisables dans les modèles.

2.2.2 Les forêts aléatoires : théorie et application

La théorie des forêts aléatoires découle de celle des arbres de décision. Elle combine arbres de décision et *Bagging*.

Arbre de décision :

Un arbre de décision est un ensemble de nœuds, desquels partent 0, 1 ou 2 branches. Un nœud sans branche est appelé feuille. A chaque nœud, une décision binaire est prise afin de partitionner de manière récursive un espace X de manière optimale pour la prédiction. L'arbre peut tout aussi bien prédire des variables catégorielles, que des variables continues. Dans le cas de variables catégorielles, on parlera d'un arbre de classification. Dans le cas de variables continues, il s'agira d'un arbre de régression. La variable à prédire dans cette étude est continue, ce sont les résidus.

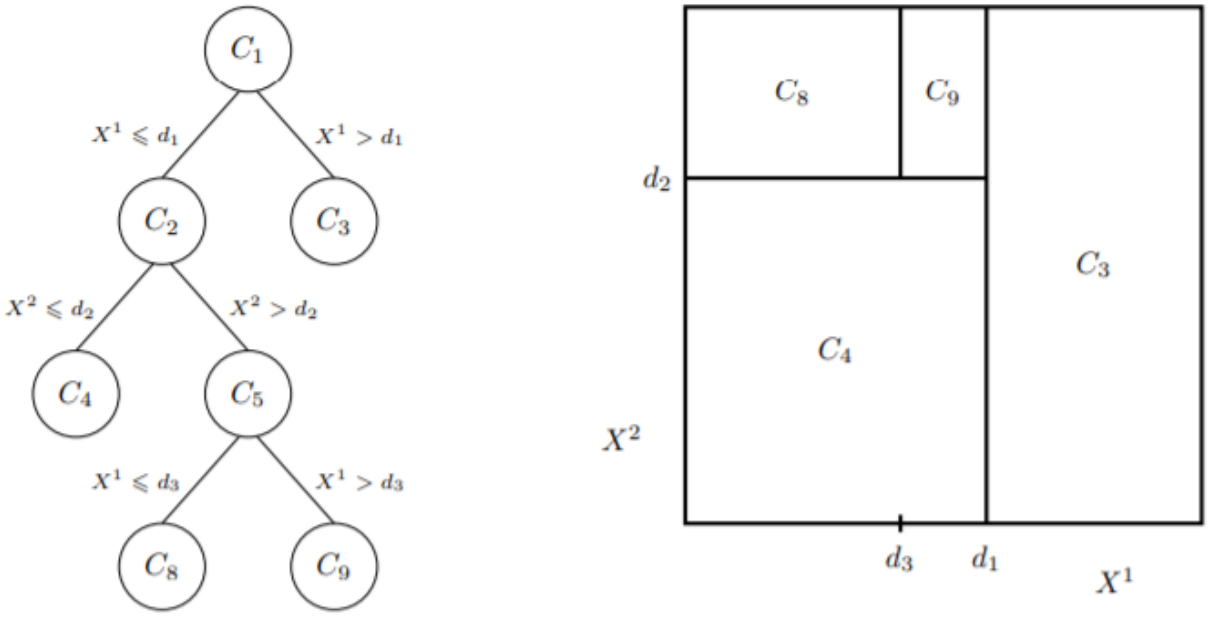


FIGURE 2.5 – Arbre de décision et partitionnement de l'espace X

A gauche, l'arbre est construit à l'aide des décisions binaires avec C_1, la racine de l'arbre qui se sépare ensuite en deux branches, qui sont : un autre nœud C_2 et une feuille C_3. Les opérations se poursuivent jusqu'à former l'arbre dit « maximal ». Cet arbre est associé à l'espace X à droite. Cet espace X est partitionné à l'aide des variables explicatives choisies pour construire l'arbre de décision de gauche.

En effet, en C_1 les individus dont la variable X^1 est telle que $X^1 \leq d_1$ sont envoyés dans le nœud C_2 et les autres dans la feuille C_3.

N.B. : X^1, ..., X^p peuvent être des variables quantitatives ou qualitatives (p est le nombre de variables).

Construction de l'arbre maximal : L'arbre de décision résultant de décisions binaires, chaque partitionnement permet de découper l'espace en deux sous-espaces. La première étape consiste à séparer N en deux nœuds N_d et N_g. La découpe est réalisée à l'aide d'un couple (j,s) qui permet de minimiser la fonction de coût : j correspond à une variable et s correspond à un seuil, tel que :

$$N_d(j,s) = \{x \in N : x_j < s\}$$

et

$$N_g(j,s) = \{x \in N : x_j \geq s\}$$

Dans le cadre d'une régression, l'objectif est de minimiser la variance intra-groupes en mesurant le gain d'information pour chaque paire (j,s) résultant de la découpe d'un noeud N en deux autres nœuds.

La variance d'un nœud N est notée :

$$V(N) = \sum_{i:x_i \in N} (y_i - \bar{y}_N)^2$$

avec

$$\bar{y}_N = \frac{1}{|N|} \sum_{i:x_i \in N} y_i$$

et

$$|N| = \#\{i : x_i \in N\}$$

Ce qui implique donc de minimiser :

$$\frac{|N_d(j,s)|}{|N|} V(N_d(j,s)) \quad + \quad \frac{|N_g(j,s)|}{|N|} V(N_g(j,s))$$

Bagging :

Le *Bagging* consiste à entraîner plusieurs arbres de décision sur un échantillon en remplaçant le jeu de données à chaque itération. Ainsi, la prédiction sera la moyenne des prédictions données par chaque arbre pour chaque individu, et non plus la prédiction d'un seul arbre de décision. Ceci permet de réduire le bruit et la variance d'un seul arbre. L'avantage est qu'en changeant le jeu de données, les arbres ne sont pas corrélés, car le *Bagging* aide à réduire la corrélation des arbres.

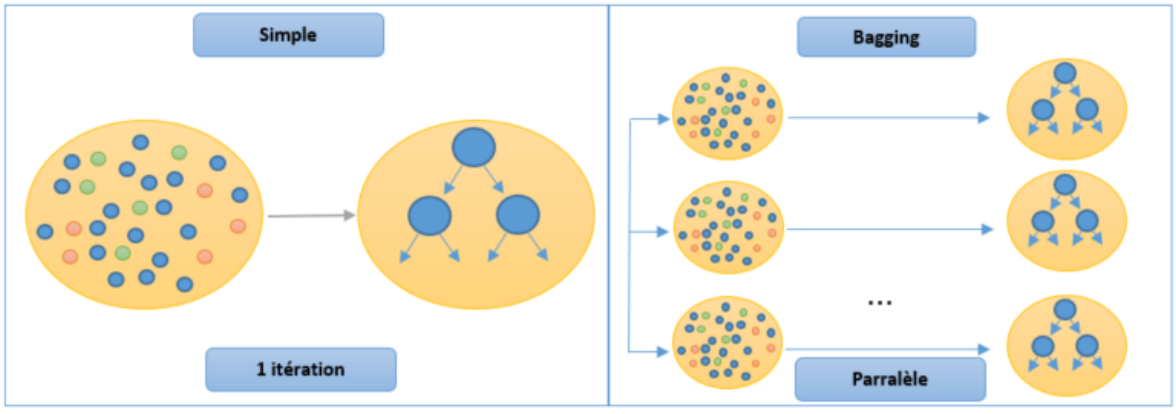


FIGURE 2.6 – Principe du *Bagging*

L'un des avantages du *Random Forest* est qu'il peut traiter un jeu de données avec un petit nombre d'observations et un grand nombre de variables. En revanche, non maîtrisé, il peut conduire à faire du sur-apprentissage.

Afin de limiter le sur-apprentissage tout en ayant la meilleure prédiction possible, plusieurs actions ont été mises en oeuvre :

- la séparation de la base en deux bases, une base d'apprentissage et une base de validation. La base d'apprentissage contient 80% des données et la base de validation contient 20% des données.
- limiter la profondeur des arbres de manière à ne pas avoir l'arbre maximal
- imposer un nombre minimum d'individus par feuille finale
- le choix de la fonction de coût

L'objectif premier est d'obtenir la meilleure prédiction possible, en optimisant les hyper-paramètres du modèle *Random Forest*. Une validation croisée est utilisée. Elle consiste à croiser un grand nombre de paramètres afin de déterminer ceux qui permettent de minimiser l'erreur de prédiction en fonction de la fonction de coût choisie. Ainsi, peuvent être testés : le nombre d'arbres optimaux, le nombre de splits optimaux, etc. Toutes les combinaisons de paramètres sont alors testées de manière exhaustive et itérative.

La méthode consiste à découper une base de données en k échantillons. Chaque échantillon est subdivisé en x morceaux pour constituer l'échantillon d'apprentissage. Les k-x échantillons restants permettent d'évaluer la performance du modèle. Pour construire le modèle, les échantillons sont sélectionnés différemment de manière à ne jamais avoir les mêmes échantillons d'apprentissage et de validation. Le choix du nombre d'échantillons k peut également être testé. Dans cette étude il a été décidé de le fixer à 5 pour respecter le 80% / 20%.

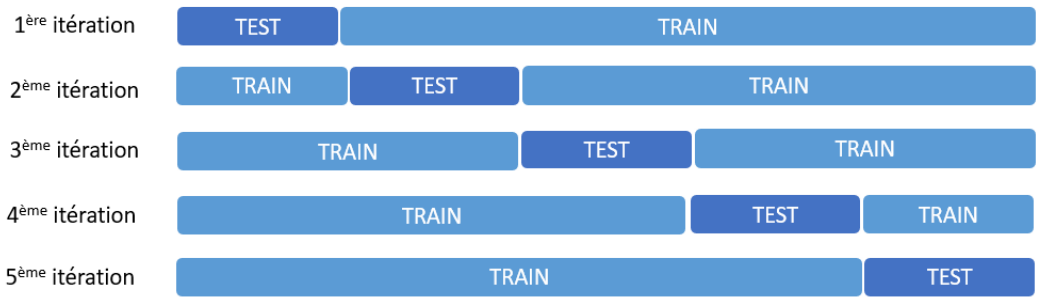


FIGURE 2.7 – Principe de validation croisée

Une fois que chaque modèle a pu être entraîné et évalué, il ne reste plus qu'à comparer la performance pour choisir le meilleur modèle.

Évaluation des modèles :

La robustesse des modèles peut être évaluée de plusieurs manières. L'évaluation des modèles est différente selon la variable à expliquer : catégorielle ou continue. Pour les modèles continus

comme dans cette étude, l'évaluation est faite à l'aide des indicateurs ci-après :

– *Root Mean Squarred Error* où l'erreur quadratique moyenne qui correspond à la racine carrée de la variance notée :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Cet indicateur permet de mesurer l'écart entre la valeur prédite par le modèle et la valeur observée. Ils permettent de choisir les hyper-paramètres optimaux. Plus la valeur du RMSE est petite, meilleure est la prédiction puisque que l'écart entre la valeur observée et la valeur prédite est faible.

Construction et évaluation du *Random Forest* :

Les hyper-paramètres testés dans la validation croisée sont les suivants :

Hyper-paramètres testés	Fréquence	Sévérité
Nombre d'arbres	50/100/200/300	50/100/200
Profondeur maximale de l'arbre	3/5/8	3/5/8

TABLE 2.4 – Critères de validation croisée

Les résultats de la *cross validation* sont les suivants :

Random Forest	Nombre d'arbres	Nombre de splits
Fréquence	200	5
Sévérité	100	3

TABLE 2.5 – Résultats de la validation croisée pour la fréquence et pour la sévérité

Le choix du nombre d'arbres peut être contrôlé de manière visuelle, en traçant l'évolution du RMSE en fonction du nombre d'arbres.

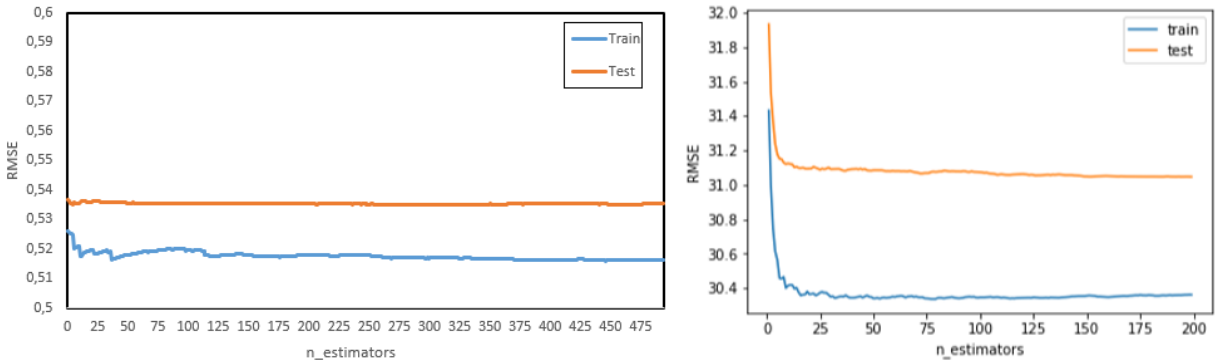


FIGURE 2.8 – Évolution RMSE en fonction du nombre d'arbre

A gauche, l'évolution du RMSE pour le modèle *Random Forest* fréquence et à droite l'évolution du RMSE pour le modèle de sévérite. Pour le *Random Forest* fréquence, le choix optimal de 200 arbres se justifie puisque le RMSE est très instable au voisinage 50 et 100 arbres. En revanche, l'erreur se stabilise à 200 arbres.

Pour le modèle *Random Forest* sévérité, l'erreur se stabilise peu avant 100 arbres. Que ce soit pour le modèle fréquence ou bien le modèle de sévérité, choisir un nombre d'arbres supérieur n'améliorerait en rien la prédiction.

Une fois les modèles calibrés, il est important de pouvoir les interpréter afin de vérifier leur cohérence. Le *Bagging* permet de limiter le biais et d'améliorer les prédictions. Cependant, ses résultats sont moins lisibles et sont plus difficiles à interpréter qu'un simple arbre de décision. Aussi, afin de faciliter la compréhension des modèles construits, les variables sont classées par ordre d'importance. L'importance correspond aux variables qui contribuent le plus aux résultats du modèle.

Le graphique ci-dessous classe les variables par ordre d'importance en termes de gains de RMSE.

Pour la fréquence ci-après, la valeur du véhicule est la variable la plus discriminante dans la prédiction du résidu puisqu'elle contribue à plus de 25% au modèle. Plus la voiture est chère plus elle sera susceptible d'avoir des aides à la conduite et donc plus la fréquence est susceptible de diminuer.

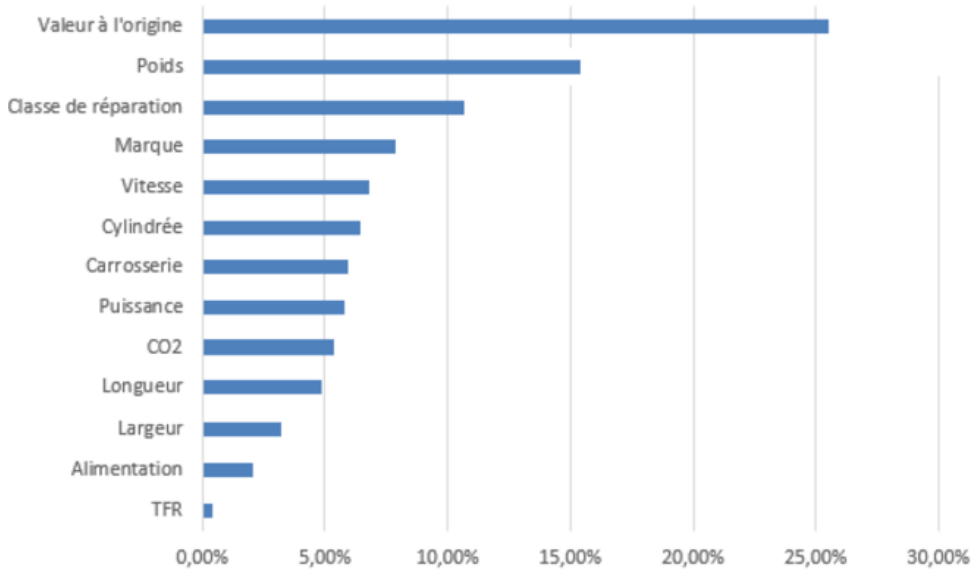


FIGURE 2.9 – Importance des variables - *Random Forest* fréquence

Concernant la sévérité, la vitesse et la valeur du véhicule sont les plus discriminantes. Viennent ensuite, la marque, la puissance, et le régime moteur. En revanche, le type de freins et l'alimentation sont les variables qui contribuent le moins aux modèles que ce soit pour le modèle fréquence ou le modèle sévérité.

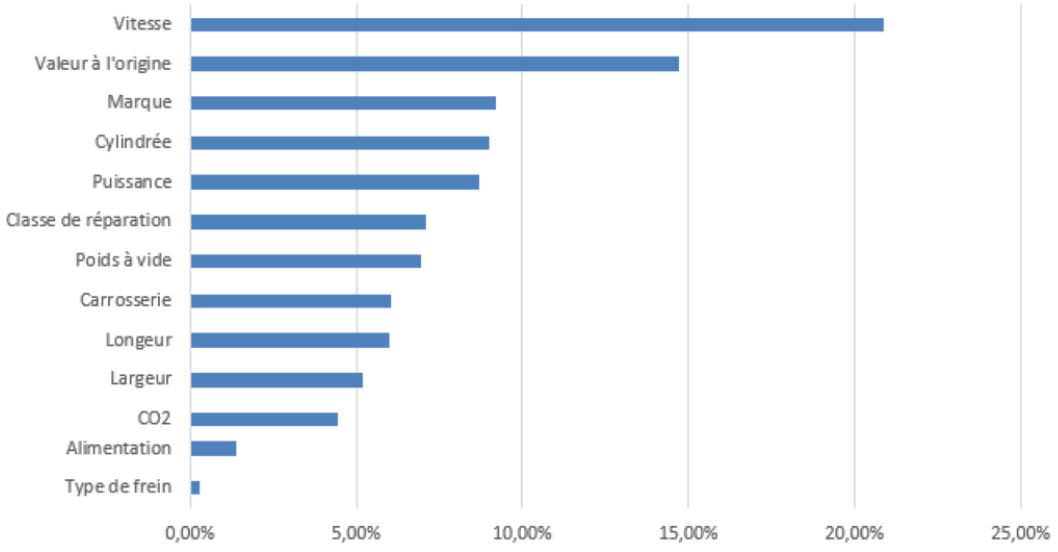


FIGURE 2.10 – Importance des variables - *Random Forest* sévérité

2.2.3 *XGBoost* : théorie et application

L'algorithme *XGBoost* (*Extrême Gradient Boosting*) est un modèle à apprentissage supervisé servant à expliquer la variable cible Y, dont y_i est la ième réalisation, à partir d'un jeu de données

x_i (les variables SRA). Cet algorithme combine linéairement des arbres de décision simples, appelés« apprenants faibles », pour en fabriquer un plus fort g_B. Ainsi, la performance est améliorée : c'est le *Boosting* (opposable au *Bagging* utilisé dans le *Random Forest*).

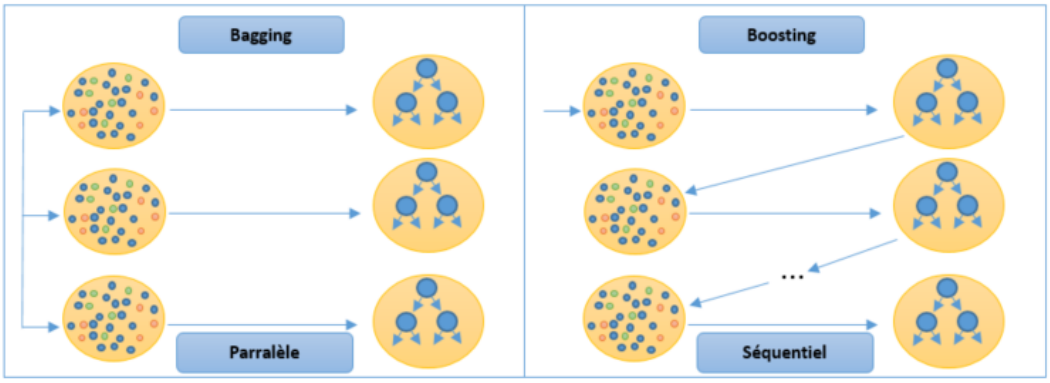


FIGURE 2.11 – Principe du *Bagging* et du Boosting

Il existe plusieurs algorithmes de boosting : *AdaBoost*, *Gradient Boosting*, ou *XGBoost*. Ce dernier est inspiré du *Gradient Boosting*.

Gradient Boosting :

L'algorithme attribue le même poids à chaque apprenant faible. Chaque apprenant faible est entraîné afin de corriger les erreurs des apprenants faibles précédents. Un premier apprenant faible est construit et représente simplement la moyenne des observations. Cette moyenne estimée sert de base pour construire le reste de l'algorithme. Le *Gradient Boosting* cherche, à réduire l'écart entre cette valeur moyenne et la valeur réelle observée.

Les apprenants faibles calibrés ensuite cherchent donc à prédire ce premier écart. La valeur prédite par ce second apprenant faible est multipliée par un coefficient inférieur à 1 et ainsi de suite. L'objectif est de se rapprocher à chaque étape de la valeur observée afin de réduire l'erreur de prédiction.

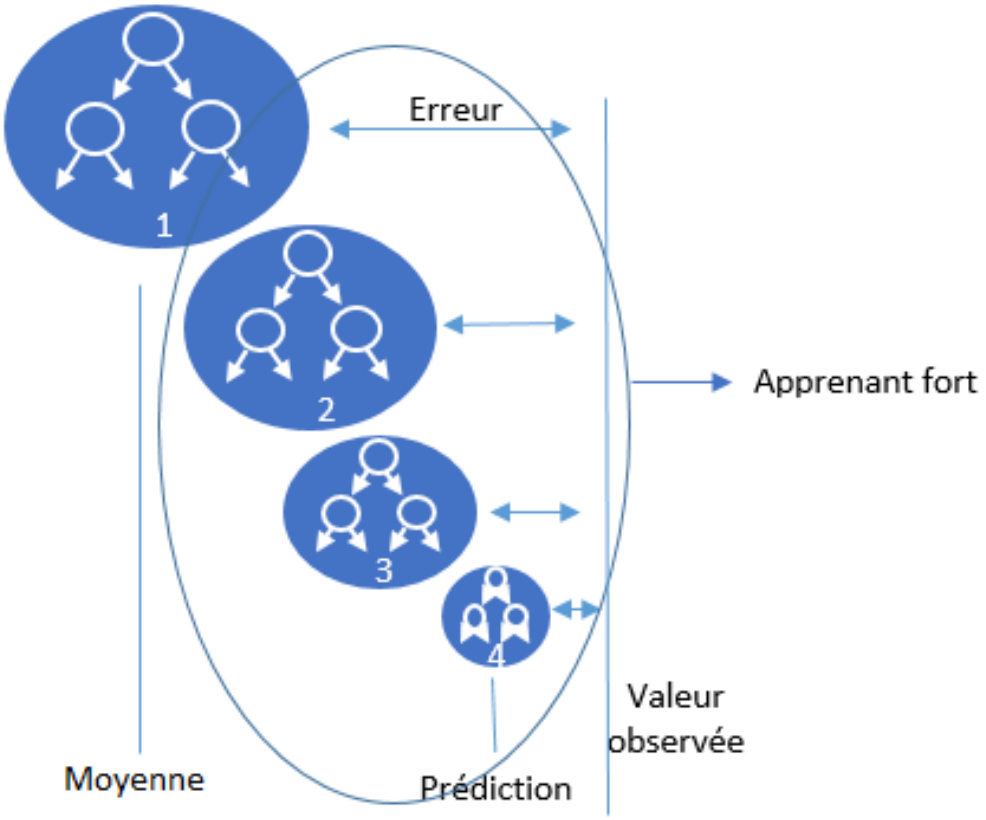


FIGURE 2.12 – Principe du *Gradient Boosting* - Descente de gradient

Soit H un nombre fini d'apprenants faibles et chaque apprenant faible est noté h : $h_1, ..., h_B \in H$. Chaque apprenant faible est basé sur la connaissance de l'apprenant faible précédent. L'apprenant fort final est une somme pondérée des apprenants faibles qui le composent. Il peut donc s'exprimer tel que :

$$g_B(x) = \sum_{b=1}^{B} \eta_b h_b(x)$$

avec b = 1, ... , B une étape de boosting et $\eta_b \in \mathbb{R}$.

L'apprenant fort a pour but de minimiser le risque empirique en donnant la meilleure approximation possible, à l'aide d'une fonction de coût notée ℓ :

$$(\hat{\eta}_{b+1}, \hat{h}_{b+1}) = argmin_{\eta \in \mathbb{R}, h \in H} \sum_{i=1}^{n} \ell(y_i, \hat{g}_b(x_i) + \eta h(x_i))$$

et les itérations entre les apprenants faibles s'expriment de la manière suivante :

$$\hat{g}_{b+1} = \hat{g}_b + \hat{\eta}_{b+1} \hat{h}_{b+1}$$

Dans un esprit de cohérence, les hyper-paramètres testés sont les mêmes que pour le *Random Forest*. Les paramètres qui minimisent le RMSE sont :

XGBoost	Nombre d'arbres	Nombre de splits
Fréquence	100	5
Sévérité	200	8

FIGURE 2.13 – Résultats de la validation croisée pour le *XGBoost*

Comme pour le *Random Forest*, la contribution de chaque variable dans la prédiction a été calculée afin d'essayer d'expliquer les résultats. Les variables les plus importantes sont différentes de celle du *Random Forest*. Pour la fréquence, la marque est la variable qui contribue le plus au modèle. En revanche, le CO2 prend la deuxième place, alors qu'elle n'était que neuvième pour le *Random Forest*. L'émission de CO2 étant proportionnelle à la valeur et au poids du véhicule ceci peut expliquer l'importance de cette dernière variable dans le classement.

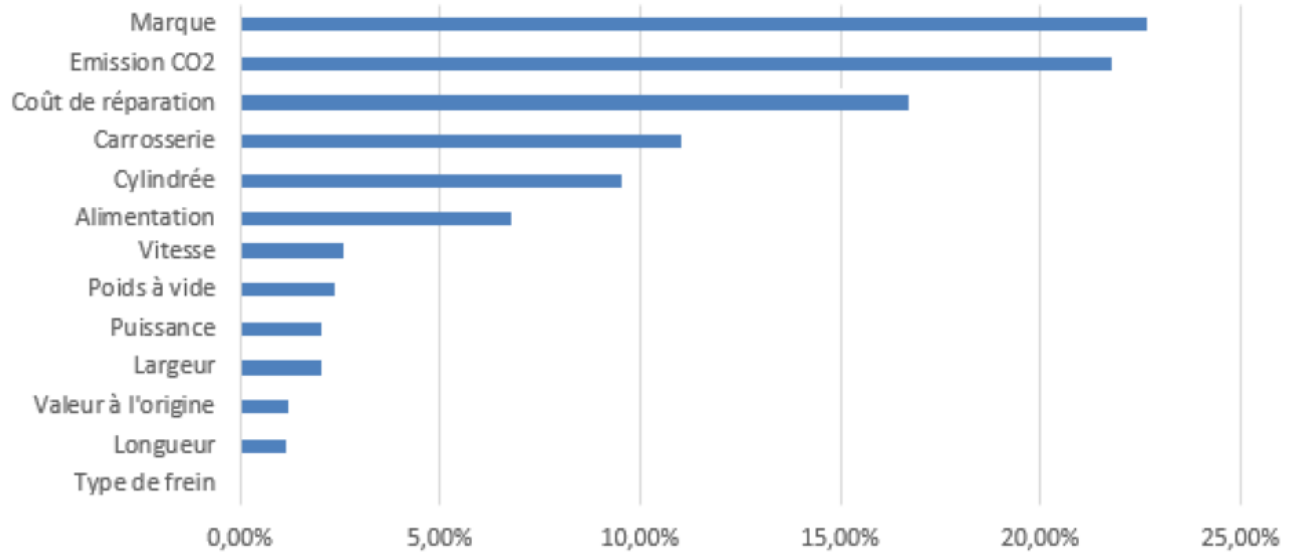


FIGURE 2.14 – Importance des variables - *XGBoost* fréquence

Pour la sévérité, ce sont le coût de réparation, la marque et la carrosserie qui ressortent comme étant les plus prédictives.

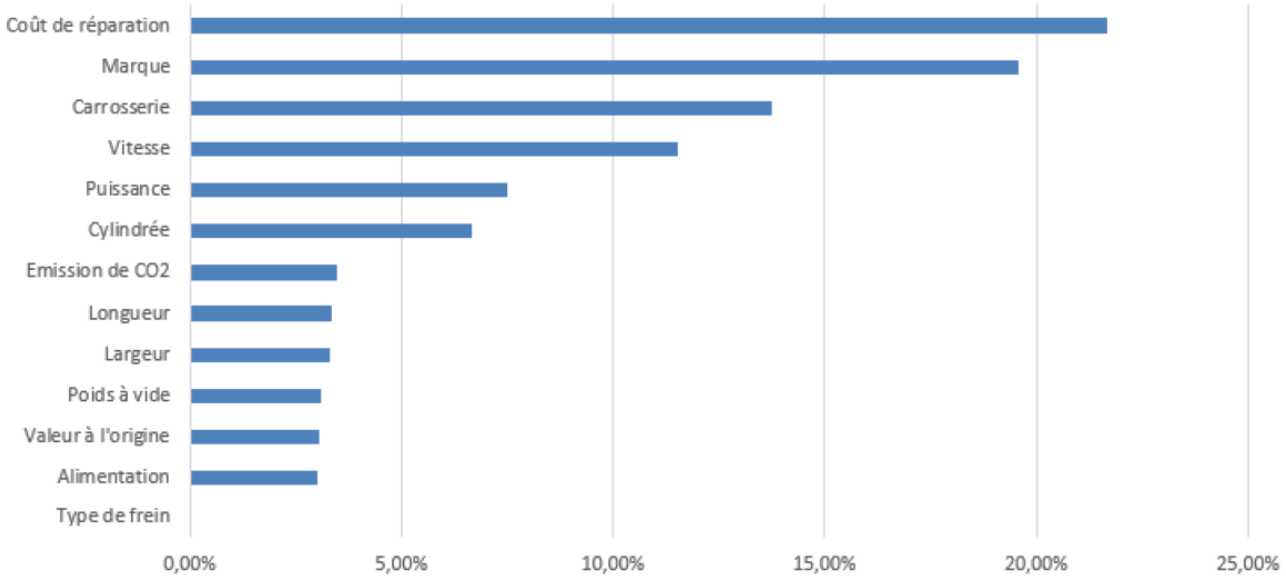


FIGURE 2.15 – Importance des variables - *XGBoost* sévérité

2.2.4 Comparaison des résultats

Les modèles *Random Forest* et *XGBoost* étant calibrés, il s'agit maintenant de déterminer lequel des algorithmes est le plus performant pour expliquer les résidus fréquence et les résidus

sévérité. L'évaluation sera faite en comparant les RMSE des modèles.

	Modèles	Critère	Train	Test
Fréquence	RF	RMSE	0,5201	0,5353
	XGBoost	RMSE	0,5226	0,5346

TABLE 2.6 – Évaluation des modèles Fréquence *Random Forest* et *XGBoost*

Les résultats de la base d'apprentissage sont meilleurs sur le modèle *Random Forest* que sur le modèle *XGBoost*, en effet le RMSE est plus faible. En revanche, les résultats du *Random Forest* sont moins bons sur la base test que pour le *XGBoost*.

	Modeles	Critère	Train	Test
Sévérité	RF	RMSE	31,0508	31,0107
	XGBoost	RMSE	31,0628	31,0375

TABLE 2.7 – Évaluation des modèles sévérité *Random Forest* et *XGBoost*

Les résultats du *Random Forest* sont meilleurs sur la base d'apprentissage et la base de test que le modèle *XGBoost*. Cependant, la performance n'est pas si éloignée des résultats du *XGBoost*.

Conclusion :

Malgré des algorithmes différents, le *Random Forest* et le *XGBoost* obtiennent des résultats satisfaisants. En effet, la validation croisée a permis de limiter le sur-apprentissage. A la vue de ces résultats, les scores obtenus par les modèles *Random Forest* sont repris pour la fréquence et la sévérité. Ce choix est conforté par le classement des variables contribuant le plus aux modèles. En effet, les variables ressortant comme les plus importantes dans les modèles semblent plus cohérentes d'un point de vue métier.

2.3 Cartographie des véhicules

La réalisation d'un véhiculier peut être comparée à la construction d'un zonier. En effet, dans les deux cas, il s'agit de créer un score à partir données externes (des données géo-démographiques dans le cas du zonier) et du résidu d'un GLM. Le contenu d'un véhiculier est en revanche plus difficile à appréhender que celui d'un zonier. Là où une simple carte de France servira de support pour représenter le zonier, il n'existe aucune carte pré-existante permettant de représenter les véhicules dans l'espace de façon simple. Il est toutefois nécessaire d'acquérir une bonne compréhension du contenu du véhiculier. Les méthodes d'analyses multi-dimensionnelles présentées dans cette partie peuvent aider en ce sens car elles permettent de créer une carte de véhicules.

2.3.1 Analyse en Composante Principale

L'ACP où l'analyse en composante principale a pour objectif de décrire la distribution de plusieurs variables quantitatives d'un jeu de données X à i individus et p variables :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ip} \end{pmatrix}$$

avec $u_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ vecteur ligne et $x_p = [x_{1j}, x_{2j}, ..., x_{ip}] \in \mathbb{R}^n$ vecteur colonne.

La description du jeu de données X se fait par le biais d'une réduction de l'espace de représentation des observations, c'est-à-dire en remplaçant les p variables par un nombre réduit de variables appelées composantes ou « factors ». L'ACP permet également de mieux visualiser les corrélations entre les variables et la proximité des variables.

La transformation des données quantitatives est essentielle. En effet, les variables n'ont pas la même unité de mesure. Afin de les rendre comparables et de leur accorder la même importance elles doivent être standardisées telles que :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

où $\bar{x}_j$ est la valeur moyenne de la variable considérée et σ_{x_j} l'écart-type de la variable, telles que :

$$\bar{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

et

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

Une fois standardisées, les données sont représentées par un point dans un espace de dimension P. Deux points proches, signifient que deux individus sont proches car leurs caractéristiques sont similaires.

Plusieurs notions sont importantes afin de comprendre la façon dont les individus sont représentés dans un plan à P dimensions : la distance euclidienne et l'inertie . La distance euclidienne est la métrique utilisée pour mesurer la distance entre deux individus i et i' :

$$d^2(i, i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

L'inertie du nuage de point traduit la quantité d'information disponible, il s'agit de la moyenne des distances euclidiennes entre chaque observation depuis le barycentre du nuage g. L'inertie s'exprime telle que :

$$I_p = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(i, i') \tag{2.3}$$

$$I_p = \frac{1}{n} \sum_{i=1}^{n} d^2(i, g) \tag{2.4}$$

Le coefficient de corrélation permet de mesurer la liaison entre deux variables X_j et X_m et se calcule de la manière suivante :

$$r_{jm} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m)}{\sigma_j \sigma_m}$$

Il est donc possible de construire un ensemble de composantes (F_1,F_2, ..., F_k) qui sont des combinaisons linéaires des variables initiales centrées réduites et de déterminer les coordonnées des individus.

La première composante principale, les coordonnées des observations et l'inertie s'expriment de la manière suivante :

$$F_1 = a_{11} z_1 + a_{12} z_2 + ... + a_{1p} z_p \tag{2.5}$$

$$\tag{2.6}$$

où Z_p correspond aux variables X_p standardisées.

En généralisant, il faut donc résoudre le système d'équations suivant :

$$\begin{cases} a_{11} z_1 + a_{21} z_2 + \cdots + a_{p1} z_p = F_1 \quad (\lambda_1) \\ \qquad\qquad \vdots \qquad\qquad\quad \vdots \quad \vdots \\ a_{1k} z_1 + a_{2k} z_2 + \cdots + a_{pk} z_p = F_K \quad (\lambda_k) \\ \qquad\qquad \vdots \qquad\qquad\quad \vdots \quad \vdots \\ a_{1p} z_1 + a_{2p} z_2 + \cdots + a_{pp} z_p = F_P \quad (\lambda_p) \end{cases}$$

avec a_1 les éléments du vecteur propre, autrement dit, les poids des combinaisons linéaires. La généralisation dans les différents plans factoriels, s'exprime donc de la manière suivante :

$$F_{ik} = a_{1k} z_{i1} + a_{2k} z_{i2} + ... + a_{pk} z_{ip}$$

donc

$$\lambda_k = \sum_{j=1}^{p} r_j^2(F_k, X_p)$$

où r^2 représente la qualité de représentation de la variable sur la composante k.

Une fois l'ACP réalisée chaque véhicule a des coordonnées associées. Le but de l'ACP sera de récupérer ces coordonnées qui seront utilisées par la suite dans la triangulation de Delaunay. Il est donc important de récupérer le maximum d'informations sur chaque axe afin d'avoir la représentation la plus fiable possible des véhicules.

L'histogramme des valeurs propres est tracé afin de visualiser la part de variance expliquée par chaque composante.

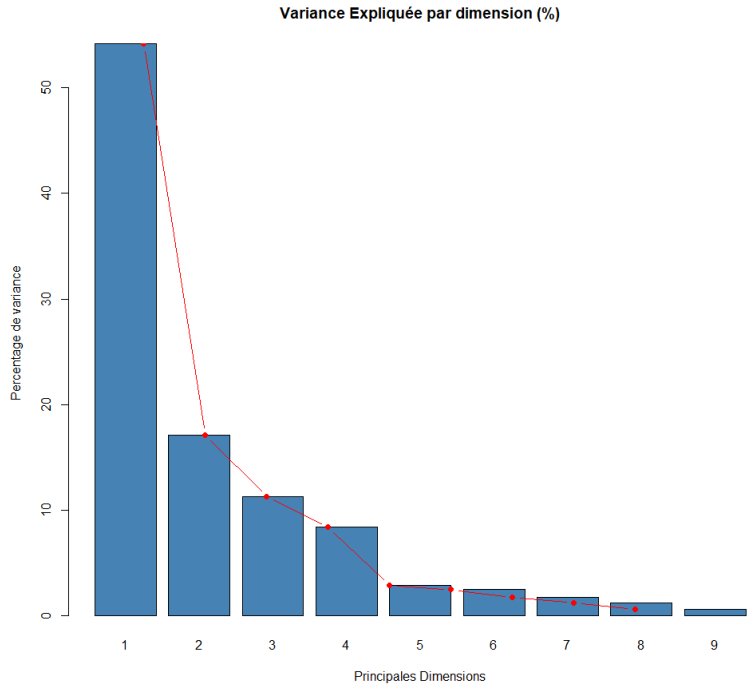


FIGURE 2.16 – Variance expliquée pour chaque composante

Les 4 premières composantes expliquent à elles seules près de 90% de l'information. Il s'agit maintenant de déterminer le nombre de dimensions à conserver à l'aide du « critère du coude ». Cette méthode consiste à conserver les valeurs propres jusqu'à un coude, le but étant de récupérer le maximum d'inertie. En revanche, il faut également prendre en compte les difficultés de représentation, car au-delà de 3 dimensions cela devient plus difficile. Aussi, seuls deux axes sont conservés, expliquant ainsi plus de 70% de l'information. Les véhicules seront donc représentés dans un plan à deux dimensions.

Le cercle des corrélations est tracé permettant une meilleure compréhension et une interprétation plus facile dans la suite de l'étude.

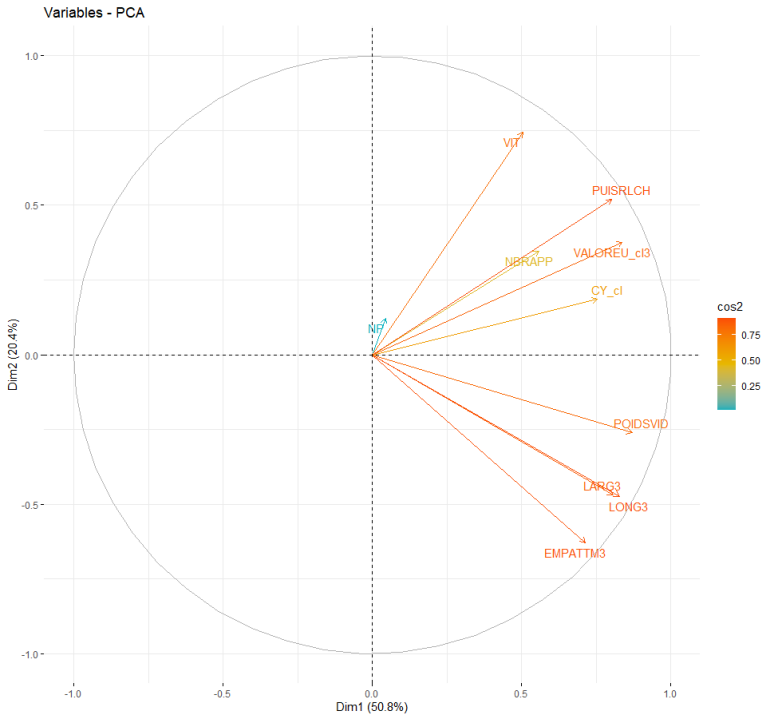


FIGURE 2.17 – Cercle des corrélations ACP

La qualité de représentation se mesure à la longueur de la flèche, plus elle est longue plus la variable est bien représentée et contribue à l'inertie des axes représentés. Ci-dessus, il est facile de voir que toutes les variables sont bien représentées sauf, la variable « NP » représentant le nombre de places du véhicule. Il est également possible de confirmer que l'axe 1 oppose les caractéristiques physiques des véhicules à leurs caractéristiques techniques. En effet, au-dessus de l'axe 1, la « puissance, la « valeur » sont représentées. Alors qu'en dessous de l'axe 1, il s'agit des variables « poids », « longueur » et « largeur ». Il est possible de conclure que plus la voiture est large, longue et lourde, plus le véhicule est cher et puissant. Ce phénomène avait déjà été constaté lors de l'analyse descriptive.

Il est possible de valider les déductions faites à partir du cercle des corrélations et dans l'analyse descriptive. A titre d'illustration, les véhicules sont représentés ci-dessous dans un plan à deux dimensions en fonction de leur puissance et de leur classe de prix.

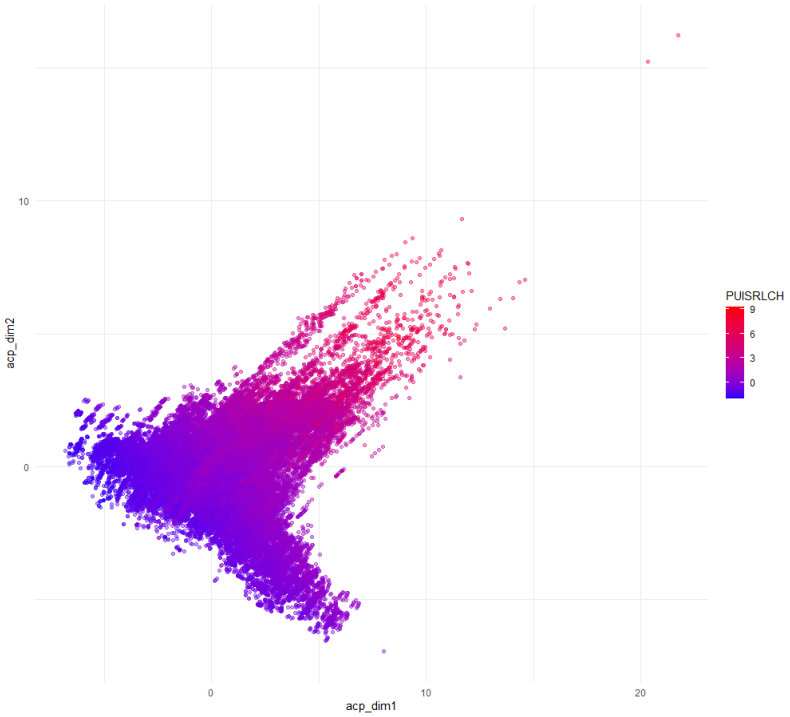


FIGURE 2.18 – Représentation des véhicules en fonction de la puissance sur les deux premières dimensions

Sur le graphique ci-dessus, il est possible de vérifier que plus un véhicule se situe sur la droite de l'axe des abscisses, plus ce dernier est puissant.

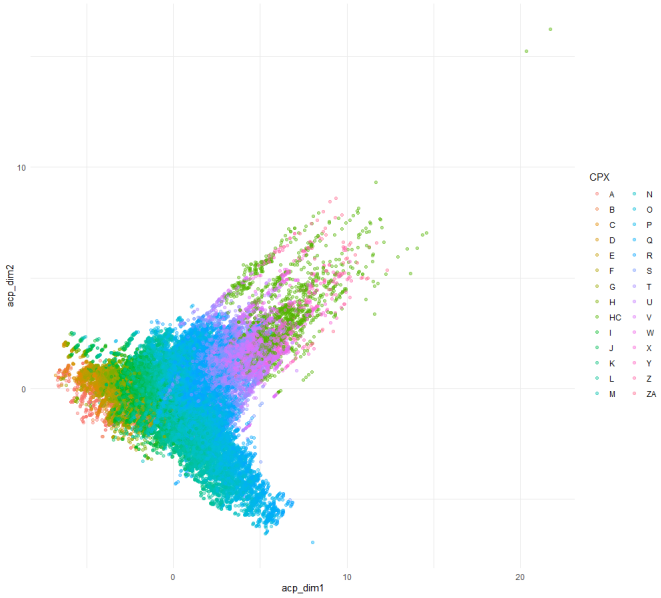


FIGURE 2.19 – Représentation des véhicules en fonction de la classe de prix sur les dimensions 1 et 2

La comparaison des deux graphiques montre clairement que les véhicules les moins chers sont les moins puissants et inversement.

2.3.2 Analyse Factorielle en Données Mixtes

L'AFDM ou Analyse Factorielle en Données Mixtes présente un double intérêt, celui de représenter les véhicules dans un plan orthogonal, mais également l'utilisation des données à la fois quantitatives et qualitatives. Ainsi, une analyse des relations entre ces deux typologies de variables est possible. En effet, l'analyse en composantes principales n'est utilisable qu'avec des données quantitatives, quant à l'analyse en composantes multiples elle n'est utilisable qu'avec des données qualitatives. L'AFDM fonctionne localement comme une ACP pour les variables continues et comme une ACM pour les variables catégorielles.

La première étape est la transformation des données quantitatives et qualitatives.

Pour les variables quantitatives, il n'y a aucun changement quant à l'ACP, elles sont standardisées. Les variables qualitatives sont transformées en indicatrices par un codage disjonctif tel que :

Individu	Var1	Var2
1	A	R
2	B	R
3	C	R
4	C	N

Var1_A	Var1_B	Var1_C	Var2_R	Var2_N
1	0	0	1	0
0	1	0	1	0
0	0	1	1	0
0	0	1	0	1

FIGURE 2.20 – Transformation des variables qualitatives en indicatrices

Il faut tout d'abord calculer le poids de chacune des modalités. Le poids dépend de la fréquence de la modalité et s'exprime de la manière suivante :

$$\omega_k = \frac{n_k}{n \times p}$$

avec n le nombre d'individus total (4 dans l'exemple ci-dessus), n_k le nombre de fois où l'indicatrice est égale à 1 dans la colonne k, pour tout $k \in \mathbb{N}$ (pour la première colonne ci-dessus : $n_1 = 1$), et avec p le nombre de variables égal à 1 (p = 2 pour la première ligne).

Une fois le poids des modalités calculé, il est possible de calculer la distance du χ^2 entre ces modalités. Plus une modalité est rare dans l'échantillon, plus elle est distante de l'origine. La distance du χ^2 s'exprime de la manière suivante :

$$d_2(k) = \frac{n}{n_k} - 1$$

Ainsi, de la même manière que pour l'ACP, l'inertie d'une modalité et l'inertie d'une variable s'expriment respectivement comme suit :

$$I(k) = \omega_k \times d^2(k)$$

et

$$I(j) = \sum_{k=1}^{m_j} d^2(k)$$

avec m_k une colonne après la transformation de la variable k en indicatrice ($M = \sum_{j=1}^{p} m_j$). La contribution d'une variable à l'inertie dépend du nombre de ses modalités. Moins une modalité est présente plus elle aura d'importance en termes d'inertie. Il est alors possible de calculer l'inertie totale pour les variables qualitatives :

$$I = \sum_{j=1}^{p} I(j)$$

Les valeurs propres s'expriment de la façon suivante :

$$\lambda_h = \sum_{k=1}^{M} \omega_k \times G_{kh}^2 \tag{2.7}$$

$$= \sum_{j=1}^{p} \sum_{k=1+\sum_{l=1}^{j-1}}^{\sum_{l=1}^{j}} \omega_k \times G_{kh}^2 \tag{2.8}$$

$$= \frac{1}{p} \sum_{j=1}^{p} \eta^2(F_h, X_j) \tag{2.9}$$

où l $\in \mathbb{N}$ et h $\in \mathbb{N}$, G_{kh} est égale à la moyenne de la modalité k pour le facteur h.

L'AFDM combine donc les méthodes de l'ACP et de l'ACM, elle permet de maximiser l'inertie apportée par les variables quantitatives et qualitatives telle que :

$$\lambda_h = \sum_{h=1}^{H_{max}} r^2(F_h, X_j) + \sum_{h=1}^{H_{max}} \eta^2(F_h, X_j)$$

Application :

Comme pour l'ACP, l'éboulis des valeurs propres est tracé afin de déterminer le nombre de dimensions à choisir. La technique du coude est utilisée aussi. Cette fois-ci, deux coudes sont constatés au niveau des 2ème et de la 4ème valeurs propres. Là où 90% de l'information était expliquée par les quatre premiers axes de l'ACP, dans ce cas ceux-ci n'expliquent plus que 9.7%.

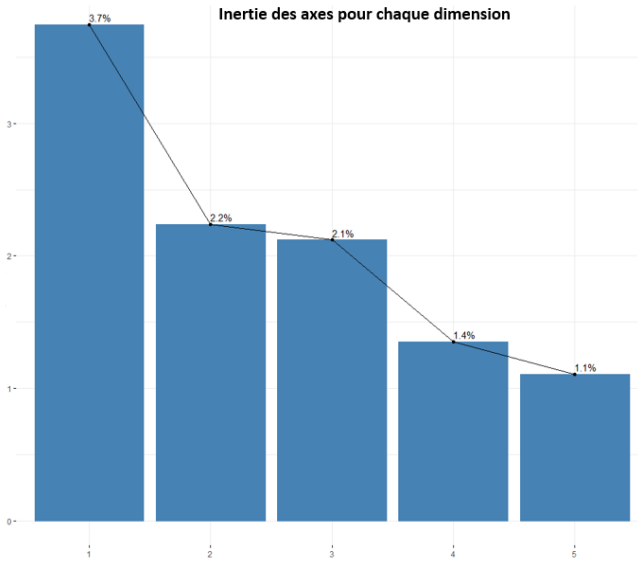


FIGURE 2.21 – Éboulis des valeurs propres de l'AFDM

Le choix aurait pu se porter sur deux dimensions, mais au vu du peu d'informations obtenues (seulement 5.9%), le choix se portera sur trois dimensions. Néanmoins, une très grande part de l'inertie reste encore à expliquer (92%). Par ailleurs, l'analyse des variables dans le plan factoriel ainsi que le tableau des contributions présenté en annexe (Figure 3.14 - Contribution des variables) ne permettent pas de faire ressortir quelles sont les variables qui contribuent le plus aux dimensions.

Afin de mieux mesurer l'impact de chacune des variables sur les trois premières dimensions et d'en tirer une meilleure compréhension, les contributions de chaque variable aux trois premiers axes sont tracées dans les graphiques ci-après.

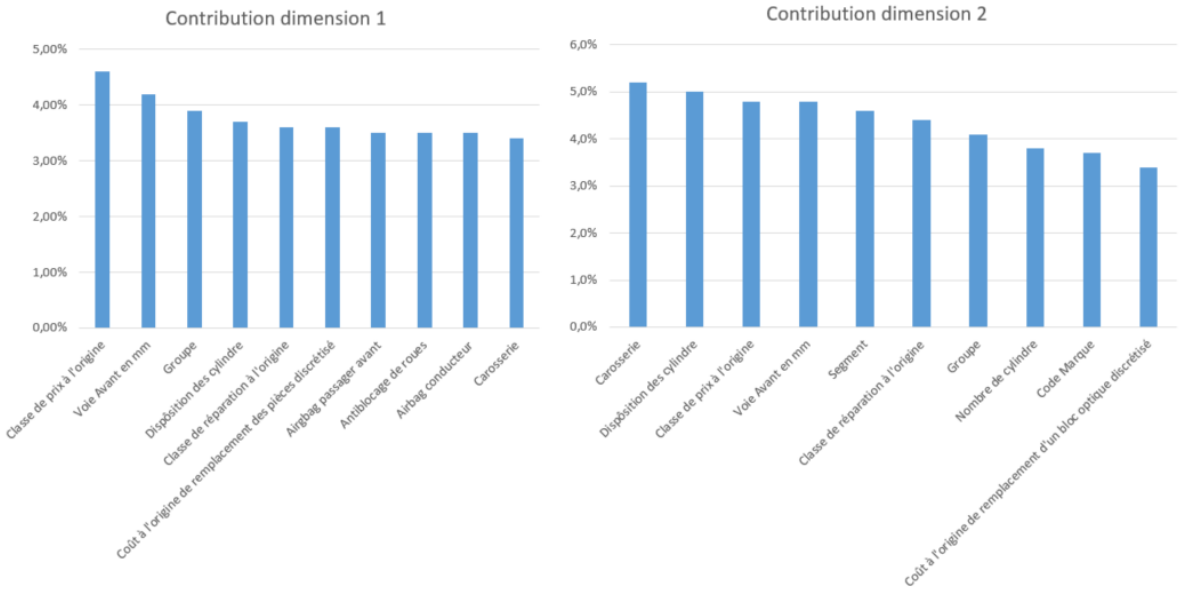


FIGURE 2.22 – Contribution des variables sur les deux premières dimensions

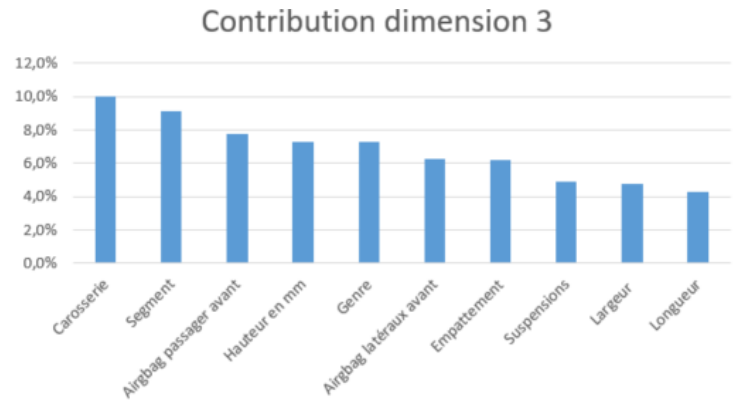


FIGURE 2.23 – Contribution des variables sur la troisième dimension

De manière globale, la classe de prix à l'origine est la variable qui contribue le plus aux deux axes. Vient ensuite la voie avant qui correspond à l'espacement entre les roues avant. Aucune véritable tendance ne ressort avec l'AFDM, contrairement à l'ACP. Le cercle des corrélations en annexe confirme cette difficulté d'interprétation.

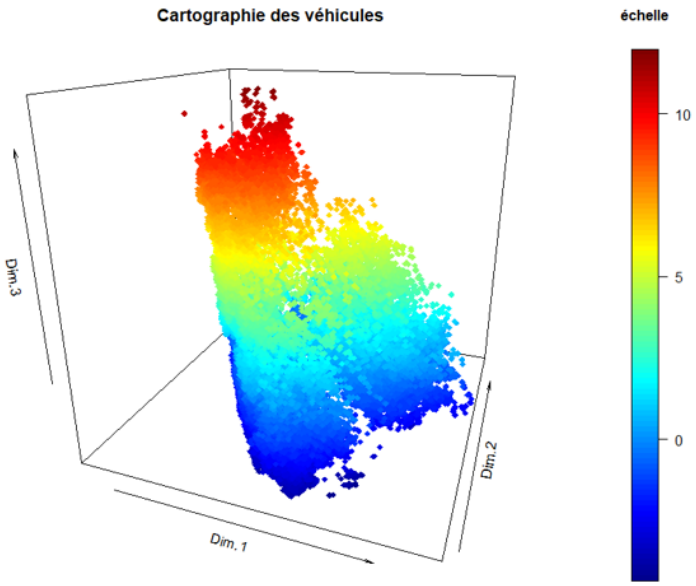


FIGURE 2.24 – Cartographie 3D des véhicules

Conclusion :

Après comparaison de ces deux méthodes, le nuage de points de l'ACP est finalement choisi pour déterminer les coordonnées des véhicules. En effet, contrairement à l'AFDM, les deux premières dimensions permettent d'expliquer 70% de l'information. La représentation des véhicules est donc plus fiable.

2.3.3 Triangulation de Delaunay

Chaque véhicule est associé à deux coordonnées notées (x,y) dans un plan. Il s'agit à présent d'identifier les points voisins. La triangulation permet de partitionner un espace en reliant des points entre eux. Ces liaisons permettent de localiser rapidement un individu dans cette espace. C'est la triangulation de Delaunay qui utilisée dans cette étude, elle permet de maximiser l'angle minimal des triangles et ainsi relier les points les plus proches. Elle s'applique dans un espace Euclidien à n dimensions qui sera noté $\mathbb{R}^n$ sur un ensemble de points $S = \{p_i, 1 \leq i \leq n\}$ (les projections orthogonales des véhicules dans cette étude).

La triangulation de Delaunay repose sur trois notions :

- la triangulation : ensemble d'arêtes qui relient les points de S sans intersection et qui est maximale (impossibilité de rajouter des arêtes) subdivisant ainsi l'espace ;
- le diagramme de Voronoï : il correspond à une partition d'un plan en plusieurs régions n, telle que chacune de ces régions $V(p_i)$ soit égale à l'ensemble des points du plan plus proche de p_i que de p_j avec $i \neq j$.

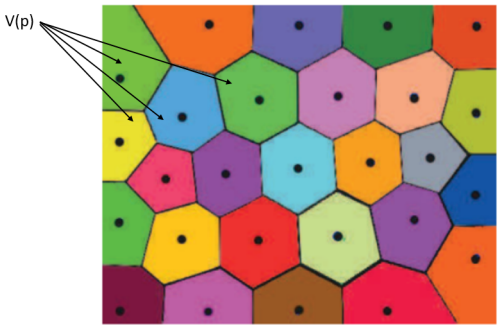


FIGURE 2.25 – Illustration d'un diagramme de Voronoï

- la décomposition de Delaunay : c'est le graphe dual du diagramme de Voronoï constitué d'un ensemble de triangles noté DT(S). La triangulation de Delaunay est formée de trois points notés p_i, p_j, p_k (i, j, k $\in \mathbb{N}$)de S qui sont les sommets des triangles. Les cercles circonscrits C(p_i, p_j, p_k) ne contiennent aucun point de S. Le centre du cercle est obtenu en traçant les médiatrices de chaque coté du triangle. La triangulation de Delaunay minimise les angles de chaque triangle et impose de n'avoir aucun autre point à l'intérieur du cercle. Ainsi, seuls les points les plus proches les uns des autres sont reliés.

  Les centres des cercles circonscrits sont représentés en rouge sur le graphique ci-dessous et sont trouvés grâce aux médiatrices des arêtes des triangles. Les sommets sont reliés grâce à la triangulation de Delaunay.

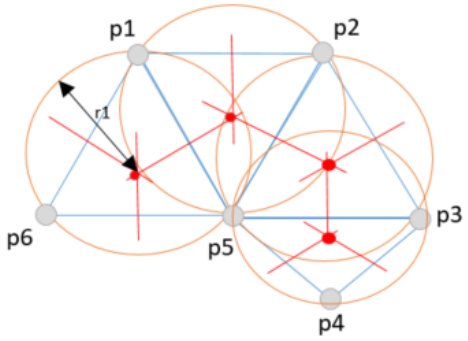


FIGURE 2.26 – Illustration de la triangulation de Delaunay

Une fois la triangulation de Delaunay appliquée, l'analyse des liaisons permet de trouver les véhicules les plus éloignés, comme les points entourés en rouge et en bleu ci-après. Conformément à ce qui a été constaté lors de la création de l'ACP, les points entourés en rouge sont des véhicules puissants à classe de prix élevée. Alors que celui entouré en bleu est petit, peu puissant, et à classe de prix moyenne.

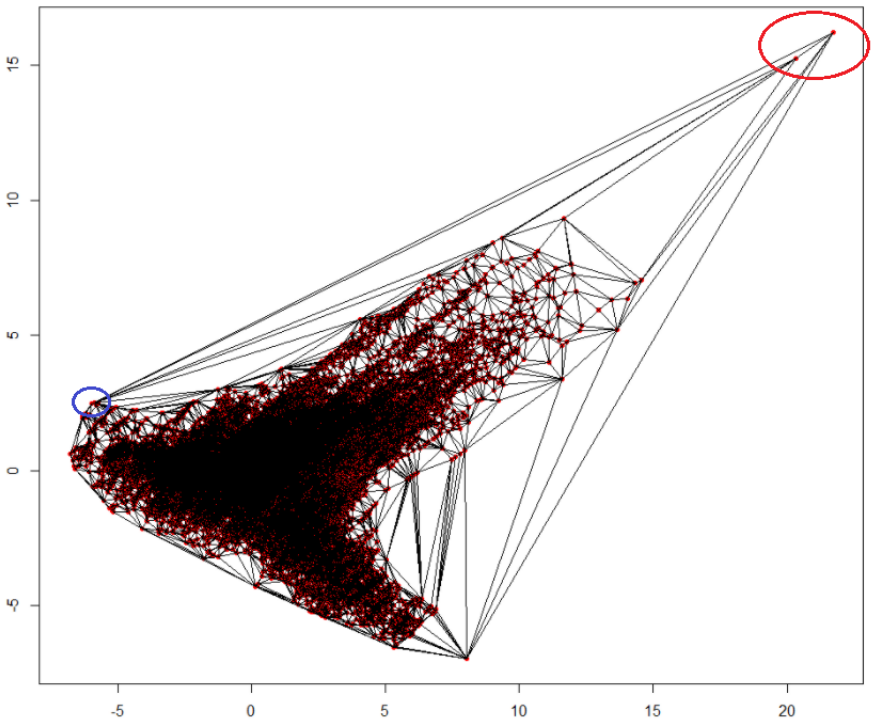


FIGURE 2.27 – Exemple de liaisons aberrantes

Les deux points rouges en haut à droite représentent deux véhicules de la marque Porsche. Le point bleu en bas à gauche représente quant à lui une Fiat. Ces liaisons sont créées puisque l'une des propriétés de la triangulation de Delaunay est d'être fermée. Certaines de ces liaisons peuvent être considérées comme aberrantes. Il s'agit maintenant de supprimer un maximum de

liens incohérents.

2.3.4 Création de la table d'adjacence des véhicules

Les véhicules sont à présent tous reliés à au moins deux autres véhicules. La première étape consiste à créer une table de voisinage. Cette table permet d'identifier les véhicules considérés comme voisins, afin de pouvoir analyser la nature des liens et éliminer par la suite ceux qui semblent incohérents.

Véhicule	Nombre de voisins	voisin_1	voisin_2	voisin_3	voisin_4	voisin_5	voisin_6	voisin_7	voisin_8	voisin_9	voisin_10
CI29004	5	CI29007	CI06001	FI19001	FI13025	AN09002					
CI29005	6	CI29007	CI29008	CI23005	FI12001	FI13025	IN02002				
CI29007	7	CI29004	CI29005	CI29008	FI19001	AB05001	FI13025				
CI29008	4	CI29005	CI29007	AB05001	FI12001						
BV01001	5	BV01002	RE36023	CI16002	AN10001	RE11003					
BV01002	5	BV01001	PA01002	PA01003	AN10001	RE11003					
BV01003	6	IA01007	IA01009	IA01008	CI10018	CI26001	TA02008				
CI06001	4	CI29004	FI19001	RE08002	CI01001						
FI19001	5	CI29004	CI29007	CI06001	AB05001	RE08002					
PA01002	7	BV01002	PA01003	RE36017	RE36020	FI13028	RE11003	CI39018			
PA01003	4	BV01002	PA01002	AN10001	RE36017						
RE45001	8	CI03005	CI16001	CI23005	FI20001	RE36009	RE36010	RE36011	RE36012		
RE45002	5	RE45003	HO02001	CI04182	PE18103	RE11001					
RE45003	4	RE45002	CI29001	HO02001	RE11001						
BV01004	8	BV01005	OP26011	SZ08017	IN10010	AU08001	MZ13005	PE07025	RE42088		
BV01005	5	BV01004	RE42099	RE42101	SZ08017	AU08001					
AB05001	6	CI29007	CI29008	FI19001	FI12001	RE08002	HO02001				
CI03004	5	CI03005	RE36001	CI23004	CI23006	AN10001					
CI03005	6	RE45001	CI03004	CI16001	RE36012	CI23004	AN10001				
CI16001	6	RE45001	CI03005	FI20001	HO06001	AN10001	RE36017				
CI23005	6	CI29005	RE45001	FI12001	FI20001	RE36011	IN02002				
CI29001	5	RE45003	CI01001	CI29010	HO02001	RE11001					
DF04001	6	FI12001	FI20001	HO06001	FI22001	AN02002	CI39018				
FI12001	8	CI29005	CI29008	AB05001	CI23005	DF04001	FI20001	HO02001	FI22001		
FI20001	6	RE45001	CI16001	CI23005	DF04001	FI12001	HO06001				
HO06001	5	CI16001	DF04001	FI20001	RE36017	CI39018					
PE21001	8	RE36002	RE36005	RE37005	OP07002	RE13001	IN02014	AO01001	AO01003		
RE08001	10	RE19001	RE36023	RE36026	RE36015	CI16002	CI10022	FI04017	FO06001	FO06002	FO07096

TABLE 2.8 – Extrait de la table d'adjacence de véhicules

Le premier véhicule « CI29004 » a cinq voisins, à savoir : « CI29007 », « CI06001 », « FI19001 », « FI13025 », « AN09002 ». Ces cinq véhicules ont aussi « CI29004 » comme voisin parmi d'autres voisins. Il est donc possible d'identifier chaque véhicule et ses voisins. Le nombre de véhicules voisins est en moyenne de 15 et au maximum de 29.

Afin de supprimer les liaisons aberrantes, les distances entre chaque véhicule voisin sont calculées et classées en vingtiles. La distance entre deux véhicules i et j, respectivement de coordonnées (x_i, y_i) et (x_j, y_j) sera notée :

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Le dernier vingtile contient les distances les plus grandes notamment la liaison entre les Porsche et la Fiat évoquée précédemment. Pour plus de cohérence ces liaisons sont supprimées. Le nouveau nombre de voisins moyen diminue donc et passe à 12 en moyenne et 22 maximum.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
0,0034	0,0051	0,0065	0,0078	0,0091	0,0103	0,0115	0,0128	0,0141	0,0155
Q11	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	**Q17**	**Q18**	**Q19**	**Q20**
0,0171	0,0189	0,0210	0,0235	0,0266	0,0310	0,0375	0,0489	0,0729	30,8124

TABLE 2.9 – Quantile de distance entre les véhicules

La liaison la plus grande entre les Porsche et la Fiat par exemple est donc supprimée. Les autres liaisons aberrantes également.

2.4 Lissage des résidus prédits

Une fois les liaisons optimisées, les résidus prédits par le *Random Forest* sont lissés. Cette homogénéisation des résidus permet de limiter les écarts entre des véhicules voisins et ainsi de réduire les sauts de classes au moment de la classification de ces derniers.

Le lissage consiste à prendre en compte à la fois l'estimation du résidu prédit de chaque véhicule r_i et la moyenne des résidus de ses voisins $\bar{r}_i$. Cela permettra de prendre en compte l'expérience du voisinage de chaque véhicule i. Pour cela, un facteur de crédibilité est calculé, ce dernier est noté Z_i. Plus l'exposition du voisinage est importante plus l'expérience du voisinage aura de poids.

Les poids utilisés correspondent à la distance par rapport au véhicule i. Ainsi chaque véhicule i aura un résidu prédit lissé par son voisinage.

Ce résidu prédit lissé sera noté :

$$R_i^* = r_i Z_i + (1 - Z_i)\bar{r}_i$$

avec :

- r_i qui correspond au résidu prédit par le *Random Forest* pour chaque véhicule i avec i $\in (1, ..., n)$;
- Z_i le poids de crédibilité définie de la manière suivante

$$Z_i = \frac{e_i}{e_i + k}$$

avec e_i l'exposition de chaque véhicule i et

$$k = \frac{\eta}{\alpha}$$

où η est la variance intra-classes et α est la variance interclasse ;
- $\bar{r}_i$ le résidu moyen au voisinage du véhicule i

$$\bar{r}_i = \frac{\sum_{i \neq j} r_j e_i d_{ij}^{-P}}{\sum_{i \neq j} e_i d_{ij}^{-P}}$$

Le coefficient P correspond au degré de lissage. Plus ce coefficient est élevé plus le lissage apporte de l'importance à la distance. C'est-à-dire que les véhicules les plus proches pèseront plus que des véhicules éloignés.

L'objectif est de réduire la variance intra-groupe de façon a obtenir une certaine homogénéité dans chaque groupe. Le fait de prendre en compte l'expérience du voisinage réduit les risques d'avoir des véhicules adjacents dont les prédictions seraient éloignées.

Le lissage est réalisé sur le score prédit en fréquence et en sévérité. Le graphique ci-après permet de visualiser l'effet du lissage sur les résidus fréquence. Les graphiques ci-après, sont un échantillon du score calculé avant et après lissage.

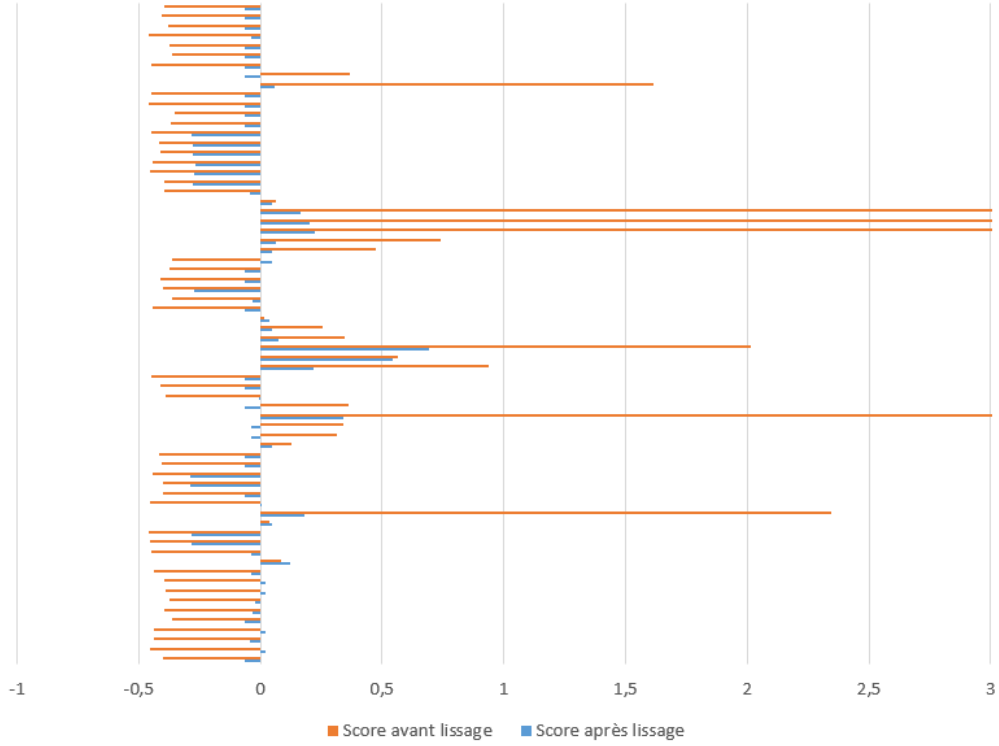


TABLE 2.10 – Échantillon des résidus avant et après lissage - fréquence

Le lissage a permis d'homogénéiser le score initial en tenant compte des voisins précédemment identifiés. Il est possible de remarquer un fort impact du lissage sur certains véhicules. Pour ces derniers, cela signifie que l'influence des voisins est très importante.

Le même effet est constaté sur les résidus du modèle de sévérité, comme le montre le graphique ci-après.

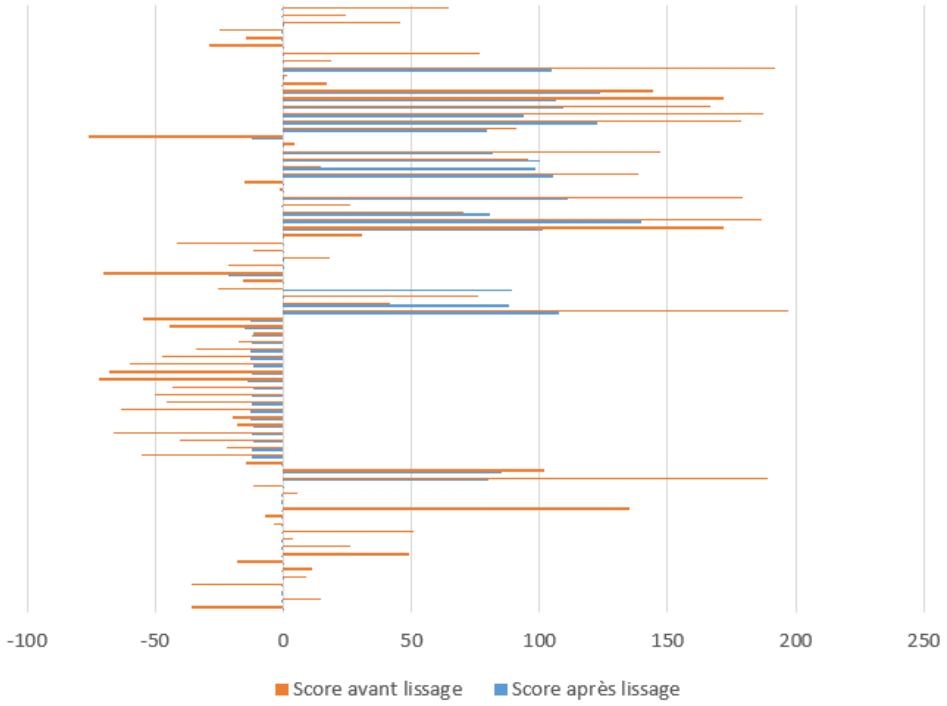


TABLE 2.11 – Échantillon des résidus avant et après lissage - sévérité

Chapitre 3

Création des classes et évaluation des véhiculiers

La première partie a détaillé les étapes de création d'un score à l'aide de méthodes de machine learning, ce dernier étant attribué à chaque véhicule compte tenu de ses caractéristiques. Dans la seconde partie un lissage a été réalisé afin de rendre ce score plus homogène entre des véhicules considérés comme similaires. Il s'agit à présent d'attribuer à chaque véhicule une classe de risque. Cette dernière partie présentera donc les méthodes de classification qui ont été utilisées pour y parvenir. Enfin, une fois introduits dans les modèles tarifaires, l'apport et l'impact des véhiculiers sont évalués et mesurés.

3.1 Création des classes

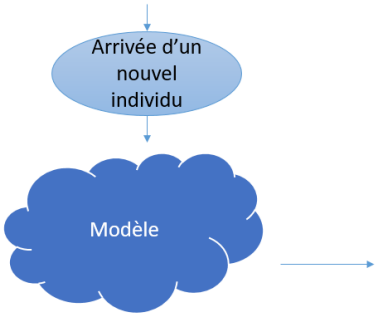
A ce stade de l'étude, un score lissé est à disposition pour chaque véhicule. Ce score a été construit à l'aide des données véhicules présentes dans la base SRA. L'objectif est maintenant d'agréger en plusieurs classes ces véhicules, en fonction de leur score de risque. Il s'agit donc, de discrétiser le score lissé précédemment obtenu, en un nombre fini de classes.

Parmi les nombreuses méthodes de classifications existantes, deux grandes typologies se dégagent : les méthodes de classification dites « supervisées » et celles dites « non-supervisées ».

La principale différence réside dans le fait que la première se base sur une donnée connue au préalable.

L'algorithme de classification supervisé permet d'établir un ensemble de règles de classification permettant d'attribuer une classe à tout nouvel individu. Ainsi, le label d'un nouvel individu peut être déterminé à l'aide de ces règles. C'est la classification supervisée qui est utilisée dans la première partie, puisque la valeur à prédire est connue. Pour rappel il s'agissait des résidus représentant l'effet véhicule extrait des modèles GLM. Le *Random Forest* et le *XGBoost* présentés en amont de cette étude sont deux illustrations performantes de ce type de classification.

Individus	Données 1	....	Données n	Prix
1	*Modalité 1*	...	*Modalité 1*	500€
2	*Modalité 2*	...	*Modalité 2*	1000€
3	*Modalité 3*	...	*Modalité 3*	1000€
4	*Modalité 4*	...	*Modalité 4*	250€
5	*Modalité 5*	...	*Modalité 5*	800€



Individus	Données 1	....	Données n	Prix
1	*Modalité 1*	...	*Modalité 1*	500€
2	*Modalité 2*	...	*Modalité 2*	1000€
3	*Modalité 3*	...	*Modalité 3*	1000€
4	*Modalité 4*	...	*Modalité 4*	250€
5	*Modalité 5*	...	*Modalité 5*	800€
6	*Modalité 6*	...	*Modalité 6*	**750€**

TABLE 3.1 – Illustration méthodologie classification supervisée

Dans l'illustration ci-dessus, le « label » est le prix. Un modèle est construit à l'aide des informations de chaque individu afin de pouvoir estimer le label d'un nouvel individu. Le prix peut donc être expliqué par les données de chaque individu.

La classification non supervisée vise au contraire à affecter une classe à chaque observation, cette dernière n'est pas connue au préalable. Celle-ci permettra de classifier les résidus prédits.

Il existe plusieurs méthodes de classification dont trois ont été testées, à savoir :

- la méthode des quantiles ;
- la méthode du poids égaux ;
- la méthode de classification à ascendance hiérarchique.

3.1.1 Quantiles

La base SRA est constituée d'un peu plus de 126 000 véhicules au moment de l'étude. La méthode des quantiles consiste à séparer les individus en k groupes de véhicules de même taille. Le quantile x_p d'ordre p, est la valeur qui découpe une série statistique ordonnée en plusieurs sous ensembles contenant chacun un nombre de véhicules égal n.

$\forall \quad i \in \mathbb{N}$ et $\forall \quad j \in \mathbb{N}$, la statistique $x_i : i = 1, ..., n$ donnant lieu à une série statistique ordonnée $x_j : j = 1, ..., k$. Le quantile x_p d'ordre p correspond à l'observation de rang [np] désignant le plus petit entier supérieur ou égale à np.

$$x_p = \frac{x_{(np)} + x_{(np+1)}}{2}$$

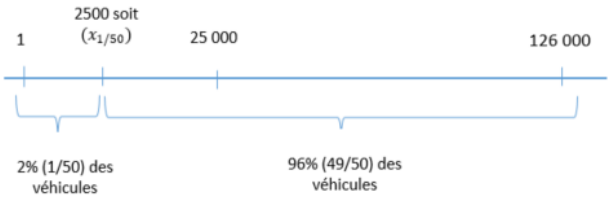


FIGURE 3.1 – Illustration - Découpage des quantiles

Cette méthode a l'avantage d'être facilement applicable. En revanche, le nombre de classes optimal pour créer le véhiculier est arbitraire. Aussi, plusieurs groupes seront testés au sein des modèles GLM. Les variables « classe de prix » ou encore le « groupe » présentes dans la base de données SRA contiennent respectivement 27 classes allant de A à ZA et de 26 groupes allant de 20 à 46. Lors de la création de la cartographie des véhicules, ces deux variables étaient celles expliquant le mieux l'inertie. Aussi il a été décidé de se fier à l'expertise métier de SRA et de faire au moins 30 classes pour chaque véhiculier.

La classification des véhicules à l'aide de la méthode des quantiles permet de séparer la base des véhicules en groupes contenant le même nombre de véhicules à partir du score obtenu par les *Random Forest*. Il est important de rappeler que la base SRA est à la maille véhicule et la base données servant à la tarification est à la maille image. Une séparation à la maille véhicule implique que l'exposition de chaque classe peut être différente, voir nulle.

Le graphique ci-après est une illustration de la problématique qu'entraîne la méthode des quantiles. Dans cet exemple, huit véhicules sont classés par ordre croissant de score. Plus le score est élevé, plus le véhicule est considéré comme ayant une espérance de sinistres plus élevée. Ainsi la dernière classe sera celle contenant les véhicules les plus risqués. Ici, 4 groupes contenant chacun deux véhicules sont créés. Chaque groupe contient donc 25% de la base des véhicules, mais ces derniers ne contiennent pas les mêmes proportions en termes d'exposition.

Division en 4 groupes de 2 véhicules

8 véhicules

Numéro	Véhicule	Quantiles sur les véhicules	Exposition	Exposition cumulée	Score
1	A	12,5%	1	25%	0,001
2	B	25%	0,75	44%	0,002
3	C	37,5%	0,25	50%	0,004
4	D	50%	1	75%	0,005
5	E	62,5%	0	75%	0,006
6	F	75%	0	75%	0,008
7	G	87,5%	0,5	88%	0,010
8	H	100%	0,5	100%	0,015

4 années polices

TABLE 3.2 – Illustration du principe de classification par quantiles

Dans le graphique ci-dessus, les véhicules 1 et 2 sont dans la classe 1 en rouge, les véhicules

3 et 4 sont dans la classe 2 en jaune, les véhicules 5 et 6 sont dans la classe 3 en vert, enfin les véhicules 7 et 8 sont dans la classe 4 en bleu.

Le groupe numéro 3 ne contient aucune exposition donc aucun contrat. L'exposition étant le poids du modèle de fréquence, l'estimation du paramètre β_k du groupe 3 ne sera pas robuste.

Afin d'être certain d'obtenir des coefficients robustes, des classes de véhicules contenant le même nombre d'expositions (respectivement de sinistres pour le coût moyen), la méthode des poids égaux est mise en œuvre.

3.1.2 Poids égaux

La méthodologie des poids égaux consiste à créer des groupes égaux non pas à partir du nombre de véhicules mais égaux en termes de poids. Il est donc nécessaire de créer de nouvelles classes contenant chacune le même poids en termes d'exposition ou de nombres de sinistres.

Dans l'illustration ci-après, 4 classes contenant chacune une année police sont créées par opposition aux quatre classes obtenues par la méthode des quantiles dans la partie précédente.

Division en 4 groupes de 1 année police chacun

8 véhicules

Numéro	Véhicule	Quantiles sur les véhicules	Exposition	Exposition cumulée	Score
1	A	12,5%	1	25%	0,001
2	B	25%	0,75	44%	0,002
3	C	37,5%	0,25	50%	0,004
4	D	50%	1	75%	0,005
5	E	62,5%	0	75%	0,006
6	F	75%	0	75%	0,008
7	G	87,5%	0,5	88%	0,010
8	H	100%	0,5	100%	0,015

TABLE 3.3 – Illustration du principe de classification par poids égaux

Dans le tableau ci-dessus chaque classe contient maintenant le même poids, avec au moins une exposition. La prédiction pour la classe 3 sera ainsi plus robuste qu'avec la méthode des quantiles. En effet, la classe 3 contenait alors 0 exposition avec la première méthode.

De la même manière qu'avec les quantiles le nombre de classes optimal n'est pas connu. Le nombre de classes sera supérieur à 30 également.

3.1.3 Classification ascendante hiérarchique

L'objectif est d'obtenir une hiérarchie dans les classes de véhicules. La première phase consiste à regrouper deux à deux chaque élément les plus proches en terme de distance. Une fois ces paires associées, elles sont à nouveau associées deux à deux, toujours grâce à la distance. Et ainsi de suite, jusqu'à n'avoir plus qu'une seule classe composée de toutes les paires.

Choix des critères de distance et d'agrégation :

Dans un premier temps il est nécessaire de définir la distance entre individus et une stratégie d'agrégation entre les groupes :

- la distance entre individus peut être définie ainsi (liste n'est pas exhaustive) :
  - distance euclidienne : Distance géométrique dans un espace multidimensionnel

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

  - distance euclidienne au carré :

$$d(x,y) = \sum_i (x_i - y_i)^2$$

- Les stratégies d'agrégation existantes sont :
  - Simple Linkage : les éléments étant les plus proches en termes de distance sont regroupés en deux classes ;
  - Complete linkage : les éléments étant les plus éloignés en termes de distance sont regroupés en deux classes ;
  - *Average* Linkage et critère de *Ward* : la distance entre deux groupes est associée à la distance entre les barycentres G_j de chaque groupe. Chaque agrégation a pour but de réduire l'inertie intra-classe. La méthode *Ward* maximise le gain d'inertie intra-classe à chaque agrégation et minimise l'inertie interclasse. Toutes les classes créées sont non vides et distinctes. A chaque agrégation, la perte d'inertie augmente du fait des regroupements. Deux groupes seront fusionnés si leur fusion correspond à la plus petite perte d'inertie. L'inertie totale se mesure selon le critère de *Ward*, celui-ci s'exprime de la manière suivante :

$$\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - X_G)^2 = \sum_{j=1}^{p}(X_{Gj} - X_G)^2 + \sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - X_{Gj})^2$$

La figure ci-dessous illustre les différences entre les trois critères de distance. Pour le critère de *Ward*, le barycentre est représenté en orange.

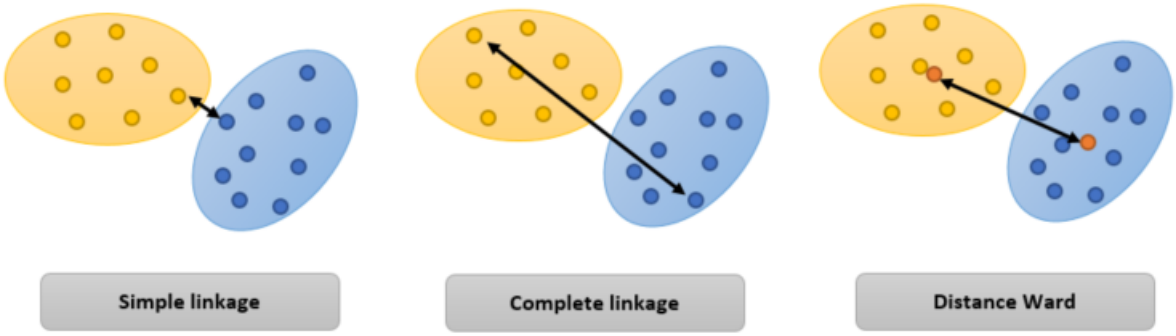


FIGURE 3.2 – Illustration des méthodes d'agrégation

Choix du nombre de classes optimales :

Lorsque les critères de distance et la stratégie d'agrégation sont choisis, il est possible de déterminer le nombre de classes optimal en traçant un dendrogramme et en représentant graphiquement la part d'inertie expliquée en fonction du nombre de groupes.

Le dendrogramme est une manière de visualiser les classes créées et la hauteur des branches représentent la proximité entre les classes. Il peut aider à déterminer de manière visuelle le nombre de classes optimal. Le nombre de classes est donné par la hauteur de la coupe. Plus la coupe est basse, plus le nombre de classes est élevé et la classification est fine. Plus la hauteur entre deux noeuds est grande plus la coupe est pertinente. Le graphique ci-dessus permet de comprendre le processus de création des groupes :

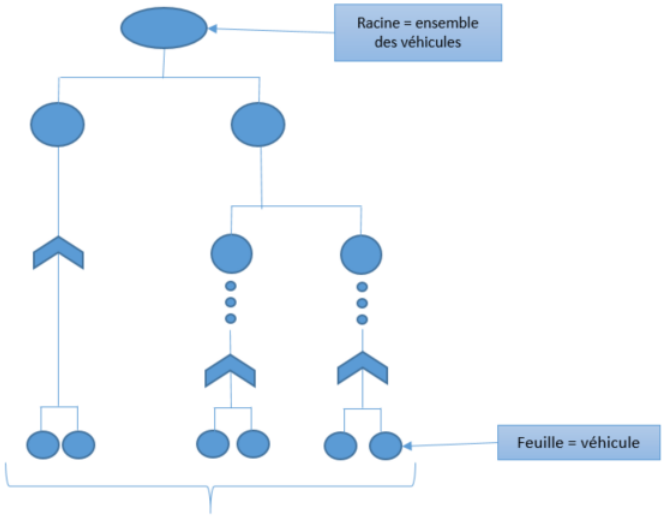


FIGURE 3.3 – Illustration du principe de hiérarchie

Cependant, avec un grand nombre d'individus, le dendrogramme peut vite s'avérer illisible rendant difficile la détermination du nombre de classes optimal. C'est pourquoi il est possible de tracer l'évolution de l'inertie totale à chaque agrégation. Le nombre de classes optimal sera alors déterminé par la méthode dite du « coude ». Une « cassure » importante indique une forte perte d'inertie ce qui signifie qu'un nombre de classes plus grand n'est pas optimal.

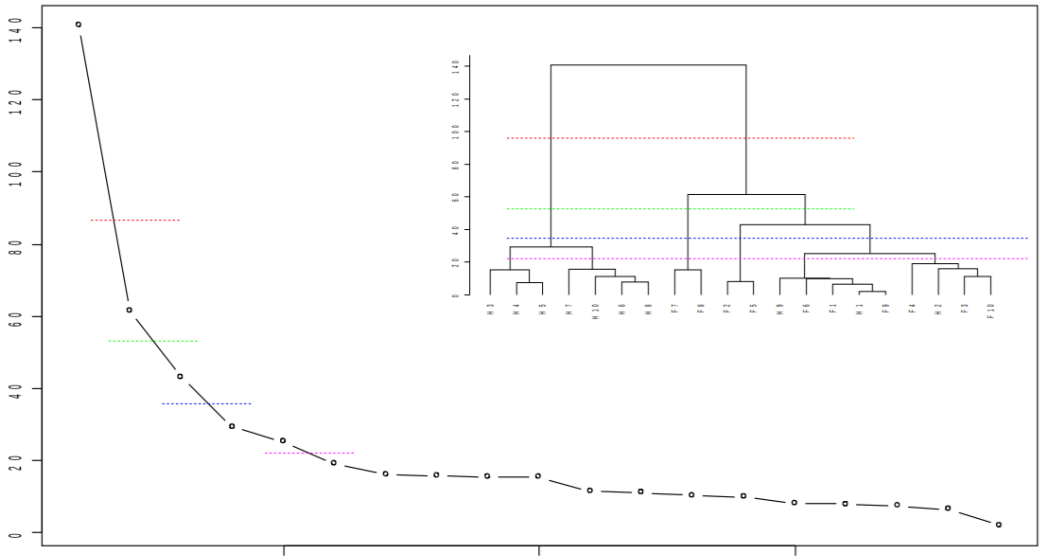


FIGURE 3.4 – Illustration de la perte d'inertie

Application :

Dans cette étude, la distance euclidienne et l'agrégation par le critère de *Ward* sont adoptées. Ces dernières semblent être les plus adéquates. Les dendrogrammes pour la fréquence et la sévérité sont tracés ci-après.

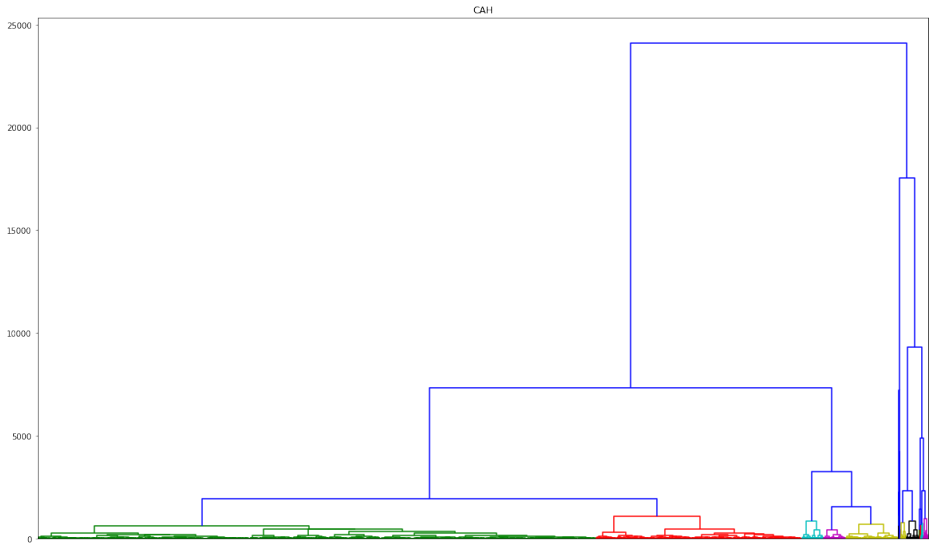


FIGURE 3.5 – Dendrogramme fréquence - CAH

La longueur des branches correspondant à la distance entre les groupes, le nombre de groupes optimal pour la fréquence serait deux groupes, ce qui est faible. De plus, les sauts semblent également importants pour les coupes trois et quatre. Un graphique représentant la perte d'inertie à chaque coupe est représenté ci-après afin de valider le nombre de classes optimal.

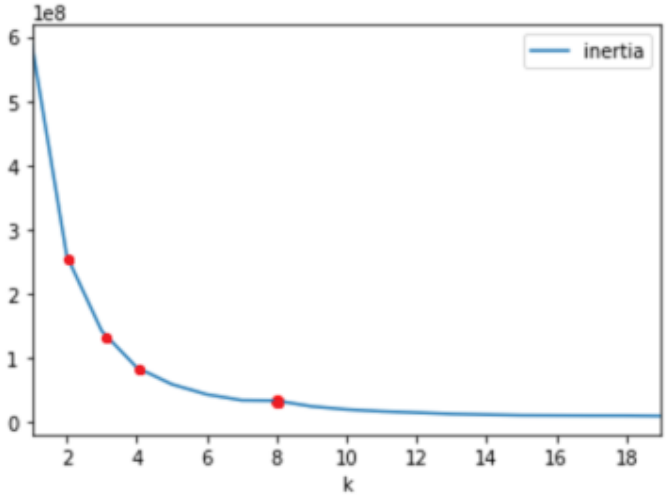


FIGURE 3.6 – Perte d'inertie fréquence

Des « cassures » sont bien présentes à 2, 3 et 4 classes. Une très légère perte d'inertie est observable à 8 classes. Mais l'inertie étant déjà très stable, la valeur ajoutée d'avoir plus de classes est très faible. Or, le but étant de segmenter au mieux le risque, le choix du nombre de classes doit être fait d'un point de vue objectif et respecter des attentes métiers. Aussi, à la vue du peu de classes obtenues avec la classification ascendante hiérarchique, cette classification n'est pas retenue pour être testée dans les modèles tarifaires fréquence. Il y avait cinq variables véhicules dans le modèle fréquence GLM. Ce qui permettait une segmentation plus fine, que ce que pourrait apporter l'utilisation des 4 classes obtenues à partie de la CAH.

La même étude est réalisée sur le score de sévérité :

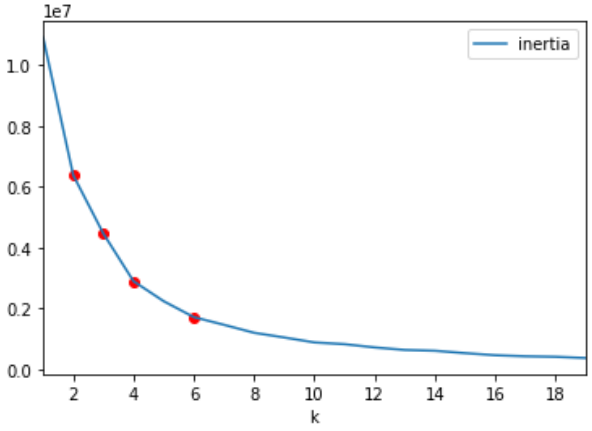


FIGURE 3.7 – Perte d'inertie sévérité

Pour la sévérité, les « cassures » s'observent aux classes 2, 3, 4 et 6. Tout comme pour la fréquence, le nombre de classes obtenu est considéré comme trop faible pour le véhiculier du modèle de sévérité, aussi cette méthode n'est également pas retenue.

Conclusion :

A la vue du faible nombre de groupes qui ressort de la méthode de classification ascendante hiérarchique, les méthodes des quantiles et des poids égaux ont été retenues pour être testées dans les modèles fréquence et sévérité.

Les deux premières méthodes de classification présentent l'avantage d'être facilement applicables. En revanche, connaître le nombre de classes optimal n'est possible qu'en testant les véhiculiers un à un dans les modèles tarifaires. La méthode *Ward* se base quant à elle sur l'inertie afin de déterminer le nombre de classes optimal, mais cette dernière ne permet pas d'obtenir suffisamment de classes d'un point de vu métier. A présent, chaque véhicule est associé à des classes dites « quantiles » et des classes dites du « poids égaux ».

3.2 Impacts dans les modèles tarifaires

Il s'agit désormais de déterminer si les véhiculiers créés sont pertinents en les introduisant un à un dans les modèles GLM d'origine à la place des variables véhicules des modèles initialement utilisés. Les modèles fréquence évoqués dans cette étude sont réalisés sous le logiciel Emblem. Le pouvoir prédictif et la qualité d'ajustement des nouveaux modèles sont évalués à l'aide de l'erreur quadratique moyenne et du coefficient de Gini.

3.2.1 Courbe de Lorenz et coefficient de Gini

Le coefficient de Gini est un indicateur permettant de mesurer la répartition d'une variable au sein d'un groupe d'individus. C'est-à-dire qu'il mesure les inégalités de répartition dans ce groupe d'individus. Il est possible de calculer ce coefficient à l'aide de la courbe de Lorenz.

La courbe de Lorenz est :

- une fonction définie sur l'intervalle [0,1] ;
- Elle est croissante et convexe sur ce même intervalle ;
- f(0) = 0 et f(1) = 1 ;
- Pour tout $x \in [0, 1], f(x) \leq x$.

Elle est représentée comme suit :

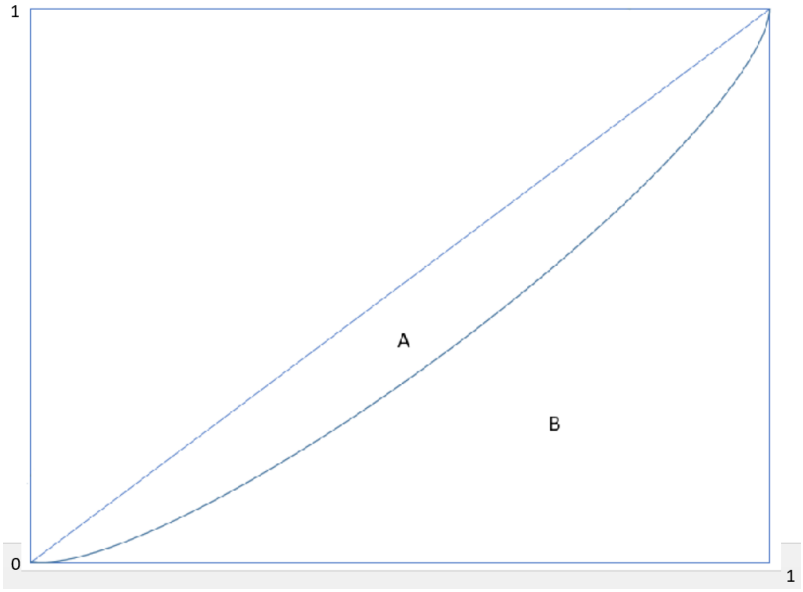


FIGURE 3.8 – Courbe de Lorenz et Coefficient de Gini

Plus la courbe de Lorenz est proche de la première bissectrice plus le modèle répartit les individus de manière égalitaire et inversement plus elle s'en éloigne, plus les individus sont repartis de manière inégalitaire. Une bonne segmentation permettant de limiter le phénomène d'antisélection, une maximisation du coefficient de Gini est recherchée.

Ce coefficient correspond à l'air entre la première bissectrice et la courbe de Lorenz, il est égal à :

$$\text{Gini} = \frac{A}{A+B}$$

Il est compris entre zéro (répartition égalitaire) et un (répartition inégalitaire). Plus il est proche de un, meilleur est le modèle dans le cadre d'une tarification.

3.2.2 Intégration des véhiculiers dans les modèles de fréquence

Les véhiculiers testés dans les modèles de fréquence et de sévérité sont les suivants :

Listes des modèles	Description
Modèle initial	Modèle référence contenant les variables véhicules
Modèle + véhiculier "Poids égaux" 50	Modèle sans les variables véhicule et avec le véhiculier de 50 classes créé avec la méthode des poids égaux
Modèle + véhiculier "Poids égaux" 40	Modèle sans les variables véhicule et avec le véhiculier de 40 classes créé avec la méthode des poids égaux
Modèle + véhiculier "Quantile" 30	Modèle sans les variables véhicule et avec le véhiculier de 30 classes créé avec la méthode des quantiles
Modèle + véhiculier "Quantile" 50	Modèle sans les variables véhicule et avec le véhiculier de 50 classes créé avec la méthode des quantiles
Modèle + véhiculier "Quantile" 80	Modèle sans les variables véhicule et avec le véhiculier de 80 classes créé avec la méthode des quantiles

TABLE 3.4 – Listing des modèles

Les tableaux ci-dessous illustrent les résultats obtenus suite à l'introduction des véhiculiers dans le modèle fréquence. Les indicateurs présentés précédemment permettent de comparer la performance de chacun des modèles en termes d'ajustement et de segmentation.

Modèle	Indicateurs	Modèle initial avec variables véhicules	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Fréquence – Base d'apprentissage	RMSE	0,14062	0,14066	0,14068	0,14067	0,14066	0,14066
	Gini	0,24360	0,23440	0,23380	0,23370	0,23790	0,23480

Modèle	Indicateurs	Modele initial avec variables véhicules	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Fréquence – Base Test	RMSE	0,14055	0,14057	0,14058	0,14057	0,14058	0,14057
	Gini	0,24680	0,23780	0,23700	0,23650	0,24040	0,23790

TABLE 3.5 – Évaluation des modèles de sévérité

L'introduction des véhiculiers dans le modèle de tarification de la fréquence dommage ne contribue pas à améliorer les résultats. Le RMSE est plus élevé pour tous les modèles avec véhiculier, ce qui indique que la qualité d'ajustement des modèles est meilleure sur le modèle initial que sur les modèles contenant les véhiculiers. Les résultats des véhiculiers sont très proches et aucune méthode de classification ne se dégage.

Le coefficient de Gini mesure le niveau de segmentation du modèle. Celui-ci permet de confirmer que le modèle initial est également meilleur en termes de segmentation puisque le coefficient de Gini est plus élevé sur le modèle initial que sur les modèles contenant le véhiculier.

Ces résultats sont confirmés sur la base de test.

Le graphique ci-après provient du logiciel Emblem et permet de visualiser l'impact des véhiculiers sur la modélisation de la fréquence et de la sévérité en termes de qualité d'estimation et de segmentation. La courbe rose représente la fréquence observée. La courbe vert foncé représente la fréquence prédite par le modèle compte tenu des autres variables du modèle. La courbe vert clair, représente les coefficients β_k pour chaque classe du véhiculier. Enfin, la courbe bleue représente la fréquence prédite du modèle initial (avec les variables véhicules libres). Les résidus sont plus élevés sur les petites classes (classes 1 à 9) et sur les grandes classes (36 à 40). Malgré une moins bonne qualité d'ajustement, le véhiculier est discriminant puisque l'amplitude des coefficients est élevée.

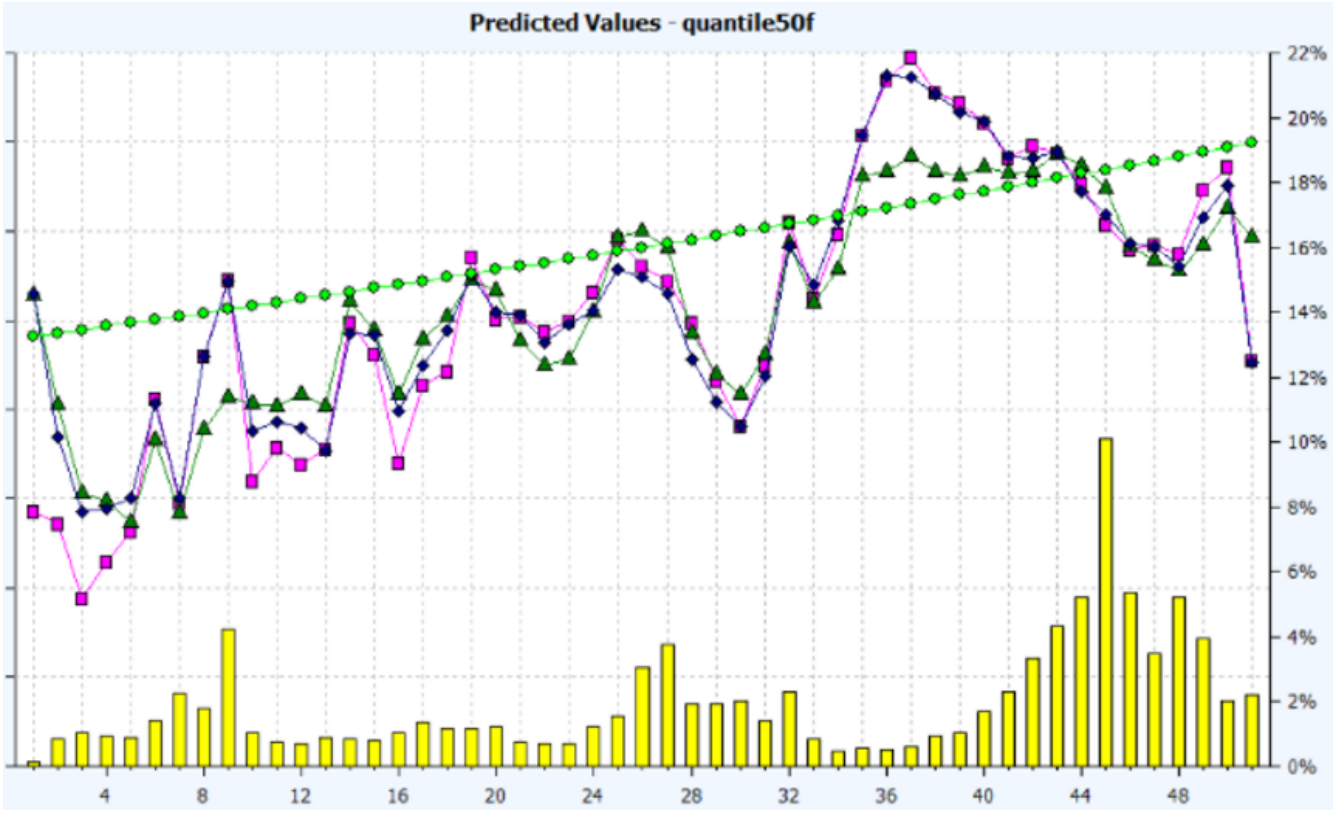


FIGURE 3.9 – Véhiculier - Fréquence

En conclusion, les véhiculiers créés pour les modèles fréquence ne peuvent remplacer les variables véhicules dans les modèles initiaux. En effet, les résultats ne s'améliorent pas une fois les véhiculiers introduits dans les modèles.

3.2.3 Intégration des véhiculiers dans les modèles de sévérité

Les véhiculiers testés dans les modèles de sévérités sont les mêmes que pour la fréquence.

Les tableaux ci-dessous illustrent les résultats obtenus suite à l'introduction des véhiculiers dans le modèle de sévérité. Les indicateurs présentés précédemment permettent de comparer la performance de chacun des modèles.

Modèle	Indicateurs	Modèle initial	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Sévérité - Apprentissage	RMSE	264,484	270,803	269,768	265,832	264,281	264,35
	Gini	0,148	0,1498	0,1495	0,1427	0,1482	0,1483

Modèle	Indicateurs	Modèle initial	Modèle + véhiculier "Poids égaux" 50	Modèle + véhiculier "Poids égaux" 40	Modèle + véhiculier "Quantile" 30	Modèle + véhiculier "Quantile" 50	Modèle + véhiculier "Quantile" 80
Sévérité - Test	RMSE	261,569	264,512	263,869	262,393	260,536	260,556
	Gini	0,1565	0,156	0,1552	0,1493	0,1535	0,153

TABLE 3.6 – Évaluation des modèles de sévérité

Contrairement aux modèles fréquence, une méthodologie se détache, puisque les modèles contenant les véhiculiers construits à partir des quantiles ont une qualité d'ajustement par rapport à la sévérité observée meilleure que le modèle initial. En effet, pour les véhiculiers avec 50 et 80 classes, les RMSE sont plus faibles que pour le modèle initial. Le véhiculier contenant 30 classes a un RMSE plus élevé que le modèle initial mais reste meilleur que les véhiculiers construits avec la méthode des poids égaux. Ces résultats sont confirmés sur la base de test.

De plus, le coefficient de Gini est meilleur sur tous les modèles contenant les véhiculiers, excepté pour le modèle à 30 classes et construit à l'aide des quantiles. Cependant, ce résultat n'est pas confirmé sur la base de test.

Les véhiculiers semblent pertinents sur les modèles de sévérité puisque la qualité d'ajustement des modèles est meilleure.

La courbe de Gini du modèle de sévérité est représentée ci-après :

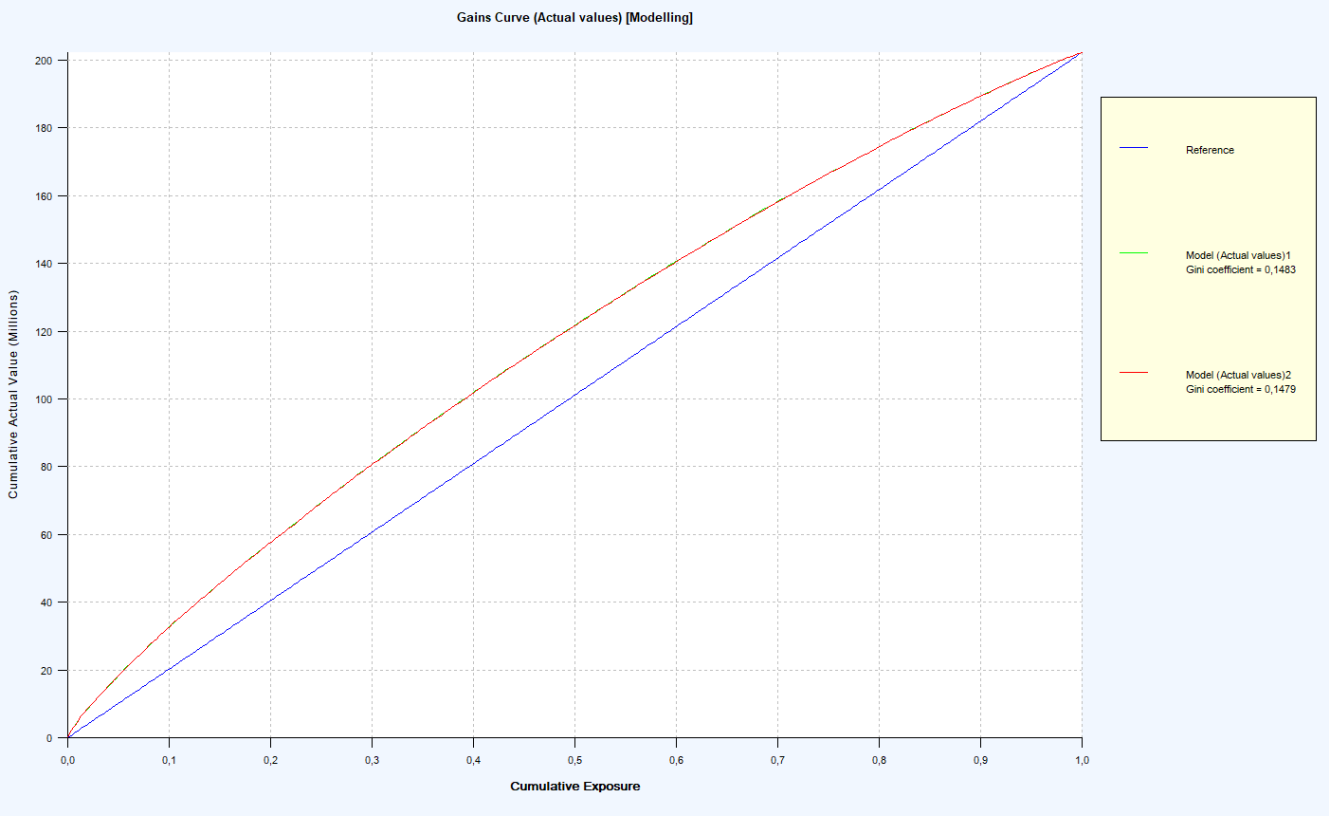


FIGURE 3.10 – Courbe de Lorenz du modèle sévérité y compris véhiculier « quantiles 50 »

Les coefficients de Gini étant très proches entre le modèle initial et le modèle contenant le véhiculier, les courbes de Lorenz sont presque superposées, ce qui signifie que le modèle initial et le modèle contenant le véhiculier répartissent de manière équivalente le portefeuille.

Le modèle sévérité « quantiles 50 » a la meilleure qualité d'ajustement (RMSE le plus faible). Afin pouvoir mesurer le niveau de segmentation de ce dernier, il est également important de regarder la tendance des coefficients du véhiculier. Ces coefficients doivent être croissants et présenter une amplitude suffisamment grande qui justifierait d'un fort pouvoir discriminant entre les classes.

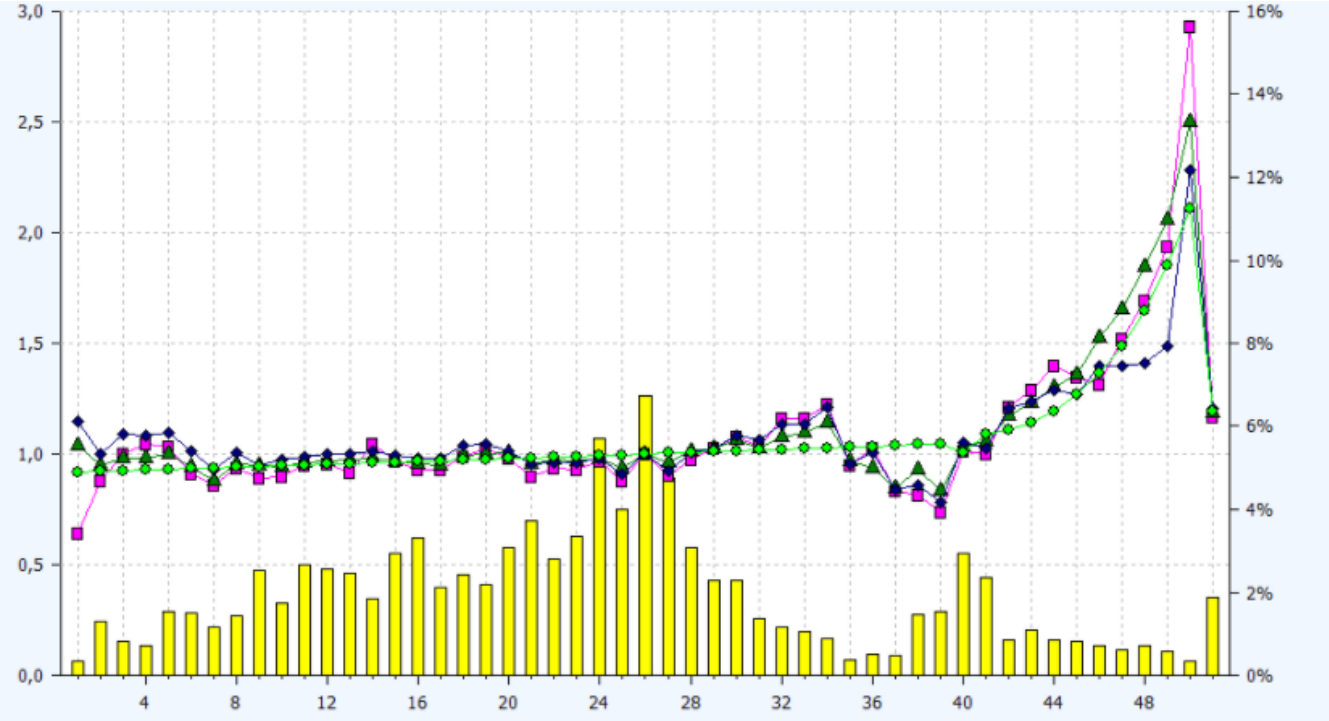


FIGURE 3.11 – Véhiculier - sévérité - apprentissage

L'écart entre la sévérité observée et la sévérité prédite dans le modèle initial est plus prononcé que l'écart avec le modèle de sévérité avec véhiculier, notamment sur les véhicules à classes très faibles ou très élevées. Ce qui signifie que le modèle avec véhiculier est bien meilleur que le modèle initial, mais il est possible de voir que cette amélioration est limitée. De plus, comme attendu, plus la classe de risque du véhicule augmente, plus les coefficients sont élevés.

L'amélioration de la prédiction ne suffit pas pour intégrer le véhiculier dans les modèles tarifaires finaux. En effet, pour des raisons métier et dans une logique de distinction, l'amplitude des coefficients doit être suffisante. L'amplitude des coefficients est calculée comme suit :

$$\text{Amplitude} = \frac{\beta_p}{\beta_1} - 1$$

avec p le nombre de modalité (ici p = 50).

Spread modèle sévérité	
Spread classes 1 à 40	0,09
Spread classes >40	1,49

FIGURE 3.12 – Véhiculier - sévérité - test

L'amplitude des coefficients est faible sur les 40 premiers groupes (+9%), ce qui signifie que

le pouvoir discriminant est faible. En revanche, l'amplitude des coefficients des classes 41 à 50 est très élevée. Le véhiculier semble donc bien fonctionner sur les véhicules à coût moyen élevé.

Cette faible amplitude est confirmée sur la base de test.

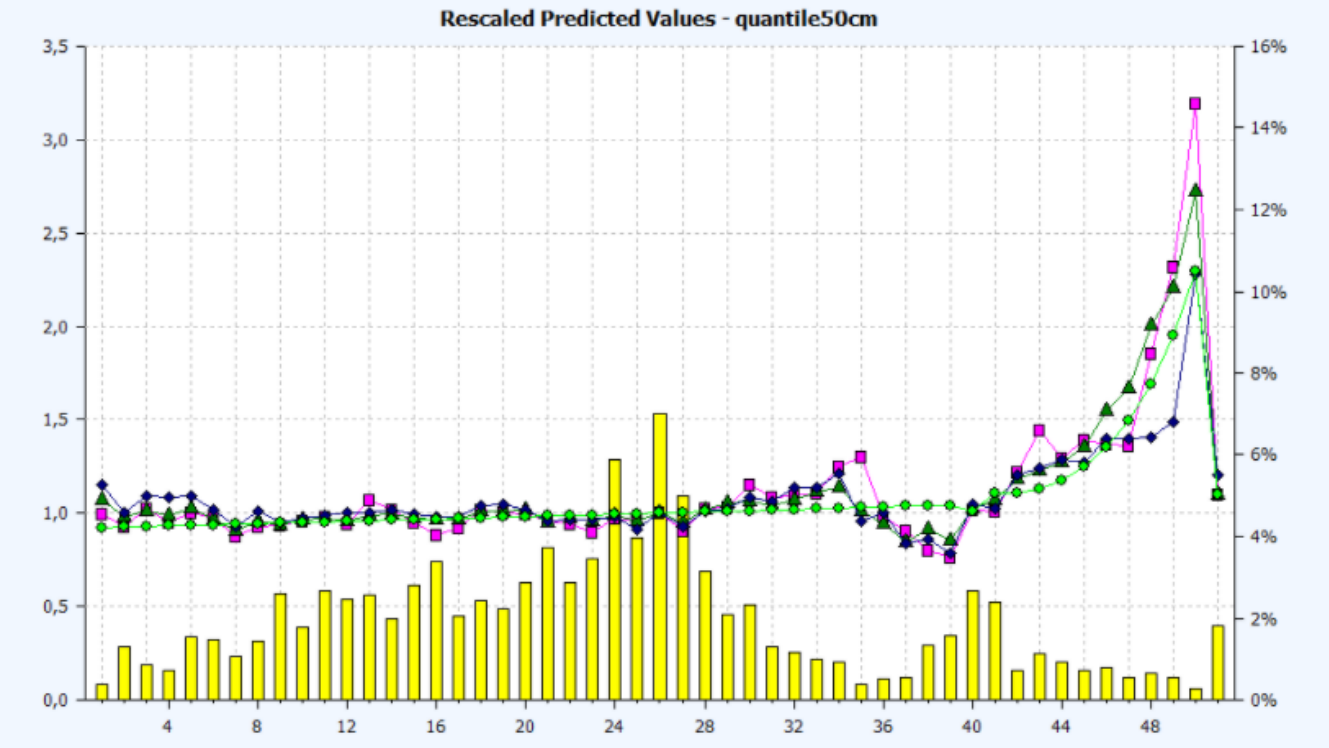


TABLE 3.7 – Spread véhiculier sévérité

Le dernier point regroupe les véhicules non classés par SRA, le travail réalisé sur ces véhicules ne sera pas présenté dans cette étude.

Conclusion

Le véhiculier n'apporte pas d'amélioration sur le modèle fréquence dommage puisque tous les indicateurs sont moins bons sur les modèles incluant le véhiculier (sur la base d'apprentissage et la base de test). En revanche, les résultats ne sont pas très éloignés du modèle initial. Ceci peut s'expliquer par le fait que la fréquence dépend plus de l'aspect conducteur et du zonier que de l'effet véhicule.

En revanche, sur le modèle de sévérité le véhiculier semble plus pertinent puisque les résultats sont meilleurs sur la base d'apprentissage et sur la base de test (en termes de RMSE, et de Gini). Ces résultats sont cependant à nuancer, puisque l'amélioration n'est pas significative et les résultats de la déviance sont moins bons sur la base de test que le modèle initial. De plus, l'amplitude des coefficients confirme que l'impact tarifaire entre chaque classe est trop faible et ne permet pas une segmentation optimale, excepté sur les classes élevées où l'amplitude est très forte.

Enfin, les résultats mitigés peuvent être expliqués par la période observée. En effet, l'une des contraintes de cette étude fut de devoir utiliser les données du portefeuille 2013 à 2017. La

sophistication des véhicules croissant avec le temps, ce sont les voitures récentes qui bénéficient des nouveautés en termes de sécurité. Celles-ci étant moins présentes sur les vieux véhicules, les données disponibles concernant ces options sont peu exploitables, du fait d'un grand nombre de valeurs manquantes. La mise à jour de cette étude sur des années plus récentes permettrait de tenir compte des derniers modèles de véhicules et donc de pouvoir augmenter le nombre de variables dans les modèles *Random Forest* et *XGBoost* et donc d'améliorer à terme la classification.

Conclusion

De nouveaux axes d'amélioration des modèles tarifaires sont essentiels aux assureurs pour réussir à sélectionner correctement leurs risques afin d'éviter l'anti-sélection. Ce besoin d'innovation incite les assureurs à tester de nouvelles méthodes et à remettre en question leurs approches tarifaires. C'est dans ce contexte que ce mémoire a été réalisé. Cette étude s'est attachée à vérifier si l'introduction de nouvelles informations véhicules pouvait être appropriée au modèle dommage. Les Modèles Linéaires Généralisés ne permettant qu'un nombre limité de variables explicatives, il était pertinent de considérer l'utilisation d'un score véhicules, plutôt que d'introduire une multitude de nouvelles variables. D'où le recours à un véhiculier.

La création d'un score véhicule à l'aide des méthodes non paramétriques telles que le *Random Forest* ou le *XGBoost*, a permis d'introduire dans la tarification un ensemble de variables véhicules auparavant inexploitées, notamment les variables continues comme la valeur à l'origine du véhicule, le poids, ou la puissance. Ces variables sont plus fines que les variables traditionnellement utilisées (telles que la classe de prix pour la valeur du véhicule, ou le poids à vide et la puissance pour le groupe). En effet, le logiciel de tarification contraint le nombre de modalités maximum par variable à 350, ce qui se révèle insuffisant compte tenu du potentiel d'informations qu'apportent certaines variables continues. En revanche, ces méthodes sont moins lisibles qu'un simple Modèle Linéaire Généralisé.

Cette étude a permis de démontrer en partie l'intérêt du véhiculier pour la garantie dommage, notamment sur le modèle de sévérité. En revanche, l'amélioration des résultats n'est que limitée et l'amplitude des coefficients est plus significative pour les grands groupes. Sur le modèle de fréquence, l'ajout du véhiculier ne permet pas d'améliorer les résultats par rapport au modèle initial. Ceci pourrait être expliqué par le fait que la fréquence de sinistres dépend plus de l'aspect conducteur que de l'aspect véhicule, contrairement au modèle de sévérité qui dépend fortement de la typologie du véhicule assuré.

Plusieurs éléments peuvent expliquer ces résultats et limiter l'impact du véhiculier sur la tarification :

Tout d'abord, la période d'observation (2013-2017) et l'âge moyen élevé des véhicules présents en portefeuille limitent l'utilisation des nouvelles données véhicules (aides à la conduite). En effet, ces informations sont présentes uniquement sur les véhicules les plus récents, ce qui implique qu'elles ne puissent pas être prises en compte dans cette étude. L'arrivée de ces nouvelles données pourrait accroître l'impact du véhiculier notamment sur le modèle de sévérité.

Ensuite, la triangulation de Delaunay est très dépendante du jeu de points qu'elle utilise. Le faible pourcentage d'inertie expliqué sur les premiers axes de l'AFDM ne rendait pas l'utilisation de cette dernière possible, c'est donc l'ACP qui a été privilégiée. L'utilisation des coordonnées de l'AFDM avec plus de trois dimensions permettrait d'augmenter la part d'inertie expliquée, d'augmenter le nombre de caractéristiques prises en compte et ainsi d'améliorer la représentation des véhicules et donc l'identification des véhicules considérés comme voisins.

Enfin, l'approche résiduelle a été choisie dans cette étude, mais il est possible de tester

une approche basée uniquement sur les caractéristiques des véhicules en créant un score à partir de méthodes de classification non supervisées, comme les *K-means*. Ces modèles permettent de regrouper des individus suivant des caractéristiques, tout en réduisant la variance intra-classe et en maximisant la variance interclasse. Cependant, l'interprétation des classes est souvent difficile.

Bibliographie

[1] ARBOGAST E. MAHUZIER A.(2014) *Performance de la crédibilité, Towers Watson*

[2] BOYER C., SANGNIER M. (2020) *Supervised learning, Non-paramétric methods, Sorbonne Université*

[3] Christophe L. (2013) *Triangulation de Delaunay et arbres multidimensionnels Synthèse d'image et réalité virtuelle , Ecole Nationale Supérieure des Mines de Saint-Etienne ; Université Jean Monnet - Saint-Etienne*

[4] FFA (2021) *L'assurance française - données clés 2020*

[5] GENUER R., POGGI J-M. (2017) *Arbres CART et Forêts aléatoires, Importance et sélection de variables, ISEP Université de Bordeaux, LMO Université Paris-Sud Orsay*

[6] HUSSON F. *Classification ascendante hiérarchique, Agrocampus Rennes*

[7] IWASZKO T. (2014) *Généralisation du diagramme de Voronoï et placement de formes géométriques complexes dans un nuage de points, HAL archives-ouvertes.fr*

[8] KRANZLIN S. (2017) *Modélisation du risque géographique en assurance automobile*

[9] LAVENU J. (2016) *Les méthodes de Machine Learning peuvent-elles être plus performantes que lávis déxperts pour classer les véhicules par risque homogène ?*

[10] McCUllagh P., Nelder J.A. FRS (1989) *Generalized linear models second edition*, p-48

[11] PALMIROTTA G. (2014) *Triangulation de Delaunay, Université de Luxembourg*

[12] PLANCHET F., MISERAY A. (2017) *Tarification : Introduction aux techniques avancés*

[13] RAKOTOMALA R. *Classification Ascendante Hiérarchique, Université Lyon Lumière 2*

[14] RAKOTOMALA R. *Analyse des Correspondances Multiples - ACM, Université Lyon Lumière*

[15] RAKOTOMALA R. *Analyse en Composantes Principales - ACP Université Lyon Lumière 2*

[16] RAKOTOMALA R. *Analyse Factorielle en Données Mixtes - AFDM, Université Lyon Lumière 2*

[17] RUIMY M. (2017) *Elaboration dún véhiculier en assurance automobile 2017*

[18] SEPULVEDA C. (2016) *Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier, Comparaison de deux méthodes de lissage spatial , (consulté en 2021)*

[19] https ://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/, *(consulté en 2021)*

[20] https ://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost, *(consulté en 2021)*

[21] *http : //www.itse.be/statistique*2010/*co*/22141$_{C}ours_{q}uantile_{1}$.*html* , *(consulté en 2021)*

[22] https ://www.sra.asso.fr/informations-vehicules/automobiles/recherche, *(consulté en 2021)*

[23] https ://xgboost.readthedocs.io/en/latest/tutorials/model.html *(Introduction to Boosted Trees), consulté le 15 Avril 2021*

Annexes

Variables	Typologie de variables	Libellées variables	Contribution Dimension 1	Contribution Dimension 2	Contribution Dimension 3
VIT	Variables quantitatives	Vitesse maximum (en Km/h)	2,2%	1,6%	3,4%
NP		Nombre de places	0,0%	0,0%	0,7%
NBRAPP		Nombre de rapports	3,0%	0,1%	0,2%
CY_cl		CY transformé en numérique	1,8%	2,4%	0,4%
LARG3		Largeur	2,2%	0,4%	4,8%
EMPATTM3		Empattement	1,4%	0,1%	6,2%
LONG3		Longueur	2,3%	0,4%	4,3%
POIDSVID		Poids à vide (en Kg)	3,0%	1,1%	3,1%
PUISRLCH		Puissance DIN (en CV DIN)	3,1%	3,3%	0,2%
VALOREU_cl3		Valeur à l'origine	3,1%	3,4%	0,0%
MRQ	Variables qualitatives	Code Marque	1,7%	3,7%	2,6%
CAR		Carrosserie	3,4%	5,2%	10,0%
ALI		Alimentation	3,2%	2,1%	0,5%
ENE		Energie	0,7%	0,9%	1,4%
GRE		Genre	0,0%	0,3%	7,3%
SER		Série limitée	0,1%	0,4%	0,2%
NBCYL		Nombre de cylindres	0,7%	3,8%	0,8%
DCY		Disposition des cylindres	3,7%	5,0%	0,2%
PMO		Position du moteur	0,0%	0,2%	0,0%
TRA		Transmission	1,0%	3,3%	0,3%
BDV		Boîte de vitesses	1,0%	1,6%	0,6%
SUS		Suspension	0,4%	1,5%	4,9%
TFR		Type de freins	3,4%	0,1%	0,0%
AFR		Assistance de freinage	0,3%	0,2%	0,0%
ACV		Airbag conducteur	3,5%	2,9%	0,4%
APV		Airbag passager avant	3,5%	3,1%	7,8%
ALV		Airbags latéraux avant	0,1%	1,1%	6,3%
ALR		Airbags latéraux arrière	0,0%	2,1%	3,0%
DAS		Direction assistée	2,7%	1,5%	0,1%
groupe		Groupe	3,9%	4,1%	1,3%
CPX		Classe de prix à l'origine	4,6%	4,8%	1,1%
groupa		Groupe APSAD	6,7%	3,3%	0,6%
AFU		Assistance de freinage d'urgence	0,1%	0,8%	0,7%
ABR		Antiblocage de roues	3,5%	1,6%	0,6%
CDS		Contrôle dynamique de stabilité	0,1%	2,3%	0,8%
CRE		Classe de réparationà l'origine	3,6%	4,4%	0,8%
ADEA		Antidémarrage actuel	3,3%	2,6%	0,3%
SLO		Système de localisation	0,1%	0,2%	0,0%
SEGMENT		Segment	1,4%	4,6%	9,1%
CO22		CO2 discrétisé	3,4%	2,3%	2,4%
CRBOEOR		Coût à l'origine de remplacement d'un bloc optique discrétisé	3,4%	3,4%	0,7%
CRPBEOR		Coût à l'origine de remplacement d'un pare brise discrétisé	3,3%	3,1%	0,7%
CRPEOR		Coût à l'origine de remplacement des pièces discrétisé	3,6%	3,1%	0,8%
HAUT		Hauteur (en mm) discrétisée	3,4%	3,0%	7,3%
VOIEAV		Voie avant (en mm) discrétisée	4,2%	4,8%	3,3%

FIGURE 3.13 – Contribution des variables

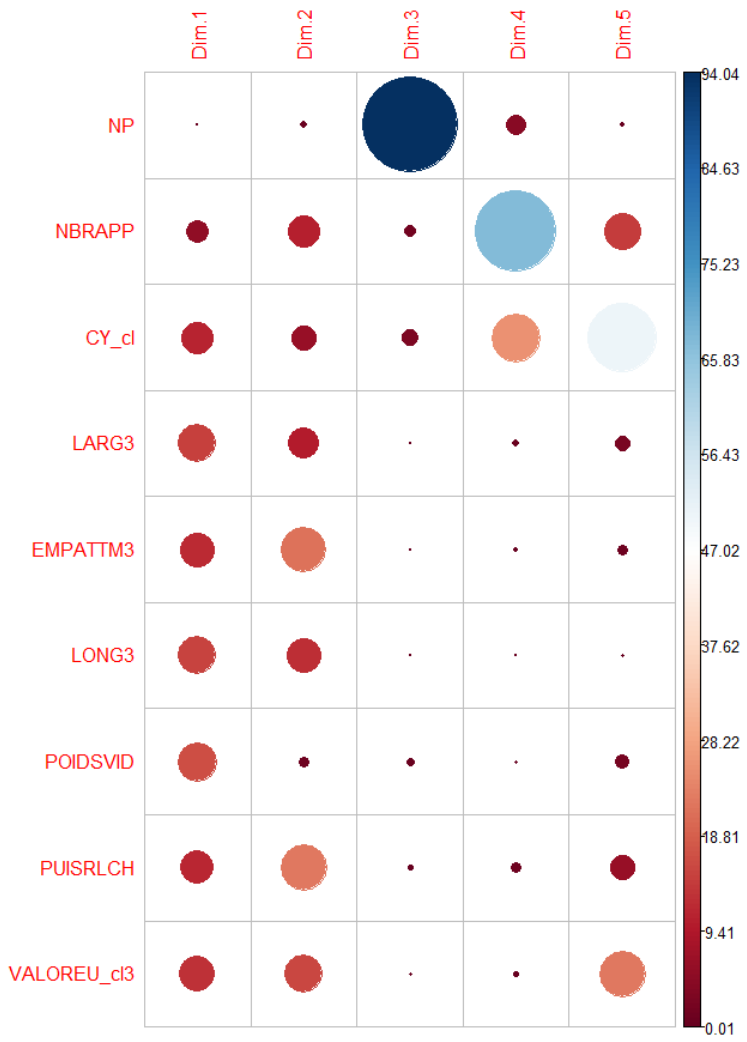


FIGURE 3.14 – Contribution des variables aux dimensions

	Puissance du véhicule	Cylindrée	Valeur à l'origine du véhicule	Largeur	Empattement	Poids à vide	Vitesse	Marque	Carrosserie	Alimentation	Classe de réparation	Emission de CO2
Puissance du véhicule		0,4419	0,4268	0,2968	0,2677	0,2641	0,4661	0,3225	0,2258	0,2938	0,2363	0,2133
Cylindrée			0,1467	0,3144	0,2958	0,2368	0,3185	0,5123	0,3269	0,3841	0,2987	0,2788
Valeur à l'origine du véhicule				0,3625	0,3727	0,3679	0,4777	0,1962	0,3855	0,2837	0,2554	0,2167
Largeur					0,6356	0,3288	0,2345	0,3876	0,4122	0,3675	0,3538	0,2876
Empattement						0,3763	0,2167	0,4794	0,4776	0,3525	0,3775	0,3496
Poids à vide							0,2216	0,2736	0,2149	0,1729	0,2675	0,2746
Vitesse								0,2184	0,2218	0,1729	0,1499	0,1516
Marque									0,2216	0,2314	0,1686	0,1131
Carrosserie										0,1716	0,1511	0,1666
Alimentation											0,1521	0,1814
Classe de réparation												0,1512
Emission de CO2												

FIGURE 3.15 – Extrait du V de Cramer

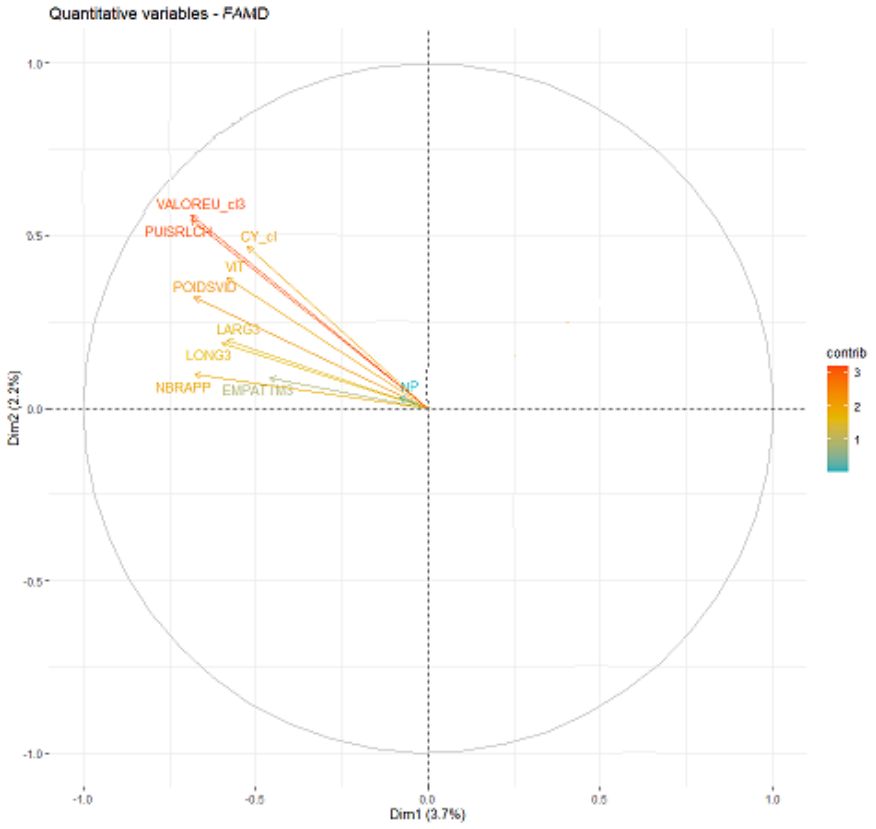


FIGURE 3.16 – Cercle corrélation - AFDM

Table des figures

Liste des tableaux