Mémoire présenté le :

pour l'obtention du titre d'Actuaire de l'Institut des Actuaires

Par : Adrien Condamin

Titre : **Construction de véhiculier et mise en perspective dans le cadre de tarification d'assurance automobile : Application sur les garanties Bris-de-Glace et Vol**

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

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

Membres du jury de l'institut des Actuaires :

–

–

–

Membres du jury de l'Essec :

–

–

–

Entreprise :

Nom :

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 du candidat

Construction de véhiculier et mise en perspective dans le cadre de tarification d'assurance automobile : Application sur les garanties Bris-de-Glace et Vol

Adrien Condamin

Décembre 2020

Résumé

Dans l'optique d'éviter l'anti-sélection et de proposer à ses clients un juste prix, les entreprises d'assurance tentent d'optimiser leur modèle de tarification. Il s'agit tout d'abord de trouver et de sélectionner de façon pertinente les variables discriminantes de leur modèle. Ce mémoire, réalisé au sein de l'entreprise Reacfin, se concentrera sur l'étude des variables véhicule en assurance automobile. Parmi les nombreuses informations accessibles à partir du modèle d'un véhicule, il s'agit de sélectionner un nombre optimal de variables afin de trouver le juste équilibre entre complexité du modèle, quantité d'information utile gardée et temps de calcul. Une des solutions possibles est de créer une variable synthétique (appelée véhiculier) contenant l'essentiel de l'information pertinente (en termes de prédiction de risque) portée par l'ensemble des variables véhicule à disposition.

La construction d'un véhiculier comporte les étapes suivantes :

- D'abord, il s'agira d'isoler la part du risque qui peut être exclusivement expliquée par les variables véhicule, que l'on nomme *effet véhicule*. Celui-ci sera extrait d'un modèle linéaire généralisé dans lequel toutes les variables pertinentes seront intégrées, à l'exception des variables véhicule. L'effet véhicule sera ainsi capté par les résidus du modèle.
- Ensuite, l'espace des variables véhicule sera réduit afin d'obtenir un sous-espace à plus petite dimension sans que la perte d'information qui en découle ne soit trop importante.
- La troisième étape correspondra à effectuer un lissage des résidus dans le sous-espace obtenu précédemment afin de donner plus de robustesse au modèle et de réduire les résidus extrêmes dont l'exposition est faible.
- Enfin, une classification supervisée des résidus sera effectuée en fonction des variables véhicule afin d'obtenir le véhiculier final. Puis les différentes classifications obtenues seront comparées sur une base de validation afin de déterminer la plus pertinente de tous.

Afin d'évaluer la pertinence du véhiculier ainsi sélectionné, il faudra comparer ses performances avec les méthodes alternatives à disposition des assureurs. Une option consiste à intégrer dans le modèle de tarification toutes les variables véhicules à disposition (si elles sont peu nombreuses) ou bien seulement certaines, jugées pertinentes par un avis d'expert. La mise en place de ce modèle (appelé modèle *benchmark*) sera l'occasion d'utiliser des pénalisations nouvelles, originales et plus complexes que celles généralement utilisées. Au-delà de la simple performance de prédiction de ces différents modèles, il s'agira aussi d'évaluer, de façon rigoureuse, l'utilité d'un tel véhiculier du point de vue du coût de mise en place de ces algorithmes et du temps de calcul nécessaire.

Mots clés : assurance automobile, classification, effet véhicule, Machine Learning, MLG, réduction de dimension, véhiculier

Abstract

In order to prevent risks of adverse selection and to offer a fair price to clients, insurance companies try to optimize their pricing models. It consists in finding and selecting the variables that are most likely to distinguish risks. The following thesis, written as an intern at Reacfin, will focus on studying variables related to vehicles for car insurance contracts. Among the vast amount of data available for a given car model, one should find the optimal number of variables that balances between model complexity, proportion of relevant information kept and computation time. One solution could be to create a variable gathering as much relevant information as possible contained by vehicle variables to predict risks.

Building a summary variable is based on the following steps :

- First, isolate the part of risk information exclusively related to vehicles (called *vehicle effect*) by building a GLM model based on all relevant variables except for the ones related to vehicles. The vehicle effect corresponds to the residuals of the model.
- Then, reduce the space made of vehicle variables into a lower-dimension subspace without losing too much information on vehicles.
- Finally, build the summary variable with a supervised classification algorithm taking the residuals as the target variable and vehicle variables as the predicting ones. For each tuple of parameters, a summary variable will be built and then be compared with others. Only the best one will be selected.

The performances of this summary variable should then be evaluated and compared to alternative models chosen by insurance companies. One option could be to put all the variables related to vehicles (or only the most relevant ones) directly in a pricing model. Building these alternative models (called *benchmark models* in the thesis) will be the opportunity to try and test the relevance of new and more complex penalty functions. Besides the performances of these models in terms of prediction, one should question the usefulness of such summary variable by taking into account the computation and implementation costs.

Keywords :car insurance, dimension reduction, GLM, Machine Learning, summary variable, vehicle effect, vehicle classification

Remerciements

J'adresse tout d'abord mes remerciements à Geoffrey Feraut et Michaël Lecuivre, mes tuteurs de stage, pour leur disponibilité, leur expertise, leurs conseils et leur aide tout au long de ce stage. Je remercie aussi Samuel Mahy en charge du pôle Non-Vie de Reacfin de m'avoir permis d'effectuer mon mémoire dans cette entreprise et pour son aide tout au long de ce stage. Je tiens aussi à remercier les collaborateurs de Reacfin pour leu accueil, leur soutien et leur disponibilité durant ce stage.

Je souhaiterais enfin remercier Marie Kratz, directrice de la filière actuariat de l'Essec et tutrice académique de mon mémoire, pour sa disponibilité, pour ses conseils concernant les aspects théoriques, mais surtout pour m'avoir transmis le goût pour l'actuariat, la rigueur mathématique et la curiosité scientifique.

Table des matières

Introduction

Dans un secteur très concurrentiel comme l'assurance où se développent les services de comparateurs et pour lequel les clients ont du mal à différencier les produits proposés par les assureurs, le prix est une variable essentielle et discriminante dans le choix des consommateurs. L'enjeu est alors pour les compagnies d'assurance de faire de la prime un outil permettant de maintenir ou augmenter ses parts de marché mais aussi d'attirer les types de risques qu'ils considèrent comme étant les plus rentables.

Cette recherche du juste prix et de définition de clients cible en assurance implique d'effectuer un choix très méticuleux des variables explicatives à intégrer dans un modèle de tarification. Dans le cas de l'assurance automobile, cette question est très importante lorsqu'il s'agit de sélectionner les variables discriminantes liées au véhicule assuré. En effet, la diversité des variables véhicule dont peut disposer un assureur implique de réfléchir en amont de toute modélisation à la pertinence de chacune. La part du risque expliquée par l'ensemble des variables liées au véhicule d'un assuré est appelé *effet véhicule*.

Une des solutions possibles pour intégrer cet effet véhicule dans le modèle de tarification consiste à synthétiser l'information détenue par toutes les variables relatives aux véhicules au sein d'une seule appelée *véhiculier*. Cette option a l'avantage de la clarté et de la simplicité, ce qui est très avantageux lorsqu'il s'agit de mettre en place des modèles parfois très complexes construits sur de très grandes bases de données. L'enjeu est d'obtenir un modèle plus simple sans pour autant perdre en capacité de prédiction portée par l'ensemble des variables véhicule.

La question à laquelle cette étude cherchera à répondre est la suivante : l'utilisation d'un véhiculier dans un modèle de tarification a priori d'un produit d'assurance automobile est-elle pertinente comparée à la simple utilisation de variables véhicule, que ce soit en termes de performances prédictives et de temps de calcul ?

La méthodologie choisie pour mettre en place un véhiculier est la suivante : dans un premier temps, il s'agit d'isoler la partie du risque de sévérité expliquée par les seules variables véhicule (appelée *effet véhicule*) en extrayant les résidus d'un premier modèle linéaire généralisé qui exclut les variables véhicule. Puis l'espace des variables véhicule sera réduit afin de faciliter la visualisation et le lissage en veillant à ne pas perdre trop d'information contenue par ces variables. Ensuite, une troisième étape consistera à retravailler ces résidus en les lissant afin de les rendre plus robustes. Enfin, grâce à un algorithme de classification, des véhiculiers seront créés (en faisant varier les différents paramètres à disposition) et seront réintégrés dans un modèle GLM afin de sélectionner le meilleur d'entre eux.

Afin d'évaluer la pertinence du véhiculier ainsi créé, il faudra par la suite comparer un modèle de prédiction de la sévérité contenant cette nouvelle variable à un modèle alternatif. Ce dernier correspondra à intégrer toutes les variables véhicule à disposition (ou bien seulement certaines) et d'ajouter un terme de pénalisation dérivé du lasso, mais plus utile et plus efficace, appelé *fused lasso*.

Ce mémoire sera organisé de la façon suivante : dans la première partie, nous définissons le cadre de l'étude, son périmètre, la méthodologie utilisée et des concepts clés. La seconde partie correspond à la définition des concepts théoriques de science actuarielle et de Machine Learning utilisés. Enfin, dans une dernière partie, les résultats obtenus à l'issue de ce mémoire seront analysés et comparés aux modèles alternatifs au véhiculier (c'est-à-dire des GLM avec toutes ou certaines variables véhicule). De plus, toutes les étapes de cette étude seront remises en question afin d'évaluer leur pertinence et leur utilité. En effet, à chacune des étapes (lors de la réduction de dimension, du lissage et de la classification), deux solutions sont développées afin de pouvoir effectuer des comparaisons, mettre en lumière les différences entre les méthodes, puis identifier et sélectionner la plus pertinente. Ce troisième chapitre sera ainsi l'occasion d'observer les bonnes performances d'algorithmes de classification communs comme les arbres de régression face à d'autres méthodes plus complexes comme les Evolutionary Trees. Sans oublier que l'utilité et la pertinence du véhiculier devront aussi être interrogées du point de vue du coût de mise en place et du temps de calcul par rapport aux méthodes alternatives.

Chapitre 1

Cadre et objectifs de l'étude

Cette partie permettra de définir le contexte dans lequel s'inscrit ce mémoire et de soulever l'intérêt de ce dernier en expliquant en quoi il répond à des thématiques importantes rencontrées par des compagnies d'assurance. Il s'agira d'abord d'énoncer quelques rappels essentiels sur le fonctionnement de l'assurance, notamment avec les concepts de segmentation et de mutualisation, puis d'expliquer en quoi la mise en place d'un véhiculier est reliée à ces principes. Enfin, le cadre de l'étude, son périmètre et la méthodologie suivie seront explicités en détails.

1.1 La segmentation en assurance

La segmentation correspond à la capacité d'un assureur à regrouper dans des groupes homogènes les assurés en fonction du risque qu'ils portent, et de variables jugées pertinentes. C'est un enjeu très important puisqu'il permet de cibler et de proposer un juste prix à chacun de ses clients et ainsi d'attirer les *bons risques* (les assurés ayant une probabilité de sinistres plus faible pour lesquels le résultat de l'assureur est positif) aux dépens des *mauvais risques* (les assurés qui ont une probabilité de sinistres plus forte pour lesquels le résultat de l'assureur est négatif). Cette segmentation permet aussi à l'assureur de lutter contre l'anti-sélection. Une bonne segmentation d'un portefeuille est donc un outil efficace face à la concurrence et permet de conserver les revenus de l'assureur.

Cet enjeu de segmentation doit être étudié en parallèle d'un autre principe, celui de mutualisation des risques sur lequel toutes les activités d'assurance se basent. Tous les assurés paient une prime qui permettra à la compagnie d'assurance d'indemniser ceux qui ont subi un sinistre. L'enjeu pour un assureur est d'obtenir un portefeuille d'assurés suffisamment développé pour que la loi des grands nombres s'applique, que la variance des primes pures (qui correspondent au produit entre la fréquence et le coût des sinistres) soit plus faible et ainsi que leur moyenne soit la proche possible de la prime pure prédite par l'assureur.

Pour effectuer une segmentation efficace, il faut d'abord obtenir des données pertinentes et fiables concernant les assurés. Dans le cas de l'assurance automobile, les variables auxquelles ont accès les entreprises d'assurance sont en général de trois types : des variables liées aux caractéristiques de l'assuré, de son véhicule et de son contrat. L'enjeu pour un assureur est de collecter le plus d'informations différentes sans que cela ne demande trop d'effort à ses assurés. Ainsi, ils ont souvent accès à

des bases de données externes qui permettent d'enrichir leurs bases de données actuelles. Dans le cas des variables liées aux véhicules, à partir du modèle d'une voiture, les assureurs peuvent avoir accès à de nombreuses caractéristiques techniques du véhicule de leurs assurés. L'enjeu de la segmentation est alors de déterminer quelles sont les informations les plus pertinentes dans l'optique de prédire la sinistralité du portefeuille d'assurance automobile.

1.2 L'intérêt d'un véhiculier

Lors de la tarification d'un produit d'assurance, l'enjeu est de développer un modèle ni trop complexe (ce qui rendrait la tarification moins lisible et transparente pour un client) ni pas assez (ce qui rendrait le modèle peu efficace et ferait courir un risque de mauvaise tarification à l'assureur). Pour cela, le nombre et le choix des variables en *input* du modèle est une étape cruciale. Cet enjeu de la complexité implique aussi de considérer les temps de calcul, de mise en place du modèle mais aussi la mise en place de matériel informatique capable d'effectuer de grandes quantités de calcul. Tous ces choix auront un impact sur les coûts internes d'une entreprise d'assurance et donc de sa compétitivité.

Pour répondre à ces enjeux, une solution consiste à créer des variables synthétiques afin de capter l'effet d'un groupe de variables sur la fréquence et le coût des sinistres. Ainsi, l'effet lié à la géographie du risque est souvent agrégé au sein d'une seule variable appelée zonier. Cela permet de réduire le degré de liberté du modèle en passant d'une variable "code postal" par exemple à une autre ne contenant que quelques modalités. De même, certains assureurs font le choix de synthétiser la part de risque portée par les caractéristiques du véhicule au sein d'une variable synthétique appelée véhiculier.

Mettre en place un véhiculier a plusieurs buts :

- Synthétiser la part du risque relative aux véhicules dans une seule variable, notamment quand l'assureur a à sa disposition une large base de données externe contenant de nombreuses variables relatives aux véhicules qui ne peuvent pas toutes être intégrées dans un modèle de tarification a priori ;
- Créer des groupes homogènes de véhicules ;
- Eviter d'utiliser des variables véhicules trop fortement corrélées ou bien redondantes dans le modèle ;
- Gagner en temps de calcul ;
- Rendre un modèle linéaire généralisé (GLM) plus robuste dans le cas où le nombre d'observations d'entraînement du modèle serait faible.

1.3 Périmètre de l'étude

L'étude suivante a été menée sur une base de données interne de l'entreprise Reacfin de contrats belges observés sur trois années consécutives (de 2014 à 2016 inclus). Parmi les garanties comprises dans cette base de données, l'étude se concentrera sur le Bris-de-Glace, cette-dernière ayant la fréquence

de sinistre la plus grande. La faible sinistralité des autres garanties ne permet pas d'obtenir de base de données de taille suffisante pour la modélisation du coût des sinistres ce qui rend la mise en place de modèles assez difficile sans traitement des données au préalable. Cependant, la mise en place d'un véhiculier pour des garanties ayant une faible sinistralité peut s'avérer pertinente puisqu'elle permet de réduire le nombre de variables. En effet, pour des modèles avec peu d'observations, le risque de surapprentissage lié au fait d'avoir trop de variables explicatives est plus important. Cette hypothèse sera étudiée à la fin de ce mémoire, à la section 3.4 pour une seconde garantie, la garantie Vol, dont le nombre d'observation est assez faible.

1.4 Méthodologie de l'étude

La méthodologie de cette étude est la suivante :

- D'une part, il s'agit de se mettre à la place des assureurs qui n'utilisent pas de véhiculier et qui intègrent directement toutes les variables véhicule ou seulement certaines dans leur algorithme de tarification. Ces modèles (décrit dans la section 2.2), qu'on pourrait appeler *modèles benchmark* (puisqu'ils intègrent toute l'information véhicule) seront ceux avec lesquels seront comparés les modèles avec véhiculier.
- D'autre part, la méthodologie de mise en place de véhiculier est la suivante :
  - Extraction des résidus (correspondant à l'effet véhicule) d'un modèle GLM ne contenant pas les variables véhicule (voir les sections 2.3 et 2.4)
  - Réduction de dimension de l'espace des variables véhicule avec l'algorithme d'Analyse Factorielle de Données Mixtes (AFDM) ou du t-distributed Stochastic Neighbor Embedding (t-SNE) (voir les sections 2.5.1 et 2.5.2)
  - Lissage des résidus par distance ou adjacence dans l'espace réduit (voir la section 2.6), la deuxième méthode nécessitant au préalable la création d'une triangulation du nuage de point (voir les sections 2.6.2.1 et 2.6.2.2)
  - Création des véhiculiers par classification des résidus lissés par Arbre de Régression (CART) et Evolutionnary Tree (voir les sections 2.7.1 et 2.7.2)
  - Comparaison des différents véhiculiers en évaluant les performances de modèles GLM contenant cette nouvelle variable sur la base de donnée de validation et sélection du meilleur véhiculier
- Enfin, il s'agira de comparer les performances du modèle GLM avec le meilleur véhiculier et celles des modèles benchmark (section 3.3.4)

Toutes ces étapes ont été réalisées sur le logiciel R.

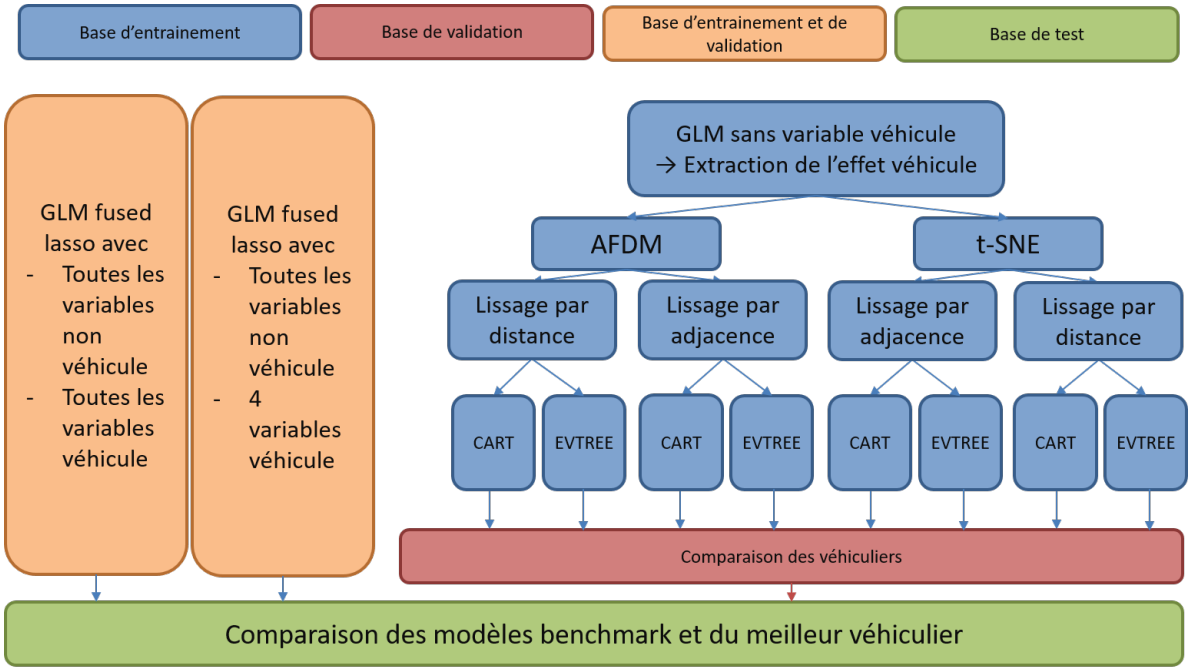


FIGURE 1.1 – Méthodologie de l'étude

Chapitre 2

Aspects théoriques de l'étude

Dans ce chapitre, nous détaillerons toutes les étapes importantes de la création d'un véhiculier et expliciterons que les concepts mathématiques théoriques sous-jacents relatifs aux différentes étapes. Les arbitrages et choix de paramétrisation de ces algorithmes seront aussi expliqués. Afin de vérifier l'importance de chacune des étapes (réduction de dimension, lissage et classification des résidus), nous développerons deux algorithmes afin de pouvoir statuer sur leur efficacité dans le troisième chapitre.

2.1 Retraitement de la base de données

Il faut d'abord découper la base de données en différentes sous-bases :

- Une base d'entraînement sur laquelle seront construits différents véhiculiers
- Une base de validation sur laquelle seront comparés ces différents véhiculiers
- Une base de test sur laquelle seront comparés le meilleur modèle véhiculier et les modèles dits optimaux (ou benchmark)

Les modèles benchmark, décrits dans la section suivante (2.2), n'ayant pas besoin d'être comparés et sélectionnés, ils sont entraînés sur la base de données d'entraînement et de validation. De plus, le meilleur véhiculier construit ne sera pas directement déterminé d'après ses performances sur la base de données de test puisque ses performances auraient été alors faussées (et jugées meilleures qu'elles ne le devraient) par rapport aux modèles *benchmark*.

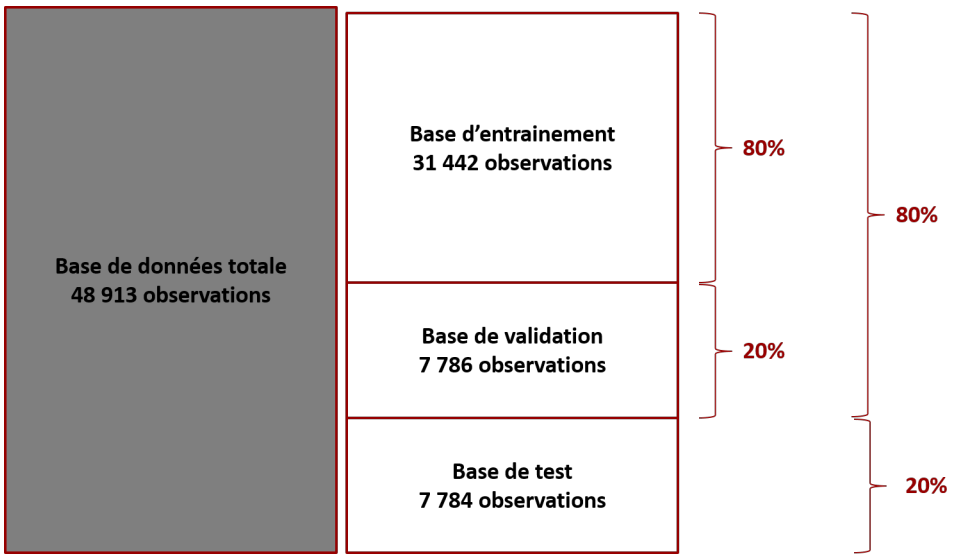


FIGURE 2.1 – Découpage de la base de données

Avant toute modélisation, les variables à disposition ont été étudiées afin de déterminer leur fiabilité, la présence ou non de valeurs aberrantes (en fonction de nos connaissances en automobile) et ainsi décider lesquelles étaient exploitables. Les retraitements suivants ont ensuite été effectués sur la base de données :

- Suppression des observations manquantes en faible quantité
- Suppression de valeurs aberrantes
- Exclusion des variables inutiles (variables redondantes, qui contiennent trop de valeurs manquantes ou bien non prédictives)
- Conversion des variables continues en variables catégorielles : une autre option aurait été de laisser ces variables telles quelles mais celles-ci ne serait considérées comme significatives que si elles sont linéairement corrélées avec la variable cible (ce qui est rarement le cas). Catégoriser ces variables permet de les garder dans le modèle si cette corrélation est bien réelle mais non nécessairement linéaire. Dans le schéma suivant, on voit une claire corrélation entre les variables mais celle-ci étant quadratique, un modèle linéaire ne pourra pas la détecter : dans ce modèle linéaire simpliste, la variable explicative est tout simplement exclue.

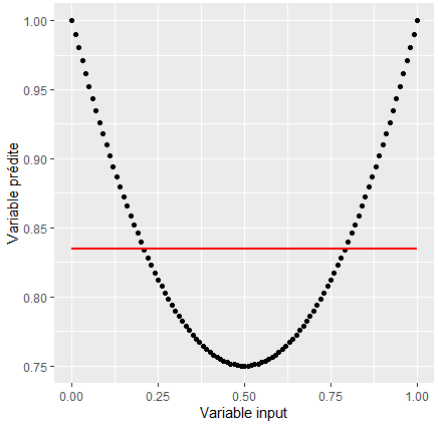


FIGURE 2.2 – Visualisation de variables quadratiquement corrélées

Or, ce passage de variable quantitative à qualitative provoque une perte d'information puisque l'on réduit le nombre de modalités à quelques unes. Le choix a donc été fait de créer un grand nombre de classes lors de la transformation en variables qualitatives. Ici les variables continues sont découpées en 50 classes de taille identique. Cela rend la base de données bien plus grande puisque chaque modalité d'une variable catégorielle devient elle-même une variable, ce qui peut poser des problèmes lors de la mise en place de modèles. Lors de la construction de modèles *benchmark*, cette catégorisation assez grossière en nombreux quantiles sera corrigée par l'ajout d'une pénalisation *fused lasso* (voir 2.2) permettant de regrouper ces nombreuses modalités entre elles lorsqu'elles ne seront pas jugées significativement distinctes.

- Retraitement des modalités manquantes des variables masse et cylindrées. Ces variables véhicule semblent pertinentes dans la mise en place d'un véhiculier. Or, elles comportaient quelques valeurs manquantes (environ 10%). Plutôt que de supprimer ces observations, nous avons considéré que des véhicules ayant la même marque et le même segment (c'est-à-dire leur catégorie : citadine, SUV, berline...) ont des masses et des cylindrées similaires. Le choix a donc été fait de remplacer ces valeurs manquantes par la moyenne de la cylindrée (respectivement de la masse) des observations non manquantes ayant les mêmes marques et segment.

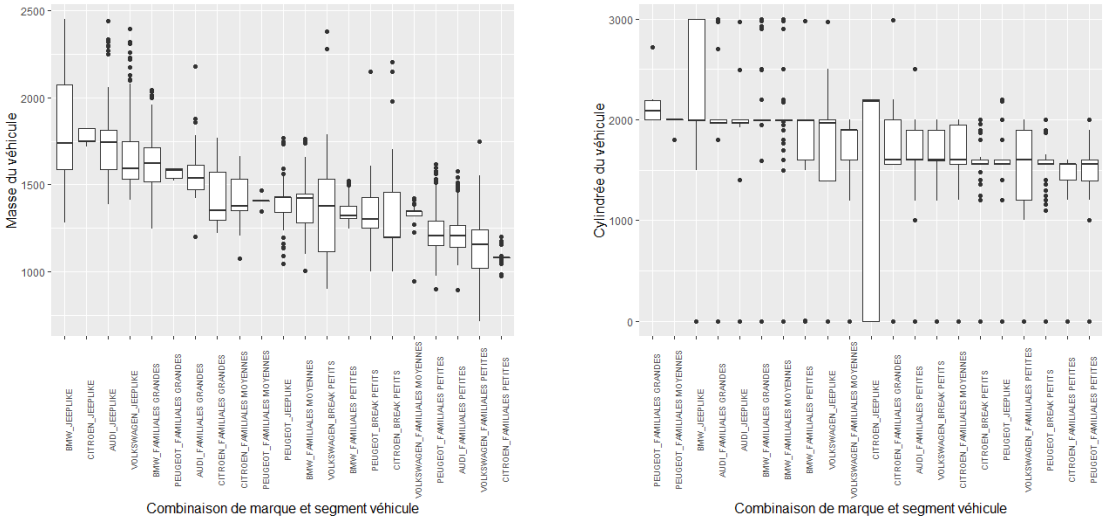


FIGURE 2.3 – Cylindrée et masse en fonction des combinaisons de marque et segment les plus présentes

Les graphiques précédents ne montrent la distribution de la masse et de la cylindrée que pour les couples marque-segment les plus représentés, mais ils permettent de vérifier que ces deux variables combinées sont corrélées à la masse et à la cylindrée puisque leurs distributions sont différentes pour chacun des couples marque-segment. Cela confirme que l'on peut en partie essayer de prédire la masse et la cylindrée de véhicules manquantes d'après leur marque et leur segment et surtout que cela semble plus pertinent que si l'on avait simplement remplacé les valeurs manquantes par leur moyenne globale.

Outre la variable "coût des sinistres", nous obtenons ainsi 17 variables non véhicule et 12 variables véhicule :

Variables non véhicule
Classe sociale de l'assuré
Sexe de l'assuré
Nombre de conducteurs assurés
Age du conducteur
Nombre d'années sans sinistres en tort
Code Mosaic (correspondant à une variable contenant des informations géographiques, démographiques et socio-économiques)
Présence d'un garage
Clause petit rouleur
Zonier
Année
Type de fractionnement
Age du contrat
Usage du véhicule
Présence d'un boîtier gobox
Titre du preneur (personne morale ou physique)
Etat civil
Conducteur principal ou secondaire

Variables véhicule
Marque du véhicule
Prix du véhicule
Puissance du véhicule
Masse du véhicule
Catégorie du véhicule
Type du véhicule
Cylindrée du véhicule
Nombre de places dans le véhicule
Segment du véhicule
État du véhicule (neuf ou occasion)
Carburant du véhicule
Age du véhicule

2.2 Élaboration du modèle de tarification comparatif

L'élaboration d'un véhiculier doit impérativement être comparée à une autre méthode. Pour cela, nous considérerons les différents modèles construits à l'aune d'un modèle dit "benchmark". Celui-ci est déterminé en intégrant dans un modèle GLM toutes les variables (qu'elles soient relatives aux véhicules ou non) c'est-à-dire un modèle disposant du maximum d'information disponible. Ici le choix est fait d'ajouter un terme de pénalisation dans le lagrangien du modèle linéaire à minimiser de façon à non seulement sélectionner les variables les plus pertinentes mais aussi à regrouper des modalités similaires au sein des variables. Sans pénalisation, certaines variables entraînées sur une base d'entraînement pourraient être considérées comme très légèrement pertinentes et seraient alors gardées dans un modèle, mais pourraient par la suite réduire la capacité prédictive de ce modèle sur une nouvelle base de données. Ajouter un terme de pénalisation permet de se prémunir contre le risque de surapprentissage en simplifiant le modèle et en ne gardant que les variables les plus pertinentes.

Le schéma suivant présente le risque de surapprentissage d'un modèle. En l'absence de membre de pénalisation dans la fonction de perte, la complexité choisie par le modèle (notée c_2) correspond à

la valeur pour laquelle l'erreur est minimale sur la base d'entraînement. Dans ce cas, le modèle aura tendance à choisir une complexité trop grande. Lorsque ce modèle sera utilisé sur une nouvelle base de données de test, l'erreur correspondante (E_2) ne sera pas minimale cette fois-ci, d'où le surapprentissage du modèle sur la base d'entraînement. L'ajout d'une pénalité permet de réduire la complexité du modèle (afin de s'approcher de l'optimum c_1) correspondant à une erreur minimale E_1.

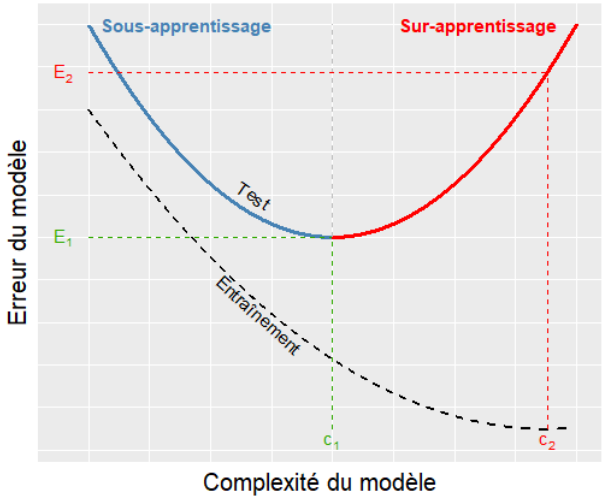


FIGURE 2.4 – Évolution de l'erreur du modèle en fonction de la complexité sur les bases d'entraînement et de test

Parmi les pénalisations les plus pertinentes, les plus connues sont la pénalisation lasso (voir [6]), la pénalisation ridge (voir [5]) et la pénalisation elastic net (voir [4]). Dans ce mémoire, le choix est fait d'introduire un autre type de pénalisation, dérivée du lasso, appelée fused lasso ([20]). Comme l'expliquent les auteurs, cette pénalisation présente l'avantage (comme le fait la pénalisation lasso) d'être *sparse* c'est-à-dire qu'elle permet d'obtenir un modèle simplifié en poussant le maximum de coefficients β à 0. Mais l'ajout d'un terme supplémentaire dans la pénalisation permet aussi de minimiser la différence entre les coefficients d'une même variable catégorielle et ainsi d'être *sparse* au sein des variables en regroupant entre elles les modalités des variables catégorielles. Cette méthode est particulièrement utile dans les cas particuliers où d le nombre de variables est grand comparé à n le nombre d'observations. Dans le cas présent, le nombre de variables est assez grand puisque d'une part, toutes les variables sont catégorielles (dont certaines ont de nombreuses modalités comme la variable *marque du véhicule*) et d'autre part, parce que lors de la transformation des variables continues en variables catégorielles, le choix a été fait de créer beaucoup de modalités afin de perdre le moins d'information possible. Ainsi, ce type de pénalisation s'avère pertinent puisqu'il permet de simplifier des modèles dont le nombre de variables est très grand, très coûteux en temps de calcul et donc portant un risque de surapprentissage.

On définit n_i le nombre de modalités de la variable i et $(\beta_{i,j})_{i\in\{1,\dots,d\},j\in\{1,\dots,n_i\}}$ le coefficient de la j-ième modalité de la variable i. Le problème revient à chercher $\hat{\beta} = (\hat{\beta}_{1,1}, \dots, \hat{\beta}_{d,n_d})$ tels que :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, l(X, Y, \beta) + \lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{n_i} |\beta_{i,j}| + \lambda_2 \sum_{i\in V_1}\sum_{j=2}^{n_i} |\beta_{i,j} - \beta_{i,j-1}| + \lambda_3 \sum_{i\in V_2}\sum_{k=2}^{n_i}\sum_{j=2}^{n_i} |\beta_{i,j} - \beta_{i,k}|$$

avec λ_1, λ_2 et λ_3 les paramètres de pénalisation à cross-valider, l la log-vraisemblance d'un modèle

GLM Gamma, V_1 l'ensemble des variables explicatives contraintes par une pénalisation *fused lasso* et V_2 celles qui le sont par une pénalisation *generalised fused lasso*.

Notons que cette pénalisation correspond à l'ajout de nouvelles contraintes à la pénalisation lasso qui est gardée (ici le premier terme de pénalisation). Pour les variables catégorielles, la pénalisation *fused lasso* compare et propose des regroupements entre deux modalités consécutives uniquement. Cette pénalisation est donc utile pour les variables chronologiques ou les variables continues transformées en variables catégorielles. Pour les autres variables qualitatives, la pénalisation *generalised fused lasso* permet de comparer et potentiellement regrouper toutes les modalités d'une même variable. Ces étapes sont très couteuses en temps de calcul, ce qui permet d'expliquer en partie pourquoi ce modèle prend plus de temps d'exécution qu'un GLM plus traditionnel.

Il s'agit à présent de créer un modèle benchmark qu'un assureur pourrait utiliser dans le cas où il ne ferait pas appel à un véhiculier. Or le modèle intégrant la totalité des variables à disposition n'est pas souvent le choix effectué par les actuaires de compagnies d'assurance. En effet, ce modèle est très coûteux en termes de temps de calcul (voire impossible à faire tourner si la base de données est très grande et que les ressources informatiques de l'assureur sont limitées) et nécessite un travail préalable de sélection des variables véhicules. Une méthode qui semblerait plus pertinente serait de ne sélectionner que quelques variables véhicule jugées pertinentes afin d'alléger le modèle sans pour autant sacrifier les performances de prédiction du modèle ni perdre trop d'information.

Dans le cadre de cette étude, ces deux modèles, (celui avec toutes les variables véhicule, l'autre avec certaines seulement) ont été exécutés pour plusieurs raisons. D'une part, lancer le modèle GLM avec toutes les variables permet d'avoir un véritable benchmark et donc d'avoir une idée des performances de prédiction dans le cas où toute l'information à disposition serait disponible. De plus, cela nous permettra de comparer les deux modèles et de connaître l'impact d'une réduction du nombre de variables véhicule en termes de prédiction. Dans le second modèle, seules les quatre variables véhicules les plus pertinentes sont utilisées (ainsi que toutes les variables non véhicule). Celles-ci correspondent aux quatres variables pour lesquelles l'Akaike Information Criterion (AIC, voir Annexe B.3) et le *Residual Sum of Squares* sont les plus faibles lorsqu'on construit un modèle de prédiction des coûts de sinistres Bris-de-Glace ne contenant qu'une variable. Ce sont les variables "Prix", "Puissance", "Marque" et "Segment du véhicule".

Variable	AIC	RSS
Prix	450 925.4	3.851×10^9
Puissance	450 927.9	3.852×10^9
Marque	451 363.7	3.888×10^9
Segment	451 578.8	3.916×10^9
Poids-lourd ou voiture	453 273.9	4.091×10^9
Carburant	453 840.7	4.151×10^9
Nombre de places	453 980.2	4.166×10^9
Catégorie de véhicule	454 033.9	4.172×10^9
Cylindrée	454 039.7	4.173×10^9
Masse	454 046	4.174×10^9
Age du véhicule	454 046	4.174×10^9

TABLE 2.2 – AIC et RSS des modèles GLM avec une seule variable véhicule

Ces modèles GLM avec pénalisation fused lasso seront déterminés sur le logiciel R avec le package *smurf*. Les résultats de ces deux modèles benchmark ainsi que la pertinence de la pénalisation *fused lasso* sont analysés dans la partie 3 de ce mémoire (voir 3.1).

2.3 Extraction de l'effet véhicule

Une des méthodes les plus utilisées pour isoler une composante du risque consiste à mettre en place un modèle intègrant toutes les variables explicatives du risque à l'exception de celles qui traitent de l'effet en question et à considérer que l'erreur de prédiction correspond aux variables exclues. Dans le cas de notre étude,il est courant de considérer un modèle GLM (voir section 2.4) avec toutes les variables non véhicule pour ensuite en extraire les résidus. Ces-derniers correspondent ainsi à l'effet véhicule, mais aussi parfois à d'autres effets que le modèle n'a pas pris en compte. C'est notamment le cas quand on n'a pas à disposition toutes les variables explicatives du risque ou que l'on oublie d'autres variables potentiellement pertinentes. Ici, nous émettons l'hypothèse que les variables à disposition permettent d'expliquer correctement le coût des sinistres, c'est-à-dire que les résidus contiennent principalement l'effet véhicule.

Cependant, cette méthode est basée sur l'hypothèse que les variables véhicule et non-véhicule ne sont pas corrélées. Dans le cas contraire, une partie plus ou moins importante de l'effet véhicule sera captée dans le GLM par les variables non-véhicule fortement corrélées aux variables véhicule. Cette hypothèse paraît a priori assez difficilement soutenable pour certains risques tels que l'automobile. En effet, on peut aisément supposer que la marque d'un véhicule (ou bien son prix) sont fortement corrélés à la classe socio-professionnelle de l'assuré. Dans le cas où cette hypothèse n'est pas vérifiée, le véhiculier construit ne contiendra pas la totalité de l'effet véhicule, puisque celui-ci sera capté par certaines variables non

véhicule.

D'autres approches ont été développées afin de créer un véhiculier sans se baser sur cette hypothèse d'indépendance entre variables véhicule et non véhicule. Une approche dite "comportementale" a été développée par Magali Ruimy dans son mémoire d'actuariat ([1]). L'hypothèse faite est que le véhicule n'est pas dissociable de son conducteur (et suppose que les caractéristiques du conducteur donnent des indications sur son véhicule et inversement). L'idée est donc de créer un véhiculier qui ne prenne pas seulement en compte les variables véhicule mais aussi un profil de conducteur correspondant. Cette méthode ne semble pas donner de meilleurs résultats et se base sur d'autres hypothèses que l'on peut aussi critiquer.

La méthode retenue dans cette étude reposera donc sur l'hypothèse qu'il est en effet possible d'isoler l'effet véhicule, notamment grâce à un retraitement préalable des données. Afin de pouvoir isoler efficacement l'effet véhicule, il faut auparavant s'assurer de l'absence de corrélation entre les variables véhicule et non véhicule et du fait que la majeure partie de l'effet véhicule sera contenue dans les résidus du modèle GLM. Pour cela, une première étape consiste à calculer le V de Cramer (voir Annexe D) de toutes les paires de variables. Puis il s'agit de déterminer un seuil au-delà duquel on supprime une variable. Ainsi, si une variable véhicule et non véhicule ont un V de Cramer supérieur à ce seuil, alors la variable non-véhicule est écartée (ce qui permet de garder toutes les variables véhicules pour la suite de l'étude). Cette étape présente aussi l'avantage d'effectuer un premier tri dans les variables intégrées dans le modèle GLM et ainsi en réduire sa complexité et donc son temps de calcul.

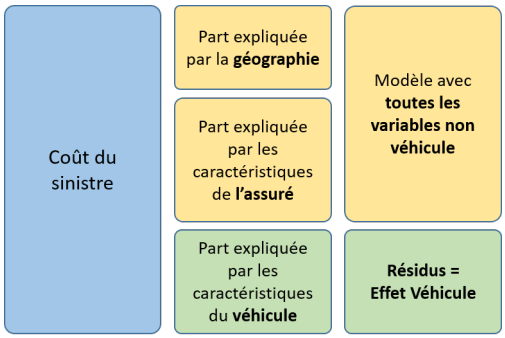


FIGURE 2.5 – Extraction de l'effet véhicule

Ici, le seuil a été choisi arbitrairement (après en avoir testé plusieurs), en fonction du nombre de variables qui sont exclues (un seuil pour lequel aucune ou toutes les variables sont exclues ne semble pas pertinent) mais aussi en fonction de la définition du V de Cramer (au-delà de 0.5 on peut considérer que deux variables sont corrélées). Ainsi, le seuil choisi est 0.25, ce qui exclut 4 variables parmi 16, c'est-à-dire qu'il y a 4 variables non véhicule pour lesquelles le V de Cramer avec au moins une variable véhicule est supérieur à 0.25. Les variables non véhicule écartées sont les suivantes : "sexe de l'assuré", "clause petit rouleur", "état civil" et "dénomination conducteur principal ou secondaire". Ces variables ne semblent pas a priori les plus pertinentes pour la prédiction de la sévérité d'une garantie Bris-de-Glace.

2.4 Modèles linéaires généralisés

Les modèles linéaires généralisés (GLM) font partie des algorithmes de tarification a priori d'assurance non-vie les plus répandus pour la modélisation de la fréquence et du coût des sinistres. Il s'agit d'expliquer une variable réponse (notée Y) grâce à d variables explicatives (notées $X = (X_1, ..., X_d)$), en particulier de modéliser $E[Y|X]$ en supposant qu'il existe une fonction dite de lien g telle que $E[Y|X] = g(X\beta)$. Les paramètres à déterminer sont donc les $\beta = (\beta_1, ..., \beta_d)$ qui vérifient cette équation.

Un modèle linéaire généralisé suppose que la variable $Y|X$ appartient à une famille exponentielle, c'est-à-dire, une loi dont la densité est de la forme

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

avec $a(\theta)$ une fonction de classe C^2 et convexe et $c_\phi(y)$ indépendant de θ. Il s'agit donc de minimiser $\log \mathcal{L}$, la log-vraisemblance des $(y_i)_{i\in\{1,...,n\}}$:

$$\log \mathcal{L}(y_1, \ldots, y_n, \theta, \phi) = \log\left(\prod_{i=1}^{n} f_Y(y_i, \theta, \phi)\right) = \sum_{i=1}^{n} \log(c_\phi(y_i)) + \left(\frac{y_i\theta - a(\theta)}{\phi}\right)$$

Le problème est donc équivalent à trouver :

$$\underset{\beta_1,...,\beta_d}{\operatorname{argmin}} \log \mathcal{L}(y_1, \ldots, y_n, \theta, \phi) = \underset{\beta_1,...,\beta_d}{\operatorname{argmin}} \sum_{i=1}^{n} \log(c_\phi(y_i)) + \left(\frac{y_i\theta - a(\theta)}{\phi}\right)$$

Ce modèle peut être légèrement modifié en ajoutant une pénalisation dans le but de le simplifier mais aussi d'éviter le surapprentissage. La fonction à minimiser devient alors :

$$\underset{\beta_1,...,\beta_d}{\operatorname{argmin}} \log \mathcal{L} + \lambda \text{ pen}(\beta_1, \ldots, \beta_d)$$

avec λ le paramètre qui attribue un poids relatif au membre lié à la pénalisation par rapport à la partie vraisemblance. Ce paramètre est généralement optimisé par cross-validation.

Dans le cadre de ce mémoire, seule l'étude sur la prédiction de la sévérité des sinistres est effectuée. Différentes lois de distribution peuvent être choisies (log-normale, gamma, tweedie...). La distribution tweedie a été écartée puisqu'elle est souvent utilisée comme une alternative à la Gamma ou à la log-normal dans le cas où certaines observations à prédire seraient nulles. Or dans le cas de l'étude sur la sévérité, aucun sinistre dans cette base de donnée n'a un coût nul. Il reste à comparer les performances des lois Gamma et Log-normal afin de déterminer la plus pertinente. Voici les performances en termes de Mean Squared Error (MSE) (voir B.1) des deux modèles GLM Gamma et Log-normal avec toutes les variables à disposition sur la base de données d'entraînement :

Modèle	MSE
GLM Gamma	92 544. 42
GLM Log-normal	110 516.4

Après avoir testé les performances des différentes lois, le choix d'une loi Gamma semble le plus pertinent.

Les résidus du modèle de coût, notés $r = (r_1, \ldots, r_n)$, sont définis de la manière suivante :

$$r_i = \frac{\text{coût observé}_i}{\text{coût prédit}_i}$$

Ainsi, afin d'extraire l'effet véhicule de la base de données, il suffit de construire un modèle GLM avec toutes les variables non véhicule (à l'exception de celles qui ont été écartées parce qu'elles étaient jugées trop corrélées aux variables véhicule) et d'en extraire les résidus définis ci-dessus et correspondant à l'effet véhicule. Ceux-ci vont ensuite être retravaillés dans les prochaines étapes.

2.5 Réduction de dimension de l'espace des variables véhicule

Avant de retravailler les résidus issus du précédent modèle, une première étape consiste à réduire l'espace des variables véhicule dont la dimension est grande (on compte 12 variables véhicule dans la base de données) à un espace de dimension plus restreinte. D'une part, cela permettra d'attribuer à chaque observation des coordonnées en fonction de ses caractéristiques véhicule, et ensuite de procéder à un lissage spatial. D'autre part, cela permettra d'obtenir des représentations graphiques de ce nouveau nuage de points dans un espace en 2 ou 3 dimensions, afin d'observer sa structure.

Les deux algorithmes de réduction de dimension choisis sont l'Analyse Factorielle de Données Mixtes (AFDM) et le t-Distributed Stochastic Neighbor Embedding (t-SNE), deux méthodes dont l'efficacité et la pertinence seront comparées dans la section 3.3.1.

2.5.1 Analyse Factorielle de Données Mixtes

Avant de lisser les résidus du modèle, il faut effectuer une réduction de dimension de l'espace des variables prédictives qui est parfois très grand si l'on a beaucoup de variables à disposition. Une des méthodes les plus utilisées est l'analyse en composantes principales (ACP). Comme nous avons à disposition des variables quantitatives mais aussi qualitatives, nous utiliserons une version particulière appelée Analyse Factorielle de Données Mixtes (AFDM) qui s'inspire de la méthode d'Analyse en Composantes Principales (ACP) (voir Annexe E) pour les variables quantitatives et de celle de l'Analyse en Composantes Multiples (ACM) (voir Annexe F) pour les variables qualitatives.

Contrairement à la partie relative à la construction du GLM permettant d'extraire les résidus, dans laquelle toutes les variables étaient catégorisées afin de détecter des corrélations non linéaires avec la variable à prédire, pour la partie réduction de dimension, le choix est fait de prendre les variables continues non catégorisées. Cela permet de ne pas perdre d'information lors du passage de variable continue à catégorielle et cela permet de réduire le nombre de variables de la base de données et ainsi de rendre l'algorithme de réduction de dimension plus rapide.

Soit K_1 le nombre de variables continues (notons $(x_{ij})_{i\in\{1,\dots,n\};j\in\{1,\dots,K_1\}}$ les observations correspondantes) et K_2 le nombre de variables catégorielles ayant chacune $(m_i)_{i\in\{1,\dots,K_2\}}$ modalités, soit $m_1 + \cdots + m_{K_2} = M$ variables dans son tableau disjonctif complet (dont les valeurs seront notées $(x'_{ij})_{i\in\{1,\dots,n\};j\in\{1,\dots,M\}}$). La matrice de données est donc la suivante :

$$
\mathrm{A} = \begin{array}{c} \\ 1 \\ \vdots \\ \vdots \\ n \end{array}
\begin{array}{c}
\begin{array}{ccc} 1 & \dots & K_1 \end{array} \\
\left(\begin{array}{ccc|ccc}
x_{11} & \dots & x_{1K_1} & x'_{11} & \dots & x'_{1M} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_{n1} & \dots & x_{nK} & x'_{n1} & \dots & x'_{nM}
\end{array}\right)
\end{array}
\begin{array}{c}
\\
\begin{array}{ccc} 1 & \dots & M \end{array}
\end{array}
$$

Dans le cas des variables continues , l'ACP essaie de maximiser

$$\sum_{k=1}^{K_1} r^2(k, F_s)$$

avec r le coefficient de corrélation et F_s les facteurs. Dans le cas des variables qualitatives, l'ACM vise à maximiser

$$\sum_{j=1}^{M} \eta^2(j, F_s)$$

avec η^2 le carré du rapport de corrélation. Ces deux analyses sont définies plus en détails dans les annexes E et F. Le critère de maximisation de l'AFDM est donc :

$$\sum_{k=1}^{K_1} r^2(k, F_s) + \sum_{j=1}^{M} \eta^2(j, F_s)$$

Une des mesures à analyser après la construction d'une AFDM est le pourcentage de variance expliquée par les différentes composantes. Celle-ci décroît logiquement en fonction de l'ordre des composantes principales. Plus ce pourcentage est grand, plus les composantes principales captent la structure des données d'origine. Voilà les résultats obtenus pour l'AFDM sur les variables véhicule (la totalité des résultats en termes de contribution, coordonnées et cosinus sont en Annexe G) :

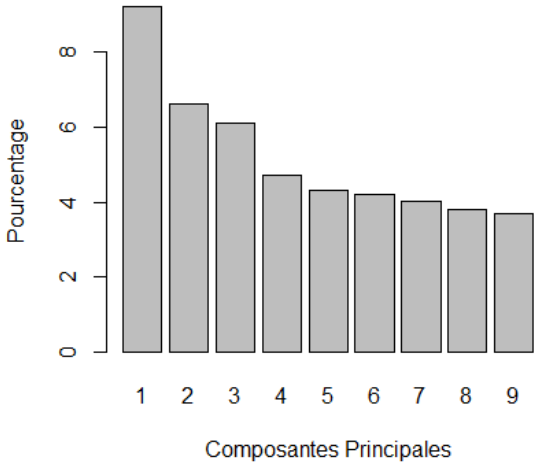


FIGURE 2.6 – Pourcentage de la variance expliquée par composante principale

Le choix a été fait de ne garder que les 3 premières composantes principales puisque cela permet de garder le maximum de dimension tout en pouvant les visualiser. De plus, on voit un coude, c'est-à-dire que le pourcentage de variance expliquée par les composantes diminue à partir de la quatrième dimension. Ainsi, 23% de la variance des données est expliquée par ces 3 dimensions.

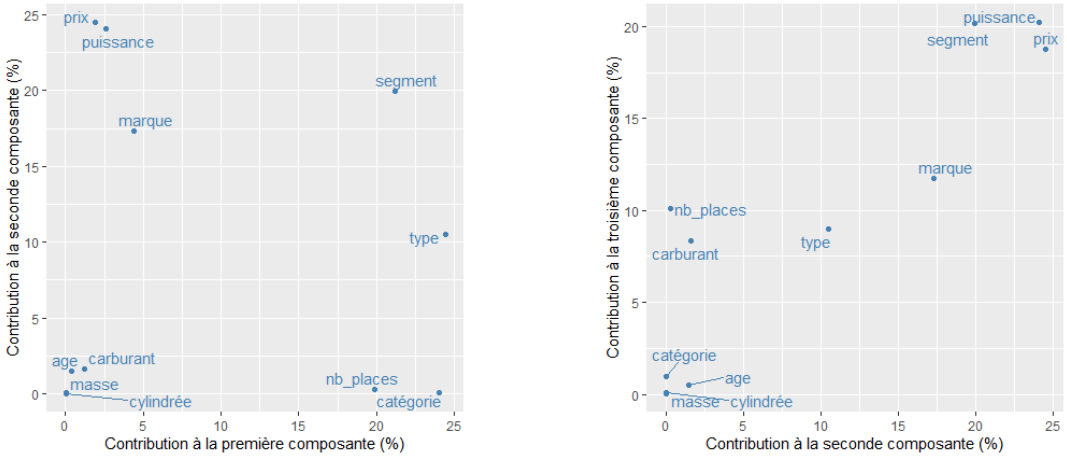


FIGURE 2.7 – Contribution des variables par composante principale

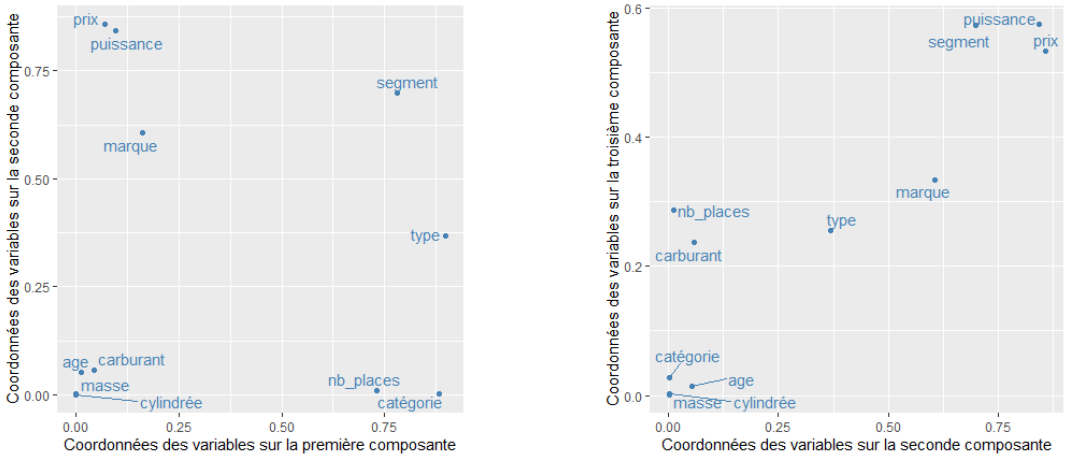


FIGURE 2.8 – Coordonnées des variables par composante principale

Lors de l'étape de réduction de dimension, celle-ci n'étant pas supervisée par les résidus, il existe un risque que certaines variables véhicule pertinentes pour expliquer ces résidus ne soient pas fortement représentées dans les trois composantes principales de l'AFDM. Or ici on remarque que les variables qui pourraient paraître pertinentes pour prédire le coût des sinistres Bris-de-Glace (c'est-à-dire le prix, la marque, le segment du véhicule...) semble aussi bien représentées dans les trois dimensions. De plus, on peut supposer que l'AFDM comporte l'avantage de donner peu de poids à des variables peu discriminantes et dont la variance est faible. Or celles-ci, par définition, ont peu de chance d'être discriminantes pour prédire les résidus contenant l'effet véhicule. Ainsi, cette étape permet de réduire l'espace des variables véhicule en enlevant la faible information contenue par des variables peu pertinentes pour la prédiction des coûts de sinistres.

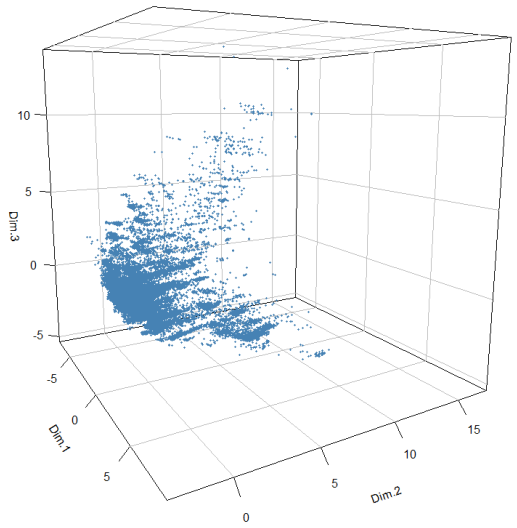


FIGURE 2.9 – Visualisation en 3 dimensions des observations dans l'espace réduit de l'AFDM

2.5.2 t-distributed Stochastic Neighbor Embedding

La seconde méthode de réduction de dimension choisie est celle du *t-distributed Stochastic Neighbor Embedding* (t-SNE). Elle était d'abord utilisée pour la visualisation de données, mais elle peut aussi être utile plus globalement dans des sujets de réduction de dimension. Alors que l'ACP (et l'AFDM) sont des techniques de réduction de dimension linéaire qui visent à garder une grande distance dans l'espace réduit pour des points très différents dans l'espace de départ, la méthode t-SNE tente de capturer, de façon non linéaire, à la fois la structure locale des données (c'est-à-dire garder proches dans l'espace réduit des points qui l'étaient dans l'espace d'origine) mais aussi de définir une structure plus globale des données à travers des clusters. Cette particularité semble donc très pertinente pour le lissage et permet d'une certaine manière de préparer à la fois le lissage par distance (du fait de la proximité des points dans l'espace d'arrivée) mais aussi par adjacence (du fait de la structure globale en clusters).

La méthodologie des t-SNE s'inspire de celle des SNE (voir Annexe H) puisqu'il s'agit de déterminer, dans l'espace de départ et d'arrivée, la probabilité conditionnelle que le point i choisisse pour voisin le point j en fonction de la distance qui les sépare, puis de faire coïncider cette probabilité dans les deux espaces. Mais le t-SNE se distingue sur deux points :

- La fonction de coût est modifiée afin d'être symétrique : au lieu de considérer $p_{j|i}$ et $q_{j|i}$ les probabilités conditionnelles que le point i choisisse pour voisin le point j dans les espaces de départ et d'arrivée, le choix est fait de calculer p_{ij} et q_{ij}, les probabilités conjointe des points x_i et x_j d'une part et y_i et y_j d'autre part (les x_i et y_i étant les coordonnées de la i-ème observation dans les espaces de départ et d'arrivée). Or dans le cas où x_i serait un point extrême, p_{ij} est très petit, pour tout j, ce qui implique que quel que soit le y_i correspondant, la fonction de coût variera très peu. Pour éviter cela, le choix est fait de définir p_{ij} par

$$\forall (i,j) \in \{1,\dots,n\}^2, \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

  ce qui assure d'avoir pour tout i entre 1 et n, $\sum_j p_{ij} > \frac{1}{2n}$

- La loi de distribution pour l'espace réduit est une Student de degré de liberté 1 (et non pas une gaussienne). Le choix d'une distribution à queues plus épaisses permet de mieux représenter dans l'espace réduit les distances entre des points éloignés dans l'espace de départ. On définit donc q_{ij} de la manière suivante :

$$\forall (i,j) \in \{1,\dots,n\}^2, \quad q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

Enfin, tout comme pour la méthode SNE, l'objectif est de rendre $(q_{ij})_{(i,j)\in\{1,\dots,n\}^2}$ et $(p_{ij})_{(i,j)\in\{1,\dots,n\}^2}$ égaux. Pour cela, on définit la divergence de Kullback-Leibler entre p_{ij} et q_{ij} comme étant la fonction à minimiser. Il s'agit donc de déterminer $\hat{Y} = (\hat{y}_1, \dots, \hat{y}_n)$ vérifiant :

$$\hat{Y} = \underset{(y_1,\dots,y_n)}{\operatorname{argmin}} \sum_i \sum_j p_{j|i} \log \frac{p_{ij}}{q_{ij}}$$

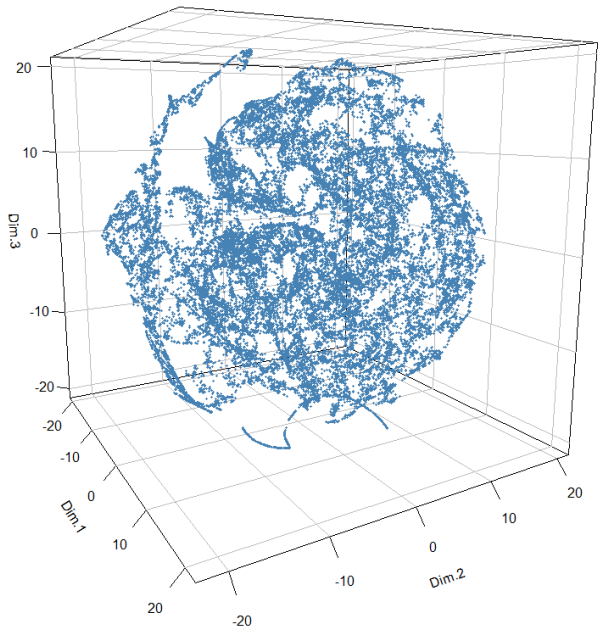


FIGURE 2.10 – Visualisation en 3 dimensions des observations dans l'espace réduit du t-SNE

2.6 Lissage spatial

Lisser les résidus du modèle a plusieurs intérêts. Tout d'abord, en supposant que des véhicules ayant des caractéristiques similaires devraient avoir une sinistralité similaire, le lissage spatial permettrait de rendre plus robustes les résidus issus d'observations dont l'exposition est trop faible. De plus, cela permet de réduire l'importance de certaines valeurs extrêmes de résidus issus d'un sinistre dont le coût est exceptionnel. Ainsi, on devrait obtenir une meilleure classification puisque lisser les résidus permettra ensuite à l'algorithme de classification de ne pas créer une classe peu robuste constituée de quelques observations extrêmes.

Pour lisser les résidus de notre modèle, nous avons deux méthodes à disposition (en fonction de ce que l'on considère être un voisin) : un lissage basé sur la **distance** et un autre basé sur l'**adjacence**. Ces lissages ne présupposant pas de modèle sous-jacent, nous ne pouvons pas évaluer ces lissages directement sur de nouvelles observations ou bien mesurer la qualité de l'ajustement. La seule manière de comparer ces différents lissages est de mesurer les performances des véhiculiers issus du lissage dans le GLM final sur la base de validation et de les comparer avec celles de véhiculiers construits sans lissage. Toutes les combinaisons possibles de ces paramètres de lissage ont donc été testées puis intégrées dans des véhiculiers. La comparaison de ces différents paramètres est détaillée en section (3.2).

Une autre question se pose : faut-il un lissage élevé ou bien faible ? A priori, rien ne semble indiquer qu'un lissage fort permettra d'obtenir un meilleur véhiculier qu'un faible lissage (et inversement). La

seule hypothèse possible est de supposer qu'un lissage extrême (où tous les points seraient lissés par tous les autres de façon à n'obtenir plus qu'une unique valeur) ou bien un lissage inexistant ne seront pas des solutions de lissage optimales. Seule une comparaison a posteriori des performances des différents véhiculiers permettra de savoir si ceux issus d'un lissage important donnent de meilleurs résultats que ceux ayant subi un lissage plus faible. Cette question sera abordée en section 3.2.

2.6.1 Lissage basé sur la distance

Un lissage basé sur la distance considère que pour chaque point étudié, les voisins intervenant dans le lissage sont ceux qui se trouvent à l'intérieur d'une boule dont le centre est le point en question et le rayon d est spécifié. Dans ce cas, la définition de voisin ne se base que sur l'étape de réduction de dimension (t-SNE ou AFDM) et on considère que cette étape a permis de garder correctement les distances entre deux observations similaires. Soient $(r_i)_{i\in\{1,\dots,n\}}$ le résidu à lisser et $(r'_i)_{i\in\{1,\dots,n\}}$ ce même résidu après lissage. Celui-ci est déterminé par l'équation suivante :

$$r'_i = 0.5\left(\frac{\sum_{j\in V_i} r_j d_{ij}^{-P}}{\sum_{j\in V_i} d_{ij}^{-P}} + r_i\right)$$

avec V_i l'ensemble des voisins du point i intervenant dans le lissage (c'est-à-dire les k points les plus proches du point i dans un rayon d) et d_{ij} la distance euclidienne entre les points i et j. Ici le choix est fait de donner un poids identique à la moyenne des résidus des voisins et à r_i ; on aurait pu faire varier ce coefficient (ici fixé à 0.5) entre 0 et 1 (comme c'est le cas pour le lissage par adjacence). Ce choix est contestable mais le nombre de paramètres à optimiser est déjà important, alors le nombre de modèles à cross-valider grandit à chaque ajout de nouvelles variables à optimiser.

Ici les paramètres à optimiser sont les suivants :

- Le rayon d de la boule délimitant le périmètre de lissage,
- Le nombre maximal, k, de voisins à étudier à l'intérieur de la boule
- La puissance P, appliquée à la distance et qui donne plus ou moins d'importance aux points les plus proches. En effet, les fonctions $f_P : x \mapsto x^{-P}$, avec $P \in \mathbb{N}$ sont convexes et leur décroissance est d'autant plus rapide sur $\mathbb{R}_+^*$ que P est grand puisque la dérivée vaut pour x strictement positif, $f'_P(x) = -Px^{-P-1}$. Supposons que le point i ait deux voisins, j_1 à une distance de 0.25 et j_2 à 0.5, soit deux fois plus éloigné. Si on prend $P = 1$, alors le poids de j_1 sera deux fois plus grand que celui de j_2. Si on prend $P = 5$, comme $0.25^{-5} = 1024$, et $0.5^{-5} = 32$, alors j_1 aura un poids 32 fois plus grand que j_2. Ainsi, plus P est grand, plus le lissage donnera proportionnellement plus de poids aux points les plus proches.

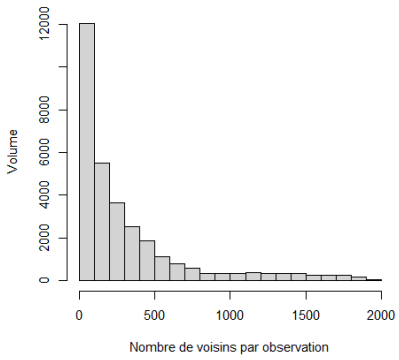
Pour déterminer les différentes valeurs du paramètre d candidates à la validation croisée, il faut d'abord avoir une idée de la distance entre ces points. La base de données étant assez conséquente, on ne pourra pas calculer la distance entre tous les points (pour une base de données de n lignes, cela reviendrait à calculer $C_n^2 = n(n-1)/2$ distances). Une solution alternative consiste à prendre différents rayons puis déterminer, en moyenne, combien de points participent au calcul de lissage pour ce rayon. Ainsi, il s'agit

non pas de déterminer différentes valeurs de d, mais plutôt de trouver différents couples (k, d) à étudier (k étant le nombre maximum de voisins entrant dans le calcul du lissage). 4 types de couples (k, d) se démarquent et seront testés :

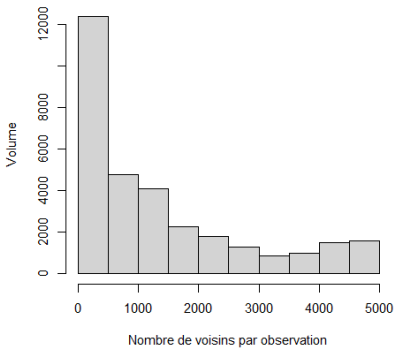
- Si k est grand et d est grand, cela signifie que l'on choisit d'intégrer beaucoup de points dans le lissage, y compris des points éloignés.
- Si k est petit et d est petit, cela signifie que l'on choisit d'intégrer peu de points dans le lissage, dans un rayon proche.
- Si k est petit et d est grand, cela signifie que l'on choisit d'intégrer peu de points dans le lissage, y compris des points parfois éloignés.
- Si k est grand et d est petit, cela signifie que l'on choisit d'intégrer beaucoup de points dans le lissage dans un rayon proche. Cette situation semble a priori la plus pertinente puisqu'on considère que tous les points proches d'une observation devraient participer au lissage. Il faut cependant veiller à ne pas choisir un rayon d trop petit qui exclurait du lissage de trop nombreuses observations.

Notons que l'AFDM et le t-SNE n'ont pas créé des nuages de points dont la dispersion et les distances sont comparables. Il faut donc choisir des valeurs de d et k spécifiques pour chaque algorithme de réduction de l'espace des véhicules.

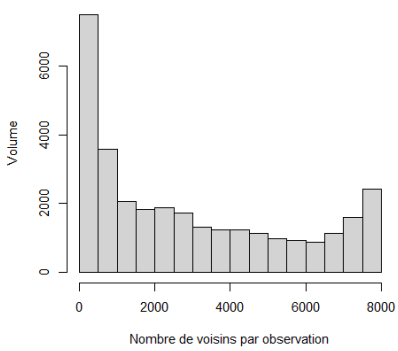
Choix des rayons et du nombre de voisins à cross-valider pour l'AFDM : Commençons par observer le nombre de voisins par observation pour des rayons d donnés :



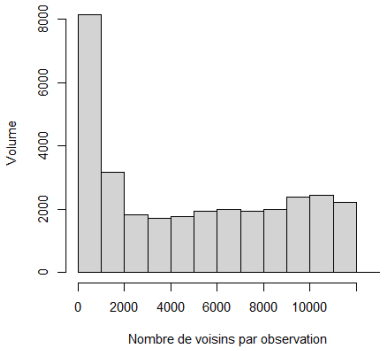
$d = 0.25$



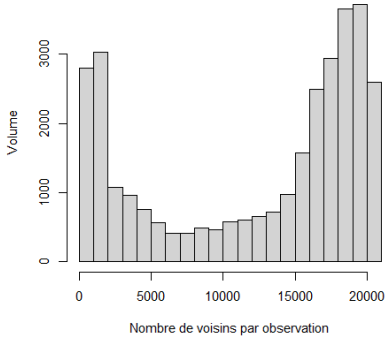
$d = 0.5$



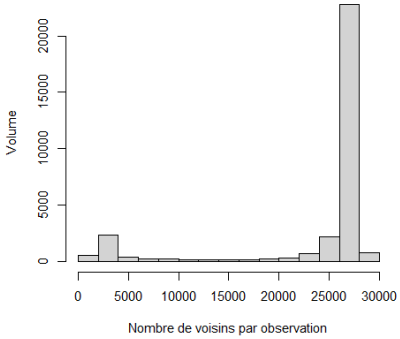
$d = 0.75$



$d = 1$



$d = 2$



$d = 5$

FIGURE 2.11 – Nombre de voisins par observation pour des rayons d donnés pour l'AFDM

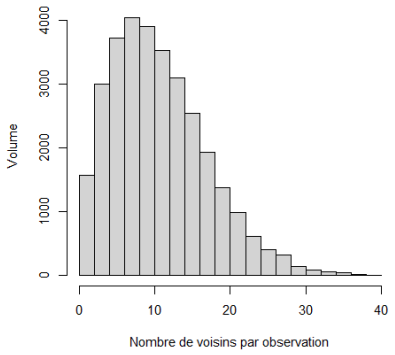
	Min	q_1	Médiane	Moyenne	q_3	Max
$d = 0.25$	1	48	164	312	394	1 964
$d = 0.5$	1	224	831	1 355	2 041	4 951
$d = 0.75$	1	557	2 201	2 933	5 038	7 974
$d = 1$	1	914	4 563	4 915	8 569	12 046
$d = 2$	1	3 997	15 851	12 331	18 572	20 773
$d = 3$	1	12 794	21 933	17 645	23 454	24 864
$d = 5$	5	25 994	27 041	23 764	27 495	28 977

TABLE 2.3 – Statistiques sur le nombre de voisins par observation pour des rayons d donnés pour l'AFDM

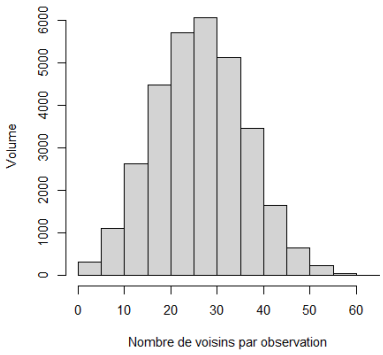
Comment interpréter les statistiques du tableau 2.3 ? Il faut comprendre que par exemple, si on prend un nombre de voisins k fixé tel que $k < 831$, la moitié des observations lissées seront identiques quel que soit le rayon d choisi supérieur à 0.5. De plus, on remarque que pour $d = 3$ (respectivement $d = 5$), 75% des observations ont au moins 12 794 voisins (respectivement 25 994), soit 41% des données (respectivement 83%). Il semble donc qu'utiliser un rayon supérieur à 3 est trop grand puisqu'il englobe une très grande part des données. Au contraire, pour $d = 0.25$, la moitié des points est reliée à moins de

164 voisins, ce qui est proportionnellement peu. Les valeurs de d choisies pour l'AFDM sont donc 0.5, 0.75, 1 et 2. On remarque que pour des rayons variant entre 0.5 et 2, le nombre de voisins par observation change beaucoup, ce qui permettra ainsi de tester différents intensité de lissage. Une fois les valeurs de d choisies, il faut choisir des valeurs de k qui conviennent à tous les rayons. Nous choisissons 100. 200. 500 et 1000.

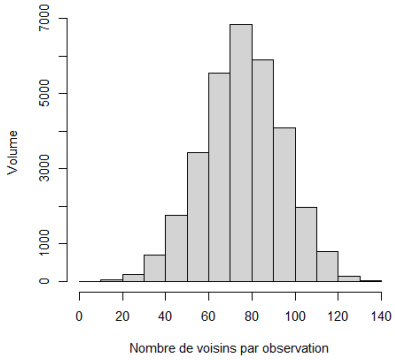
Choix des rayons et du nombre de voisins à cross-valider pour le t-SNE :



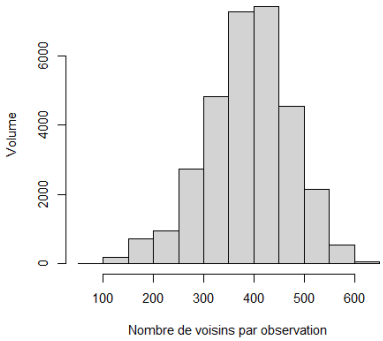
$d = 0.5$



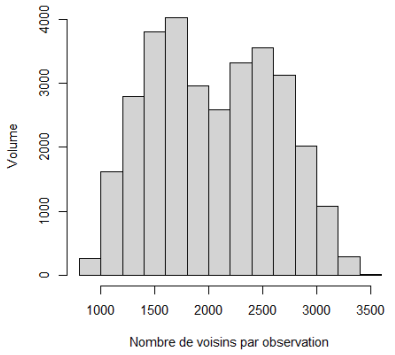
$d = 1$



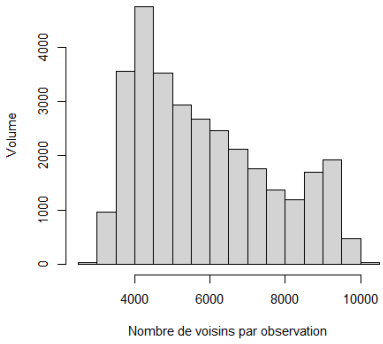
$d = 2$



$d = 5$



$d = 10$



$d = 15$

FIGURE 2.12 – Nombre de voisins par observation pour des rayons d donnés pour le t-SNE

	Min	q_1	Médiane	Moyenne	q_3	Max
$d = 0.5$	1	6	10	11	15	39
$d = 1$	1	20	27	27	34	63
$d = 2$	4	64	77	76	89	140
$d = 5$	99	337	394	389	446	647
$d = 7$	218	682	823	803	937	1334
$d = 10$	861	1 570	2 022	2 049	2 516	3 502
$d = 15$	2 948	4 335	5 498	5 873	7 140	10 163

TABLE 2.4 – Statistiques sur le nombre de voisins par observation pour des rayons d donnés pour le t-SNE

De même que pour l'AFDM, les rayons trop petits ($d = 1$ et $d = 0.5$) sont écartés parce qu'ils n'intègrent que très peu de voisins. Un rayon $d = 15$ semble lui trop grand puisqu'en moyenne un point a pour voisin 5873 observations, soit 19% des observations. Les rayons sélectionnés sont donc 2, 5, 7 et 10. Les valeurs de k correspondantes sont 100, 200, 500 et 1000.

On remarque d'abord que les deux algorithmes de réduction de dimension produisent des nuages de points assez différents, ce qu'on pouvait remarquer d'après les représentations graphiques Figure 2.9 et Figure 2.10 avec des distances bien différentes entre les deux méthodes. De plus, la répartition est différente. D'après les histogrammes décrivant le nombre de voisins par rayon donné, on se rend compte que l'AFDM crée un nuage de point très concentré et quelques autres points dispersés (d'où la répartition bimodale des histogrammes). Au contraire, le nuage de points issu du t-SNE semble réparti de façon plus uniforme : à mesure que le rayon s'élargit, les histogrammes restent unimodaux et plutôt symétriques.

2.6.2 Lissage basé sur l'adjacence

Le lissage par adjacence suppose de construire un graphe permettant de relier les observations les unes aux autres afin de définir les voisinages de chacun. Pour cela, il faut construire une première triangulation (ici le choix est porté sur la triangulation de Delaunay) puis d'écarter certains liens jugés aberrants.

2.6.2.1 Triangulation de Delaunay

Soit $X = \{x_1, ..., x_n\}$ un ensemble de points dans $\mathbb{R}^d$. On appelle une triangulation de X un ensemble d'arêtes reliant les points de X sans intersection et maximale (on ne peut pas rajouter d'arêtes). On appelle triangulation de Delaunay (notée $DT(X)$), une triangulation telle qu'il n'y a pas de points

de X à l'intérieur du cercle circonscrit de chacun des triangles de $DT(X)$.

Dans le cadre de cette étude, utiliser une triangulation de Delaunay est pertinente pour différentes raisons :

- La triangulation de Delaunay maximise le plus petit angle des triangles, ce qui, combiné avec le fait qu'aucun point ne se trouve à l'intérieur des cercles circonscrits, permet d'obtenir une triangulation qui reliera principalement des points proches.
- D'autres triangulations permettent de minimiser la longueur de ses segments, mais ils sont bien plus coûteux en temps de calcul. La triangulation de Delaunay semble un bon compromis, puisqu'elle est assez rapide à effectuer.
- Pour un échantillon donné, la triangulation de Delaunay est unique. Cela assure de pouvoir relancer cet algorithme (sur R dans le cas présent) et d'obtenir les mêmes résultats.

2.6.2.2 Réduction des liens de la triangulation

La triangulation de Delaunay crée de facto de nombreux liens entre les points, parfois même entre deux points particulièrement éloignés. Il paraît donc difficile de ne pas retraiter un certain nombre de ces liens afin de ne relier que des points ayant véritablement des similitudes. Pour cela, nous procédons en deux étapes :

- D'abord nous supprimons un certain pourcentage des liens les plus longs. En effet, on observe visuellement d'après la distribution des distances entre les points reliés par triangulation ci-dessous qu'il existe une petite quantité de liens dont la distance est particulièrement élevée. Il s'agit alors de déterminer le pourcentage idéal de liens les plus longs à éliminer. D'après les graphiques suivants (Figure 2.13), le choix s'est imposé entre 1% et 5% puisqu'au-delà on risque d'enlever beaucoup de liens (dont certains ne sont pas très longs). La décision a été prise de supprimer 1% des liens les plus longs permet de réduire fortement la distance maximale sans perdre la majeure partie de la triangulation. D'autant plus que dans le cas du t-SNE, on observe un second petit pic entre les quantiles à 95% et 99% qu'il semblait pertinent de ne pas éliminer. De plus, dans le cas de l'AFDM, éliminer 5% des liens aurait pour conséquence de supprimer des liens qui graphiquement ne semblent pas extrêmes.

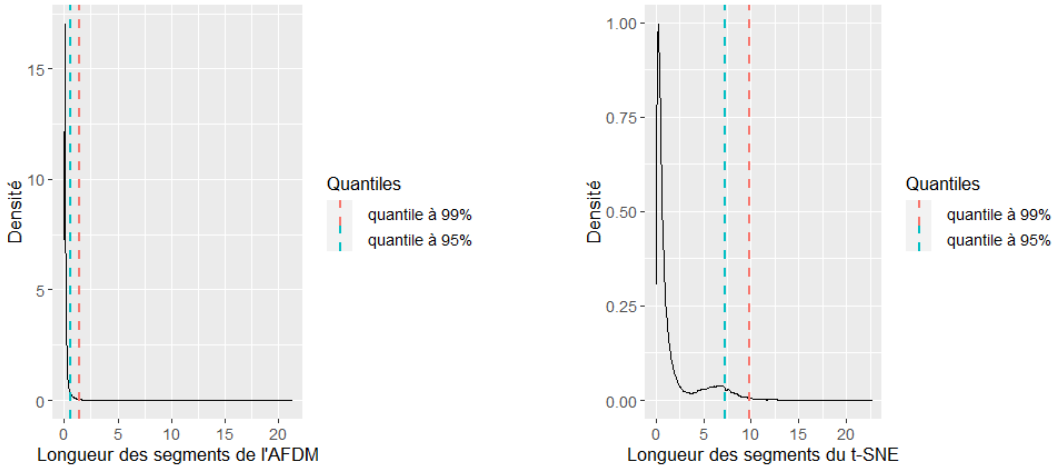
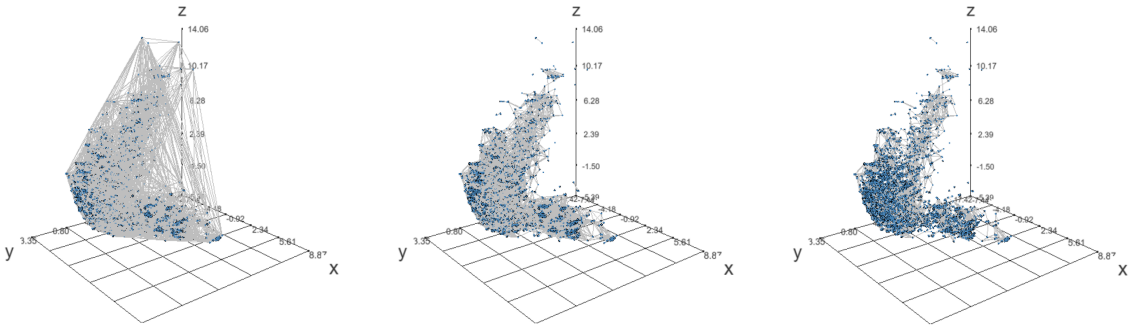


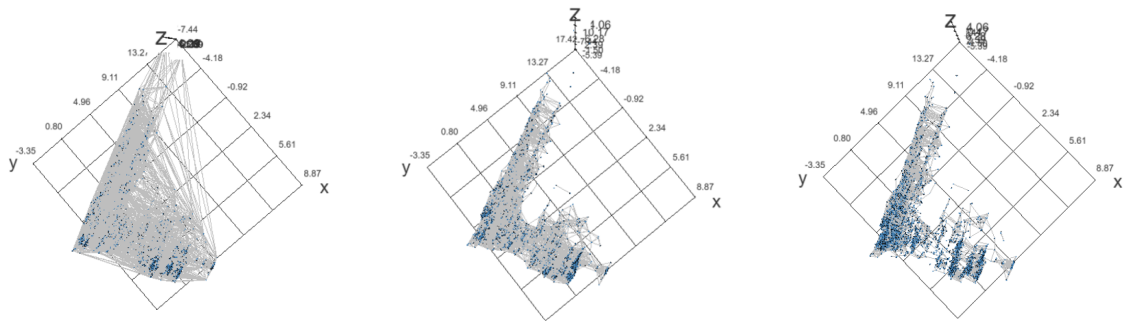
FIGURE 2.13 – Distribution de la longueur des liens de l'AFDM et du t-SNE

- Comme l'AFDM et le t-SNE sont construits indépendamment des résidus et des coûts observés, il faut s'assurer que les liens créés par la triangulation restent pertinents en termes de prédiction des résidus. Pour cela, 3 variables véhicules jugées pertinentes dans l'explication des coûts des sinistres sont sélectionnées, puis catégorisées si elles sont continues. Ici il s'agit des variables *prix du véhicule*, *segment du véhicule*, *marque du véhicule*, les trois variables véhicules les plus pertinentes à l'issue du GLM *benchmark*. Enfin, nous comparons les modalités de ces 3 variables pour chaque couple de points reliés. Si pour au moins deux des trois variables, les modalités diffèrent, alors nous considérons que le lien n'a pas lieu d'être entre ces deux points et nous le supprimons.

Voici les représentations graphiques des graphes de liens avant et après réduction des liens ainsi que des statistiques de base concernant le nombre de liens pour les deux nuages de points (issus du t-SNE et AFDM). N'ayant pas de moyen mathématique de savoir quels liens supprimer et quelle structure de graphe est la plus pertinente, nous ne pourrons faire que des remarques concernant la visualisation graphique des liens obtenus et juger en fonction de l'aspect du nuage de points.



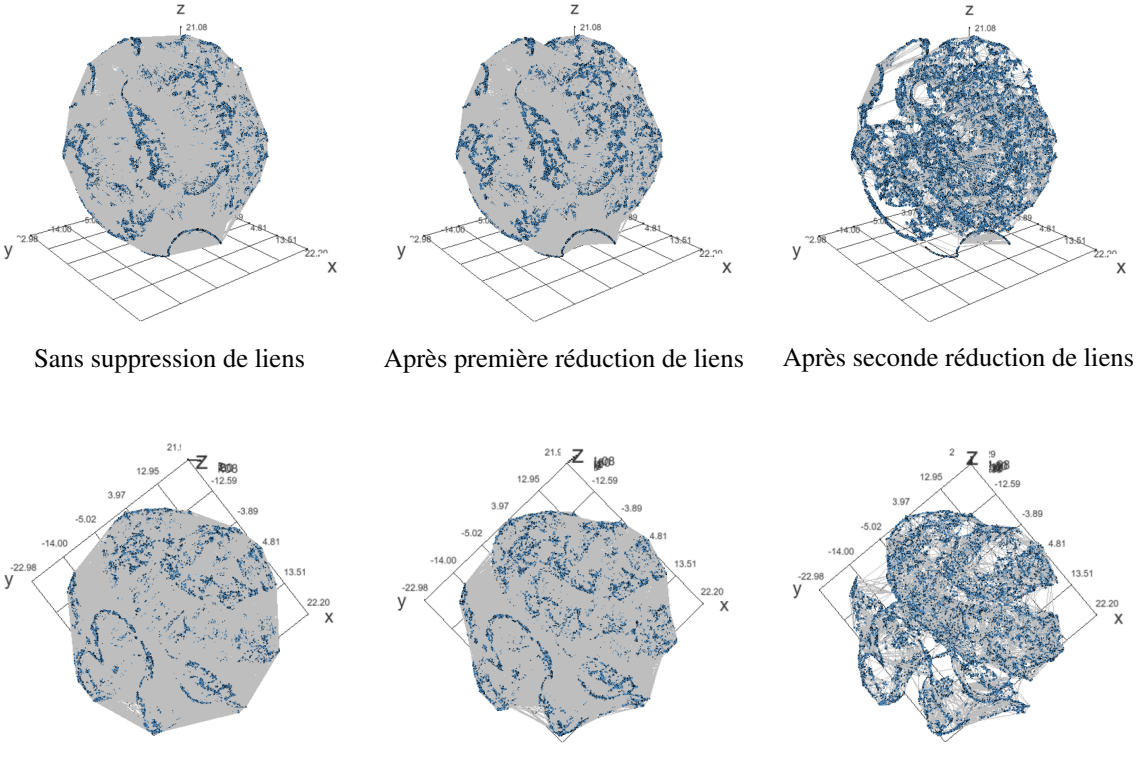
Sans suppression de liens | Après première réduction de liens | Après seconde réduction de liens



Sans suppression de liens | Après première réduction de liens | Après seconde réduction de liens

FIGURE 2.14 – Réduction de liens de l'AFDM (différents angles de vue)

Les représentations graphiques du graph de liens pour l'AFDM montrent que la réduction de liens semble pertinente : un certain nombre de liens particulièrement longs avaient été créés lors de la triangulation, il fallait en supprimer certains. Les deux étapes de réductions semblent avoir éliminé les liens les plus aberrants tout en gardant les liens les plus proches. A l'issue de la seconde réduction de liens, les deux sous-ensembles de points que l'on voit en vue de dessus semblent plus disjoints que dans les étapes précédentes.



Sans suppression de liens | Après première réduction de liens | Après seconde réduction de liens

FIGURE 2.15 – Réduction de liens du t-SNE

Graphiquement, on peut voir que la seconde étape de réduction de liens semble pertinente, pour le t-SNE aussi, parce qu'elle permet d'obtenir un graphe qui découvre la structure du nuage de points (sans pour autant supprimer trop de liens), ce qui n'était pas le cas après avoir retiré 1% des points les plus longs.

	min	q_1	médiane	moyenne	q_3	max
Avant réduction des liens	5	12	14	14.93	17	77
Après première réduction des liens	0	11	14	14.78	17	70
Après seconde réduction des liens	0	5	7	7.64	10	35

TABLE 2.5 – Statistiques du nombre de liens à chaque étape de la suppression des liens pour l'AFDM

	min	q_1	médiane	moyenne	q_3	max
Avant réduction des liens	4	10	13	14.65	17	236
Après première réduction des liens	4	10	13	14.5	17	225
Après seconde réduction des liens	0	1	3	3.09	4	37

TABLE 2.6 – Statistiques du nombre de liens à chaque étape de la suppression des liens pour le t-SNE

Le nombre de voisins à l'issue des deux étapes de réduction se réduit de façon conséquente. Par rapport au lissage par distance, le nombre de voisins moyen intervenant dans le lissage de chaque observation est assez faible. Cela ne pose cependant pas de problème majeur si l'on considère que la méthodologie utilisée est robuste. Les liens ainsi obtenus, bien qu'en faible quantité, sont alors jugés pertinents.

2.6.2.3 Lissage

Un lissage basé sur l'adjacence considère que pour chaque point étudié, les voisins intervenant dans le lissage sont ceux qui sont reliés au point en question à l'issue de la triangulation. Dans ce cas, on considère que deux points peuvent être considérés comme voisins même si leur distance n'est pas la plus faible. De même, deux points jugés proches par l'algorithme de réduction de dimension pourraient ne pas être reliés par la triangulation (dans le cas où ils différent selon certaines variables véhicules, comme cela a été expliqué précédemment).

Le résidu lissé est défini de la manière suivante :

$$r'_i = \alpha_i \left(\frac{\sum_{j \in V_i} r_j d_{ij}^{-P}}{\sum_{j \in V_i} d_{ij}^{-P}} \right) + (1 - \alpha_i) r_i$$

avec V_i l'ensemble des voisins du point i intervenant dans le lissage (c'est-à-dire les points pour lesquels la triangulation a créé un lien avec le point i mais qui n'ont pas été supprimés par la suite). Les paramètres de lissage par adjacence à optimiser sont :

- la puissance P attribuée à la distance
- le poids α relatif de l'ancien résidu par rapport à la moyenne des résidus des voisins : plus α est grand plus les voisins ont un poids important par rapport à la valeur du résidu non lissé. Les valeurs

0.2, 0.3, 0.4, 0.5, 0.6, 0.7 ont été testées, choisir des valeurs plus grandes ou plus petites impliquait de ne donner que très peu de poids soit au résidu non lissé soit au voisins, ce qui ne semble pas souhaitable. Cependant, on peut considérer que la valeur de ce paramètre ne devrait pas être fixe mais devrait dépendre du nombre de voisins et de leur proximité. En effet, si le point i a beaucoup de voisins et que ceux-ci sont proches alors ces derniers devraient avoir plus de poids que dans le cas où le point i a peu de voisins et que ceux-ci sont éloignés.
Ainsi, outre des valeurs fixes de α testées, on teste aussi une version non fixe. En effet, pour chacune des observations i, on calcule la valeur suivante :

$$l_i = \sum_{j \in V_i} d_{ij}^{-P}$$

l_i constitue, dans le calcul du résidu lissé, la partie du lissage des voisins correspondant à normaliser la moyenne des résidus des voisins. Si un point a de nombreux voisins et que ceux-ci sont proches, cette valeur sera grande. Au contraire, si un point a peu de voisins et qu'ils sont loin, cette valeur sera petite. Ensuite, on calcule les quantiles du vecteur $l = (l_1, \ldots, l_n)$ et on attribue la valeur de α selon la règle suivante :

Valeur de l_i	α correspondant
$\leq q_{10}$	0.1
$\in]q_{10}; q_{20}]$	0.2
$\in]q_{20}; q_{30}]$	0.3
$\in]q_{30}; q_{40}]$	0.4
$\in]q_{40}; q_{50}]$	0.5
$\in]q_{50}; q_{60}]$	0.6
$\in]q_{60}; q_{70}]$	0.7
$\in]q_{70}; q_{80}]$	0.8
$\in]q_{80}; q_{90}]$	0.9
$\geq q_{90}$	1

Schématiquement, voilà à quoi ressemblent les deux types de lissage pour un nuage de points donné. Les voisinages peuvent en théorie être très différents. Dans la troisième partie, ces deux méthodes seront comparées afin de déterminer à quel point elles sont semblables.

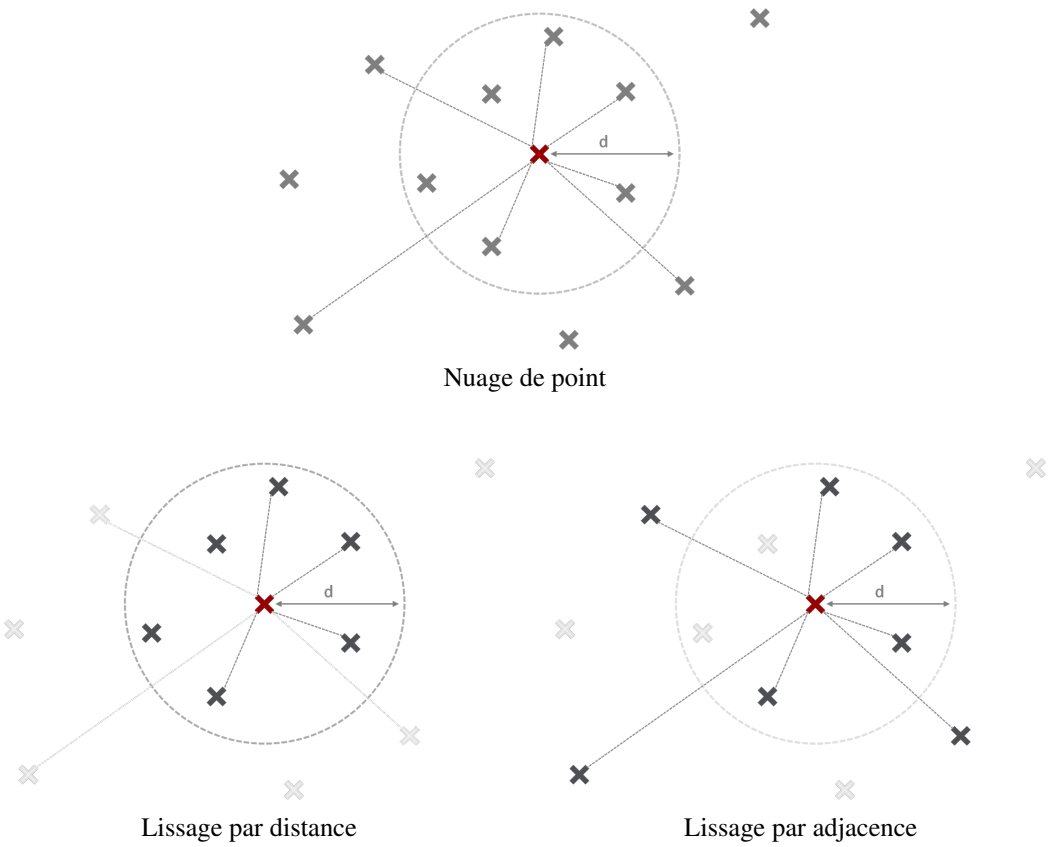


FIGURE 2.16 – Comparaison des voisins choisis par lissage

2.7 Création de véhiculier par classification des résidus

Une fois les résidus lissés, une dernière étape consiste à les classifier en fonction des variables véhicule. Ici encore, afin d'obtenir les véhiculiers les plus efficaces possibles, deux méthodes différentes ont été utilisées : les *Classification and Regression Trees* (CART) et les Evolutionnary Trees. Les performances de ces deux méthodes seront ensuite comparées au chapitre suivant (voir 3.3.2).

2.7.1 Arbre de classification et de régression

Les arbres de classification et de régression (CART) sont des algorithmes d'apprentissage supervisé. Le principe est de segmenter les données par itération en deux groupes appelés branches afin d'obtenir des sous-groupes plus homogènes relativement à la variable cible. Lorsque l'algorithme s'arrête, les dernières branches segmentées sont appelées des feuilles. Il existe deux types d'arbres : les arbres de régression qui ont pour cible une variable continue et les arbres de classification qui permettent de prédire une variable catégorielle (binaire ou multimodale). Dans le cadre de cette étude, la prédiction se portant sur les résidus du modèle GLM (qui constituent une variable continue), c'est donc un arbre de régression qu'il faut utiliser. Cependant, nous ne nous intéresserons pas aux prédictions de résidus de l'arbre (c'est-à-dire la moyenne des résidus par feuille) mais plutôt à la classification effectuée lors de la répartition des observations dans les feuilles finales.

Soient Y la variable à prédire et $X_1, ..., X_d$ les variables explicatives. Supposons que l'on ait n observations (x_1, y_1), ..., (x_n, x_n), avec $i \in \{1, ..., n\}$, $x_i = (x_{i1}, ..., x_{id})$ correspondant à la i-ème observation pour chacune des variables explicatives et y_i la i-ème observation de la variable cible. On note pour $\theta = (v_1, s_1, \ldots, v_{M-1}, s_{M-1})$ les paramètres à déterminer pour un arbre avec M feuilles finales, v_i la variable segmentée au i-ème nœud et s_i le seuil correspondant.

L'algorithme CART consiste à partitionner les variables en input $X_1, ..., X_d$ de façon à regrouper les observations de la variable prédite Y dans des groupes homogènes. A chaque étape de l'arbre, il s'agit donc de déterminer une variable à partitionner X_j et un seuil qui segmentent au mieux l'espace des variables prédictives. Il faut ainsi définir une mesure de l'homogénéité des branches créées à chaque étape ainsi que le gain d'information apporté par la partition des variables prédictives. Le package "rpart" est utilisé dans ce mémoire, celui-ci utilise l'impureté comme mesure d'homogénéité. Pour un nœud noté N et N_L et N_R les branches filles, l'impureté (notée I) est définie de la façon suivante :

$$I = V(N) - \left(\frac{|N_L|}{|N|} V(N_L) + \frac{|N_R|}{|N|} V(N_R) \right)$$

avec

$$V(N) = \sum_{j:x_j \in N} (y_j - \bar{y}_N)^2$$

$$V(N_L) = \sum_{j:x_j \in N_L} (y_j - \bar{y}_{N_L})^2$$

$$V(N_R) = \sum_{j:x_j \in N_R} (y_j - \bar{y}_{N_R})^2$$

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

$$\bar{y}_N = \frac{1}{|N|} \sum_{j:x_j \in N} y_j; \quad \bar{y}_{N_L} = \frac{1}{|N_L|} \sum_{j:x_j \in N_L} y_j; \quad \bar{y}_{N_R} = \frac{1}{|N_R|} \sum_{j:x_j \in N_R} y_j$$

A chacun des nœuds, plus l'impureté est grande, plus les observations qui la composent sont éloignées de sa valeur moyenne, plus la variance est grande et moins il est homogène. Un découpage efficace est celui qui permet donc de réduire l'impureté contenue dans les branches filles comparée à celle du nœud père, donc il s'agit de maximiser I à chaque étape.

Un autre paramètre à étudier lors de l'utilisation d'un CART est sa complexité. En effet, plus un arbre a de branches, plus il est complexe. Un arbre trop petit ne serait pas efficace puisqu'il aurait tendance à sous-apprendre les données. De même, un arbre trop grand ne serait pas meilleur puisqu'il pourrait sur-apprendre les données. Il s'agit donc (ici comme pour le modèle GLM) de trouver un juste équilibre entre la performance du modèle et sa complexité. L'idée étant qu'un arbre maximal aura une variance élevée (il comporte de nombreuses feuilles qui contiennent parfois très peu d'observations) mais un biais faible ; alors qu'un arbre minimal aura une variance faible (une seule modalité) mais un biais fort (les prédictions sont très mauvaises). Il s'agit donc de sélectionner un modèle intermédiaire. La méthode la plus souvent utilisée pour prendre en compte ce risque de surapprentissage et sous-apprentissage est d'"élaguer" cet arbre, ce qui correspond à :

- Construire un arbre maximal

- Afin de déterminer la taille optimale, on calcule non pas l'impureté seule, mais l'impureté à laquelle on additionne le nombre de feuille multipliée par un coefficient de complexité noté α correspondant au coût d'ajout d'un nouveau nœud à l'arbre. Cette mesure sera notée "coût" et définie de la façon suivante :

$$\text{coût}_\alpha(T) = \sum_{e=1}^{|T|} \sum_{i:x_i \in N_e} (y_i - \hat{y}_{N_e})^2 + \alpha|T|$$

  avec $|T|$ le nombre de feuilles de l'arbre T, $(N_e)_{e\in\{1,\dots,|T|\}}$, les feuilles de l'arbre T, $\hat{y}_{N_e}$ la prédiction de y dans la feuille N_e. Le premier terme correspond à la somme des erreurs au carré dans les feuilles terminales de l'arbre (ce terme diminue quand la taille augmente), le second sert de régularisation. De plus, plus α est grand, plus ce seront des arbres courts qui minimiseront la fonction "coût".

- Pour différentes valeurs de α, notées $(\alpha_1, \dots, \alpha_k)$, la fonction "coût" est calculée pour l'ensemble des sous-arbres issus de l'arbre maximal. On détermine les arbres $T_{\alpha_1}, \dots, T_{\alpha_1}$ correspondant puis par cross-validation, le meilleur arbre est choisi ainsi que la valeur de α correspondante.

- Relancer l'arbre en ajoutant la contrainte du coefficient α optimal.

Notons que le package "rpart" utilisé dans ce mémoire, permet d'ajouter une pondération au CART. Celui-ci sera systématiquement pondéré par le vecteur de prédictions du GLM qui a permis d'extraire l'effet véhicule. Cela permet de donner un poids plus important aux observations dont le coût prédit est grand, puisqu'il paraît très important de les prédire correctement et de ne pas les sous-estimer.

Afin de rendre notre algorithme plus pertinent, les paramètres de l'arbre sont optimisés par cross-validation. Ici, le paramètre le plus important à faire varier est le nombre d'observations dans les feuilles finales (les autres paramètres comme la longueur de l'arbre étant déjà optimisés par l'élagage). En effet, s'il y a trop peu d'observations dans une feuille, il y a un risque de surapprentissage. De plus, le véhiculier créé à l'issue de cette étude doit contenir suffisamment de classes possibles afin de ne pas perdre trop d'information (notamment parce qu'on le compare à un modèle contenant tout ou partie des variables véhicule à notre disposition). La contrainte du nombre d'observations par feuille ne devra donc pas être supérieure à 5% d'observations par feuille (ce qui limite au maximum à 20 classes dans le véhiculier, ce qui est assez peu). En d'autres termes, on espère limiter la perte d'information relative au passage de 12 variables véhicule à une seule en autorisant la construction d'arbres potentiellement grands. Le choix est donc fait de tester les contraintes de 1% et 2% d'observations au minimum par feuille finale. Mais des modèles qui suppriment cette contrainte seront aussi testés, puisque l'élagage permet déjà, d'éviter le surapprentissage.

2.7.2 Evolutionnary Trees

Les CART fonctionnent par partitionnement *forward* par itérations en maximisant l'homogénéité des branches filles, mais sans possibilité de retour en arrière, ce qui signifie que le partitionnement à un point de l'arbre ne tient pas compte des partitionnements futurs. Cette méthode, bien que rapide à exécuter, ne permet donc d'obtenir que des arbres optimaux localement et peuvent donc parfois être

très éloignés de la solution optimale. Au contraire, les Evolutionary Trees permettent d'évaluer l'impact d'une segmentation non pas seulement sur les deux branches créées, mais aussi plus profondément dans l'arbre et ainsi de trouver la solution globale. Bien que le nombre d'arbres possible soit très grand, cette méthode permet d'éviter ce problème.

Le principe de l'algorithme d'Evolutionary Tree est le suivant : à chaque étape de l'arbre, on réalise des opérations de variations qui peuvent être :

- Un split : sur un nœud terminal choisi au hasard, on effectue une segmentation
- Un élagage : un nœud intermédiaire est choisi puis élagué (on supprime les nœud issus de celui-ci)
- Mutation de règle de segmentation majeure : un nœud intermédiaire est sélectionné aléatoirement et la règle de décision (v_r, s_r) est changée
- Mutation de règle de segmentation mineure : un nœud intermédiaire est sélectionné aléatoirement et le seuil de segmentation s_r est modifié (mais la variable segmentée reste inchangée)
- Croisement : Cette étape correspond à un échange aléatoire de deux sous-arbres entre deux arbres.

A l'issue de chaque opération de variation, la meilleure solution est gardée.

La mesure permettant de choisir le meilleur arbre est appelée *evaluation function*, celle-ci est définie de la façon suivante :

$$\begin{aligned} \text{evaluation function} &= \text{loss}(Y, f(X, \theta)) + \text{comp}(\theta) \\ &= n \log(MSE(Y, f(X, \theta)) + \alpha \times 4(M+1) \log(n) \end{aligned}$$

avec M+1 le nombre de paramètres et α un paramètre de réglage qui pénalise la grandeur de l'arbre. Ici encore, il s'agit de trouver un équilibre entre la capacité de prédiction du modèle et sa complexité afin d'obtenir de bonnes performances sur la base d'entraînement sans sur-apprentissage et pouvoir faire de bonnes prédictions sur une autre base de données.

Cette méthode d'arbre étant plus complexe que les arbres classiques, son principal inconvénient est qu'elle est bien plus coûteuse en temps de calcul.

Chapitre 3

Evaluation de la pertinence du véhiculier

Une fois les différentes étapes de la construction du véhiculier effectuées, il s'agit à présent d'en évaluer la pertinence et l'efficacité, à chacune des étapes de sa construction. En effet, le choix d'utiliser au moins deux alternatives à chaque étape permet ainsi de les comparer et de déterminer la meilleure option. Mais nous avons aussi mesuré la pertinence même de chaque étape de la construction du véhiculier :

- La pertinence de la pénalisation fused lasso sera mesurée en comparant ses performances avec celles d'un modèle non pénalisé ou bien utilisant une pénalisation lasso traditionnelle.
- La pertinence du lissage sera mesurée en comparant les performances d'un véhiculier avec et sans lissage.
- La pertinence du véhiculier sera mesurée en comparant les performances d'un modèle avec le meilleur véhiculier (en fonction de ses performances sur la base de validation) et les modèles *benchmark* définis précédemment (voir 2.2).

Ici et dans la suite du mémoire, les mesures de performances des différents modèles produits sont le Mean Squared Error (MSE), l'Area Between Curve (noté ABC) et l'Akaike Information Criterion (AIC) dont les descriptions sont détaillées en Annexe B. Notons que ce qu'on appellera Area Under Curve dans la suite de ce mémoire correspondra à la mesure basée sur l'indice de Gini traditionnel et construite en Annexe B.2. Cette mesure alternative correspond à l'aire entre la courbe de Lorenz des prédictions d'un modèle et celle des observations correspondantes. Ainsi, dans ce cas, plus faible est l'ABC, meilleur est le modèle.

Les MSE et ABC sont déterminées sur la base de données de test pour les modèles benchmark (sur celle de validation pour les véhiculiers) alors que l'AIC est déterminée sur la base de données d'entraînement. Utiliser plusieurs statistiques rend la comparaison plus précise que si l'on se contente d'une seule, car ces trois mesures se distinguent les unes des autres et permettent d'évaluer différents aspects d'un modèle. En effet, le MSE est une simple moyenne des erreurs, ce qui permet de déterminer quel modèle propose les meilleurs prédictions en moyenne. L'ABC permet de mesurer le fait que la distribution des prédictions est plus ou moins similaire à la distribution des observations. Enfin, l'AIC permet d'évaluer l'ajustement du modèle avec les observations d'entraînement tout en prenant en compte la complexité du modèle et en pénalisant par le nombre de paramètres à déterminer.

3.1 Pertinence de la pénalisation fused lasso

Dans un premier temps, il faut évaluer la pertinence de l'utilisation d'une pénalisation fused lasso. En effet, à cause de l'ajout d'un terme supplémentaire dans le modèle, celui-ci devient bien plus complexe puisqu'une étape supplémentaire (par rapport au modèle sans pénalisation) de cross-validation du paramètre de pénalisation devient alors nécessaire et parce que de nouvelles contraintes sont ajoutées. Cela rend ce modèle GLM bien plus coûteux en temps de calcul par rapport à un modèle GLM classique. La question se pose donc de savoir si cela s'avère bénéfique en termes de prédiction. Pour cela, les deux modèles (avec toutes ou seulement 4 variables véhicule) sont exécutés (sur la base de données regroupant les données d'entraînement et de validation) ainsi qu'un GLM avec une pénalisation lasso standard et un GLM sans pénalisation (mais avec une sélection de variables backward). Les résultats sont présentés ci-dessous.

Modèle	**Mean Squared Error** (Base de données test)	**ABC** (Base de données test)	**AIC** (Base de données d'entraînement)
Modèle GLM avec • toutes les variables non véhicule • toutes les variables véhicule • pénalisation fused lasso	90 828.46	0.244 788	552 387.3
Modèle GLM avec • toutes les variables non véhicule • toutes les variables véhicule • pénalisation lasso	91 046.69	0.246 278	552 778.3
Modèle GLM avec • toutes les variables non véhicule • toutes les variables véhicule • sélection de variable backward	91 075.05	0.246 328	552 517.0
Modèle GLM avec • toutes les variables non véhicule • 4 variables véhicule • pénalisation fused lasso	91 002.94	0.246 008	552 473.8
Modèle GLM avec • toutes les variables non véhicule • 4 variables véhicule • pénalisation lasso	91 086.11	0.246 518	552 582.4
Modèle GLM avec • toutes les variables non véhicule • 4 variables véhicule • sélection de variable backward	91 348.1	0.248 188	552 598.2

TABLE 3.1 – Comparaison des performances des différents modèles GLM benchmark

Dans le modèle où l'on utilise toutes les variables véhicule ou bien seulement 4, les performances du GLM avec pénalisation (qu'elle soit lasso ou fused lasso) sont meilleures que celles du GLM avec sélection de variables, quelle que soit la mesure sélectionnée. Cela permet d'abord d'en conclure la pertinence de l'ajout d'un terme de pénalisation afin d'éviter le surapprentissage. De plus, la pénalisation fused lasso fait elle-même mieux que la pénalisation lasso standard. Ainsi ce modèle permet à la fois de mieux prédire en moyenne les coûts de sinistre (d'après le MSE) mais aussi de mieux prédire la distribution des sinistres et de discriminer les sinistres graves des autres (d'après l'ABC). Le modèle fused lasso avec seulement 4 variables véhicule fait même mieux que le modèle lasso avec toutes les variables véhicule.

De plus, on peut étudier l'intérêt de regrouper des modalités d'origine de variables en observant la pertinence des nouvelles modalités créées à l'issue du regroupement grâce à la pénalisation fused lasso et en les comparant aux modalités non regroupées dans le modèle lasso. On peut d'abord vérifier que les regroupements de modalités effectués sur la base de données d'entraînement demeurent pertinents sur la base de données de tests.

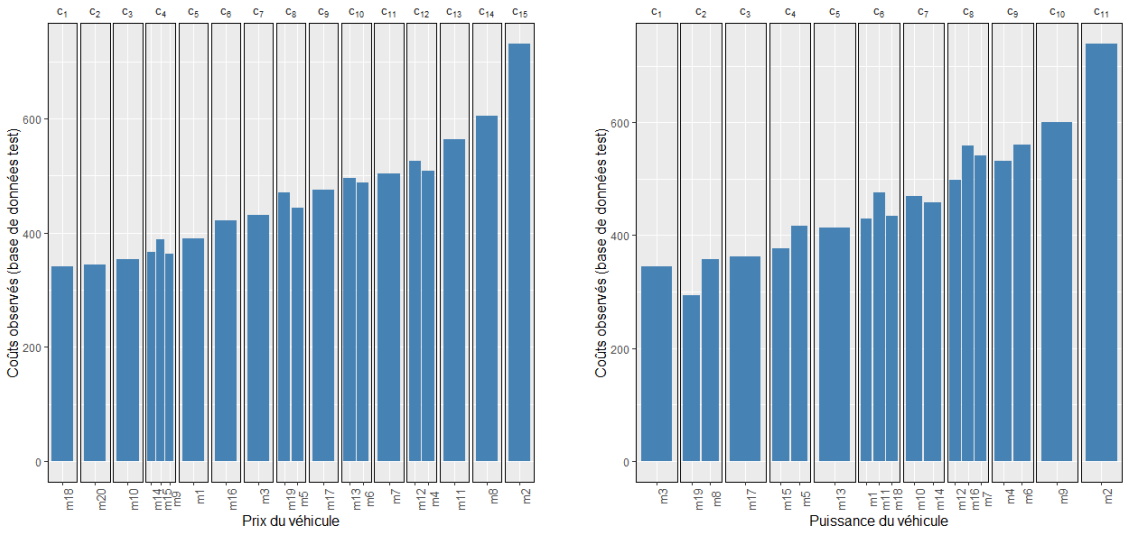


FIGURE 3.1 – Coûts observés en fonction des modalités du prix et de la puissance

Dans les graphiques précédents, chaque bâton représente une modalité de chaque variable, numérotées m_1,m_2 ... Les regroupements effectués par la pénalisation fused lasso correspondent à un bloc de bâtons notés c_1, c_2... Par exemple, pour la variable "marque", l'algorithme a regroupé les modalités m_9, m_{14} et m_{15} dans la nouvelle modalité c_4. Au sein d'un regroupement de modalité, les bâtons restent dans le même ordre de grandeur et varient faiblement. Ainsi ce graphique permet de vérifier que ces regroupements restent pertinents sur la base de données test puisque les modalités regroupées ont des coûts moyens similaires. En d'autres termes, le modèle fused lasso, construit sur la base d'entraînement, a réussi à regrouper des modalités qui restent similaires sur la base de test.

On peut aussi comparer les performances de chacune des variables en comparant les prédictions des modalités d'origine (dans le modèle lasso) à celles qui sont regroupées lors du modèle fused lasso.

FIGURE 3.2 – Prédiction des coûts de sinistres en fonction des modalités de la marque et de la cylindrée

Dans les graphiques précédents, pour chaque modalité des variables ”marque” et ”cylindrée”, le ratio entre les coûts observés et la prédiction (sur la base de données test) pour les modèles lasso et fused lasso sont calculés. Ainsi, plus ce ratio se rapproche de 1, meilleure est la prédiction. Un ratio de 1.1 ou de 0.9 signifie que les erreurs de prédictions sont de 10% en moyenne. On remarque ainsi que le modèle fused lasso est meilleur parce que les prédictions sont en général plus proches de 1 que pour le lasso. Pour toutes modalités de la variable ”cylindrée” les prédictions du modèle fused lasso ont une erreur moyenne inférieure à 5%, ce qui montre l'efficacité de ces regroupements.

De plus, pour certaines modalités, le modèle avec pénalisation lasso se montre particulièrement peu efficace : pour les m_{19} et m_{37} de la variable ”marque” par exemple, le ratio entre coût observé et prédiction est de 1.84 et 0.31 respectivement (alors qu'il est de 1.13 et 1.05 pour le modèle fused lasso), ce qui signifie que l'écart moyen entre la prédiction et le coût observé pour ces deux modalités est de 84% d'une part et 69% de l'autre (comparé à 13% et 5% pour le modèle fused lasso), ce qui est considérable. Cela s'explique par le fait que ces modalités ne représentent qu'environ 1% des observations, mais cela pose tout de même des problèmes dans le cadre d'un modèle de tarification d'assurance automobile. En effet, un assureur doit être capable de tarifer correctement tous les véhicules de son portefeuille, y compris ceux dont l'exposition est faible. De plus, pour toutes les modalités de ces deux variables, le quotient entre les coûts observés et prédits est systématiquement plus proche de 1 pour le modèles fused lasso que pour le modèle lasso.

Pour confirmer cette impression observée sur les variables ”marque” et ”cylindrée”, nous avons calculé pour chaque observation le quotient entre le coût observé et le coût prédit puis en avons déduit le pourcentage d'observations pour lesquelles les erreurs de prédictions (en valeur absolue) sont supérieures à 50% pour les deux modèles. Pour le modèle fused lasso, 20% des observations ont une erreur de prédiction supérieure à 50% alors que c'est 26% des observations pour le modèle lasso.

A présent, il s'agit de comprendre en quoi les prédictions des modèles fused lasso et lasso diffèrent

afin de comprendre pourquoi le modèle fused lasso a de meilleures performances.

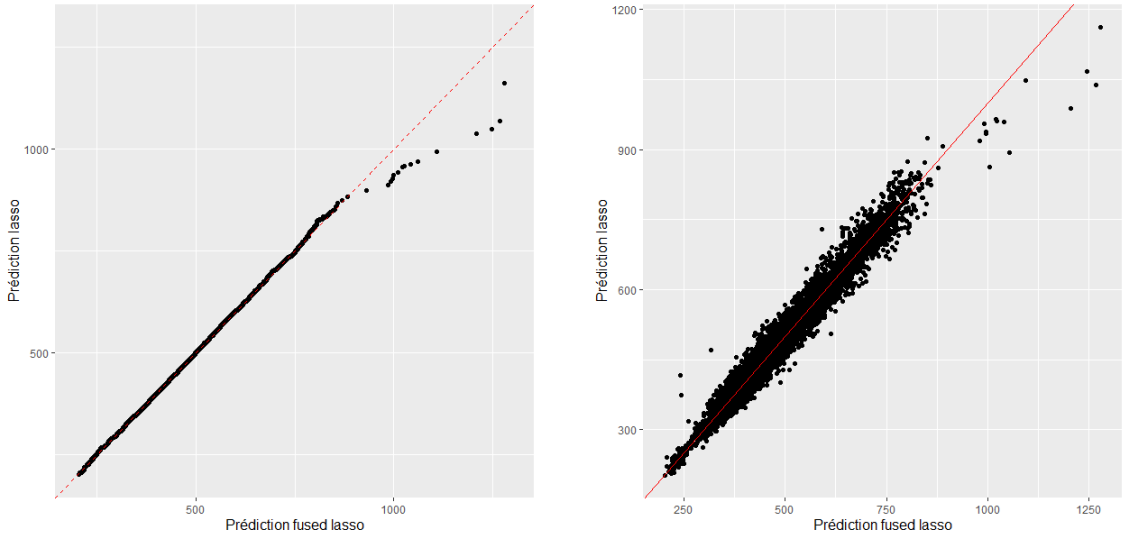


FIGURE 3.3 – Nuage de point et QQ Plot des prédictions des modèles lasso et fused lasso sur la base de test

Ici sont représentées les prédictions du modèle lasso en ordonnée et celles du fused lasso en abscisse. D'abord, il n'est pas surprenant de remarquer la forte corrélation entre ces deux variables puisque ce sont des modèles dont les fonctions à minimiser sont très proches. Il est plus intéressant d'étudier en quoi les prédictions des deux modèles diffèrent. Même si les deux modèles semblent identifier les mêmes sinistres en queue de distribution, ce sont principalement les montants de ces prédictions concernant ces sinistres au coût élevé qui diffèrent. Le modèle fused lasso a tendance à prédire un coût plus élevé pour ces sinistres que le modèle lasso. Il s'agit à présent de se demander quels sont les coûts réels correspondant.

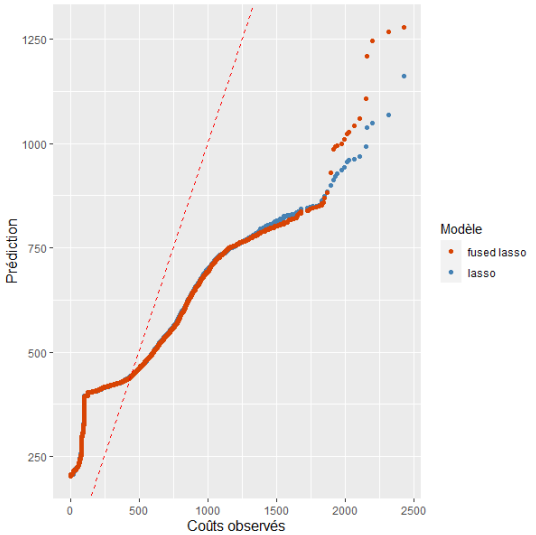


FIGURE 3.4 – QQ plot des prédictions des modèles lasso et fused lasso en fonction des coûts observés sur la base de test

Le graphique ci-dessus (Figure 3.4) montre les quantiles des coûts prédits des deux modèles en fonction des quantiles des coûts observés. Ce graphique permet donc d'étudier la distribution des prédictions

(tout comme le mesure l'ABC). Dans ce cas, un modèle parfait aurait des quantiles confondus avec la droite rouge $y = x$. On remarque tout d'abord que les modèles lasso et fused lasso ne semblent pas parfaits puisque les graphiques des quantiles sont assez éloignés de cette droite. De plus, on remarque que ces deux modèles ont tendance à surestimer les sinistres dont le coût est faible (parce que les points du QQ plot sont au-dessus de la droite) et à sous-estimer ceux dont le coût est élevé. Ce graphique permet aussi de confirmer les différences observées entre les deux modèles. Pour des sinistres dont le coût observé est relativement faible (inférieur à 1 000 euros), les deux modèles proposent des prédictions très similaires. Mais pour les sinistres plus graves, les deux modèles diffèrent. Le modèle lasso a de meilleures prédictions autour de 1 500 euros mais au-delà, c'est le modèle fused lasso qui est meilleur (puisque les points sont plus proches de la droite $y = x$).

Un autre moyen de vérifier l'hypothèse selon laquelle les modèles lasso et fused lasso diffèrent principalement en queue de distribution, consiste à découper le Mean Squared Error de ces deux modèles afin d'identifier sur quelles parties les modèles diffèrent. Soit $(y_i)_{i\in\{1,\dots,n\}}$ le coût observé du sinistre i et $(\hat{y}_i)_{i\in\{1,\dots,n\}}$ la prédiction correspondante, pour chaque intervalle de coût I (prenant les valeurs $[0, 1000]$, $]1000, 1500]$ et $[1500, \infty[$) on calcule :

$$\frac{1}{n}\sum_{i:y_i\in I}(y_i - \hat{y}_i)^2$$

Modèles	MSE (coûts observés inférieurs à 1 000)	MSE (coûts observés entre 1 000 et 1 500)	MSE (coûts observés supérieurs à 1 500)	TOTAL
Modèle fused lasso	68 886	14 522	7 420	90 828
Modèle lasso	69 070	14 449	7 528	91 047
Différence	184	-73	108	219

TABLE 3.2 – MSE des modèles lasso et fused lasso décomposé

Ce tableau permet de confirmer les remarques précédentes : les prédictions du modèle fused lasso sont meilleures sur l'intervalle $[1500, \infty[$ alors que c'est l'inverse sur l'intervalle $[1000, 1500]$. Cependant, on remarque que sur l'intervalle $[1000, \infty[$ la différence de MSE n'est pas très importante, ces différences se compensent en partie. La majeure partie de la différence de MSE s'explique par des différences de prédictions des sinistres attritionnels qui correspondent à la majorité des sinistres (94% des sinistres ont un coût inférieur à 1 000 euros) et des coûts (86% du coût total).

Conclusion partielle : La pénalisation fused lasso remplit donc bien son rôle en regroupant des modalités similaires afin de rendre ses prédictions plus robustes et plus pertinentes. Le modèle fused lasso avec 4 variables fait même mieux que la version lasso standard avec toutes les variables véhicule. Cette pénalisation est d'autant plus utile quand il s'agit de déterminer des coefficients pour des modalités dont le nombre d'observations est faible : le regroupement de modalités permet de limiter la forte volatilité liée à ces faibles expositions.

Cependant, la version avec une pénalisation fused lasso est très coûteuse en temps de calcul puisqu'elle est plus complexe et nécessite une validation croisée du paramètre de pénalisation. Il a fallu environ 5 heures de calcul pour faire tourner ce modèle sur une base de données relativement petite (19 variables et environ 38 000 observations). Les GLM avec pénalisation lasso ou sélection de variable quant à eux sont moins complexe et n'ont nécessité qu'une heure de calcul. Bien que la pénalisation fused lasso permettent à la fois de regrouper les modalités (ce que permettent pas de faire les modèle lasso et celui avec sélection backward) et de sélectionner des variables, étant données la faible différence de performances, il faut conclure que l'avantage de cette pénalisation fused lasso dépend grandement de l'usage qui en est fait. En effet, celle-ci se montre pertinente dans le cas d'un modèle pour lequel une première sélection de variables a priori est effectuée, comme c'est le cas du modèle avec toutes les variables non véhicule et seulement 4 variables véhicules. Le Mean Squared Error de ce modèle reste inférieur au modèle avec pénalisation lasso avec 4 variables mais aussi à celui qui utilise toutes les variables véhicule c'est-à-dire avec 8 variables supplémentaires.

3.2 Évaluation du lissage

L'évaluation du lissage peut se faire sous deux aspects : d'une part, on peut se demander lequel des deux types de lissage (entre adjacence et distance) est le plus pertinent. D'autre part, on peut aussi mesurer l'efficacité en soi du lissage en le comparant à des véhiculiers construits sur des résidus qui n'ont pas été lissés. Commençons par analyser à quel point ces deux lissages sont différents.

3.2.1 Comparaison des deux types de lissages

Avant de mesurer les performances des deux types de lissages, on pourrait d'abord se demander en quoi ils se ressemblent et si les voisinages qui entrent en compte dans les deux cas sont différents. Une première étape consiste à mesurer le nombre de liens dont la distance est inférieure à un certain rayon pour l'AFDM et le t-SNE.

Rayon	0.5	0.75	1	2
Pourcentage des liens AFDM	97%	98%	99%	100%

TABLE 3.3 – Pourcentage de liens de la triangulation issus de l'AFDM dont la distance est inférieure à un rayon donné

Pour l'AFDM, on n'observe que très peu les cas pour lesquels la triangulation a créé des liens entre deux points très éloignés l'un de l'autre. Cela tient dans la définition et la construction de cette triangulation de Delaunay : comme il n'y a aucun point à l'intérieur du cercle circonscrit de chacun des triangles, on peut imaginer que les points sont reliés à des voisins proches (bien que cette méthode de triangulation ne soit pas celle qui minimise les longueurs des segments). Ainsi, si on prend un rayon $d = 2$, cela signifie que tous les voisins d'un point pour le lissage par adjacence sont dans la boule de rayon 2, donc que tous les voisins du lissage par adjacence sont candidats pour être aussi les voisins du lissage par

distance. Cependant, cela ne signifie pas que les deux lissages seront identiques : dans la boule de rayon 2, le lissage par distance ne choisira que les k voisins les plus proches alors que le lissage par adjacence ne choisira que ceux qui sont reliés. Ce n'est pas une preuve que les voisins sont les mêmes dans les deux lissages, mais un pourcentage faible aurait prouvé que les voisinages sont différents.

Rayon	2	5	7	10
Pourcentage des liens t-SNE	91%	98%	99%	100%

TABLE 3.4 – Pourcentage de liens issus du t-SNE dont la distance est inférieure à un rayon donné

De même, pour le t-SNE, 91% des liens construits lors de la triangulation sont dans un rayon inférieur à 2, ce qui signifie qu'a priori, peu de voisins du lissage par adjacence sont en dehors du périmètre du lissage par distance.

De plus, on peut vérifier le nombre de voisins choisis par le lissage par distance qui sont aussi dans le voisinage du lissage par adjacence et vice versa pour les deux modèles :

d	0.5				0.75			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	8.05% [1]	4.32%	1.97%	1.16%	7.79%	4.06%	1.77%	0.97%
d	1				2			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	7.70%	3.96%	1.68%	0.90%	7.57%	3.85%	1.56%	0.80%

TABLE 3.5 – Pourcentage de liens de lissage par distance participant aussi au lissage adjacence (pour l'AFDM)

d	0.5				0.75			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	96.2% [2]	96.9%	97.1%	97.1%	97.3%	98.1%	98.4%	98.5%
d	1				2			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	97.9%	98.8%	99.2%	99.2%	98.4%	99.4%	99.9%	100%

TABLE 3.6 – Pourcentage de liens de lissage par adjacence participant aussi au lissage distance (pour l'AFDM)

[1] Cela signifie qu'en moyenne, 8% des voisins d'un point pour le lissage par distance pour $d = 0.5$ et $k = 100$ sont aussi les voisins de ce même point pour le lissage par adjacence.

[2] Cela signifie qu'en moyenne, 96% des voisins d'un point pour le lissage par adjacence sont aussi les voisins de ce même point pour le lissage par distance pour $d = 0.5$ et $k = 100$.

d	2				5			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	3.74%	3.70%	3.70%	3.70%	2.90%	1.50%	0.79%	0.78%
d	7				10			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	2.90%	1.49%	0.62%	0.39%	2.90%	1.49%	0.61%	0.31%

TABLE 3.7 – Pourcentage de liens de lissage par distance participant aussi au lissage adjacence (pour le t-SNE)

d	2				5			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	91.28%	91.31%	91.31%	91.31%	93.81%	96.71%	98.16%	98.17%
d	7				10			
k	100	200	500	1000	100	200	500	1000
Pourcentage de liens	93.81%	96.71%	98.64%	99.16%	93.81%	96.71%	98.68%	99.62%

TABLE 3.8 – Pourcentage de liens de lissage par adjacence participant aussi au lissage distance (pour le t-SNE)

Finalement, la grande majorité des voisins du lissage par adjacence sont inclus dans l'ensemble des voisins par distance mais l'inverse n'est pas vrai. En effet, le lissage par adjacence crée assez peu de liens (8 liens en moyenne par observations pour l'AFDM, 3 pour le t-SNE), et ceux-ci sont pour la quasi-totalité dans un rayon proche du point observé. Cependant, pour le lissage par distance, il y a en général plus de voisins étudiés (puisque les paramètres de rayon d et le nombre maximum de voisins k ont été choisis de sorte à avoir un nombre non négligeable de voisins, voir les tables 2.3 et 2.4) alors que pour le lissage par adjacence, nous n'avions pas le choix de paramétrer le nombre de liens lors de la triangulation.

Ainsi, le lissage par adjacence peut être quasiment considéré comme une version sélective du lissage par distance : parmi les k plus proches points dans le rayon d, le lissage par adjacence rajoute une sélection supplémentaire et ne choisit que les points reliés à l'issue de l'étape de triangulation. La comparaison des deux lissages permettra donc de répondre à la question suivante : vaut-il mieux lisser un point avec de nombreux voisins (comme c'est le cas pour le lissage par distance) ou bien seulement avec quelques-uns bien choisis (comme le fait le lissage par adjacence) ?

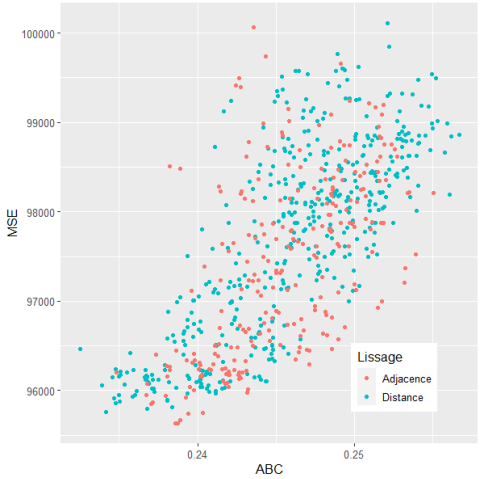


FIGURE 3.5 – Graphique du ABC et du MSE des modèles véhiculiers selon la méthode de lissage

Ce graphique permet de visualiser le MSE et le ABC des différents véhiculiers par méthode de lissage. Plus le MSE et l'ABC sont faibles, meilleur est le modèle. Ainsi, les meilleurs modèles sont ceux qui se trouvent en bas à gauche du graphique. Ici il paraît assez difficile de distinguer un modèle meilleur qu'un autre, les deux nuages de points sont assez similaires.

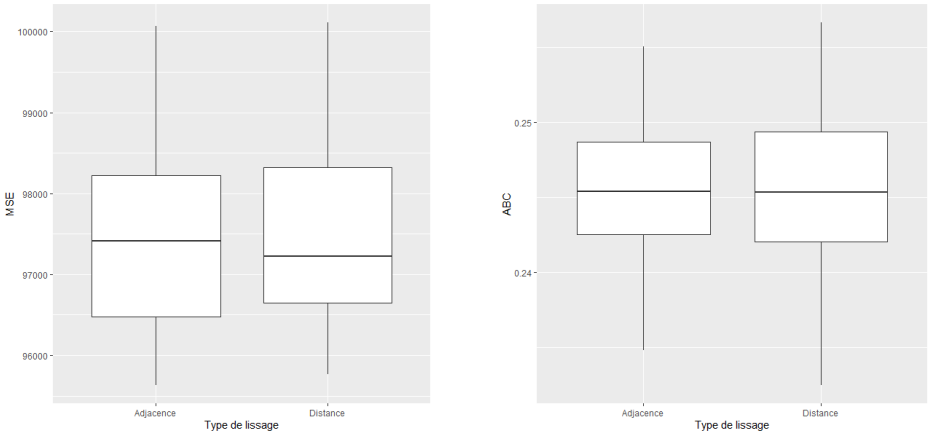


FIGURE 3.6 – Graphique du ABC et du MSE des modèles véhiculiers selon la méthode de réduction de dimension

Les boxplots ici ne permettent pas véritablement de déterminer une méthode plus efficace qu'une autre : pour le MSE, les deux graphiques se chevauchent, même si celui du lissage par adjacence est légèrement plus bas (donc meilleur), la médiane des MSE est plus élevée que pour le lissage par distance. De même, pour l'ABC, les deux boites à moustaches se superposent, les médianes semblent être similaires. Notons simplement la plus faible variabilité de l'ABC pour le lissage par adjacence.

Le tableau ci-dessous présente les performances moyennes des véhiculiers selon le type de lissage ainsi que celle des meilleurs véhiculiers. Les trois mesures de performances utilisées (MSE, ABC et AIC) ne déterminent pas les mêmes caractéristiques d'un bon modèle. Il arrivera donc parfois qu'un modèle jugé bon selon une mesure ne le soit pas selon une autre. Lorsqu'il faudra choisir le meilleur modèle, il

faudra prendre en compte ces différentes variables. Notons qu'à partir d'ici, jusqu'à la fin de ce mémoire, le meilleur véhiculier sera choisi de la façon suivante. Pour chaque modèle, on détermine le classement du MSE et du ABC. Le meilleur véhiculier est défini comme étant celui qui a le meilleur rang en termes de MSE et de ABC combinés. Le choix a été fait de ne pas intégrer l'AIC dans le classement puisqu'il ne mesure pas les performances sur une base de données test mais détermine seulement l'ajustement du modèle sur la base d'entraînement.

Mesure de comparaison	Adjacence	Distance
Meilleur véhiculier en termes de		
– MSE (base de validation)	– 95 847.2	– 95 630.93
– ABC (base de validation)	– 0.237 031	– 0.232 525
– AIC (base d'entrainement)	– 442 500.72	– 442 466.3
Moyenne en termes de		
– MSE (base de validation)	– 97 412.5	– 97 496.5
– ABC (base de validation)	– 0.245 542	– 0.245 389
– AIC (base d'entrainement)	– 442 979.7	– 442 987.0

TABLE 3.9 – Comparaison des performances des véhiculiers issus du lissage par distance et par adjacence

En moyenne, les résultats sont très proches pour les deux types de lissage : le MSE est légèrement meilleur pour le lissage par adjacence, mais légèrement moins bon pour l'ABC. Par contre, le meilleur véhiculier issu du lissage par distance est meilleur que celui par adjacence quelle que soit la mesure.

Conclusion partielle : Finalement, il semble assez difficile de déterminer le meilleur type de lissage. Notons tout de même que le lissage par adjacence nécessite plus de travail (triangulation puis réduction des liens) sans pour autant donner de meilleurs résultats. En termes d'efficacité, on pourrait donc en conclure que le lissage par distance est plus pertinent parce que plus rapide à effectuer. Cette section permet aussi de remarquer que bien que ces deux méthodes soient différentes, les voisinages obtenus sont assez proches.

3.2.2 Pertinence du lissage

La pertinence du lissage peut se mesurer en comparant les résultats des véhiculiers issus d'un lissage et ceux qui ne le sont pas. Les tableaux ci-dessous permettent de comparer les moyennes des performances des modèles GLM avec véhiculier construits après lissage et celles de ceux sans lissage, en fonction du type d'algorithme de classification (CART ou Evolutionary Tree) et des paramètres associés (c'est-à-dire le nombre d'observations par feuille).

	Modèle sans lissage			Moyenne des modèles avec lissage		
Nombre d'observations par feuille	1%	2%	Aucune contrainte	1%	2%	Aucune contrainte
MSE	96 702.5	96 682.3	96 741.4	96 440.51	96 618.3	96 817.5
ABC	0.241 882	0.241 853	0.250 365	0.239 848	0.241 458	0.245 633
AIC	442 756.6	442 780.1	443 085.4	442 644.1	442 764.9	442 810

TABLE 3.10 – Comparaison des modèles lissés ou non lissés CART

	Modèle sans lissage			Moyenne des modèles avec lissage		
Nombre d'observations par feuille	1%	2%	Aucune contrainte	1%	2%	Aucune contrainte
MSE	98 526.5	98 980.3	98 383.0	98 120.6	98 311.2	98 517.4
ABC	0.246 746	0.248 491	0.244 818	0.248 638	0.251 498	0.245 683
AIC	443 095.5	443 378.2	443 131.1	443 203.7	443 281.6	443 204.3

TABLE 3.11 – Comparaison des modèles lissés ou non lissés Evolutionnary Tree

A première vue, ces tableaux pourraient laisser penser que le lissage ne semble pas véritablement pertinent puisqu'en moyenne les modèles lissés (quels que soient ensuite les paramètres choisis lors de l'étape de classification) ont des résultats légèrement meilleurs que les véhiculiers non lissés. Le lissage ne semble pas améliorer les performances autant qu'espéré. Or, comme seule la meilleure classification est gardée par la suite, il s'agit plutôt de comparer le véhiculier issu de résidus non lissés au meilleur véhiculier issu de résidus lissés. En effet, il semble difficile de construire des véhiculiers (après un lissage) qui soient tous meilleurs que celui sans lissage puisque les paramètres ne sont pas toujours optimaux et que l'intérêt de cette étude est justement de trouver les meilleurs paramètres.

Il faut donc comparer le meilleur véhiculier sans lissage (en choisissant la meilleur méthode entre CART et Evolutionnary Tree et le paramètre optimal associé) au meilleur véhiculier avec lissage. Le meilleur véhiculier avec lissage a les paramètres suivants :

- Réduction de dimension par AFDM
- Lissage par distance
- Paramètres de lissage : $d = 2$, $k = 1000$, $p = 2$
- Classification par CART
- Nombre d'observations par feuilles mimimum : 1%

Le meilleur véhiculier sans lissage a les paramètres suivants :

- Classification par CART
- Nombre d'observation minimum par feuille finale : 1%

	Meilleur véhiculier sans lissage	Meilleur véhiculier avec lissage
MSE (Base de validation)	95 702.51	95 630.93
ABC (Base de validation)	0.243 521	0.232 525
AIC (Base d'entrainement)	442 756.5	442 446.3

TABLE 3.12 – Comparaison des meilleurs véhiculiers issus de résidus lissés ou non lissés

Quelle que soit la mesure (MSE, ABC ou AIC), le meilleur véhiculier construit avec un lissage donne de meilleurs résultats que le meilleur véhiculier sans lissage, notamment pour l'ABC pour lequel la différence est significative. Cela permet de confirmer la pertinence du lissage.

Conclusion partielle : Cette étape de lissage semble pertinente puisqu'elle permet en moyenne d'améliorer légèrement les performances mais surtout parce qu'elle permet d'améliorer le meilleur véhiculier. Pour ce qui est du choix de la méthode de lissage, du fait que les deux méthodes ne se distinguent pas vraiment l'une de l'autre en termes de performances, le lissage par distance sera choisi car plus rapide à mettre en place.

3.3 Pertinence du véhiculier

3.3.1 Performances du t-SNE face à l'AFDM

Afin de comparer l'efficacité des deux méthodes de réduction de dimension choisies dans cette étude, une solution consiste à étudier les résultats en termes de MSE, ABC et AIC des modèles avec véhiculier issus d'un t-SNE ou d'une AFDM. Pour cela, on peut :

- Comparer les MSE, AIC et ABC moyens des modèles issus de ces deux algorithmes de réduction de dimension afin de savoir lequel des deux a les meilleures performances en moyenne.
- Pour le lissage par adjacence, comparer une à une les performances des véhiculiers ayant les mêmes paramètres selon qu'ils aient été construits après un t-SNE ou bien une AFDM. Cette comparaison n'est pas possible pour le lissage par distance puisque les paramètres d et k diffèrent selon les deux méthodes de réduction de dimension.
- Regarder laquelle des deux méthodes de réduction de dimension permet d'obtenir le véhiculier avec les meilleures performances. En effet, ici seul le meilleur véhiculier sera retenu dans la suite de l'étude et comparé aux modèles benchmark et il importe finalement peu de connaître les performances des autres véhiculiers.

On peut commencer par observer graphiquement les performances des deux méthodes de réduction de dimension.

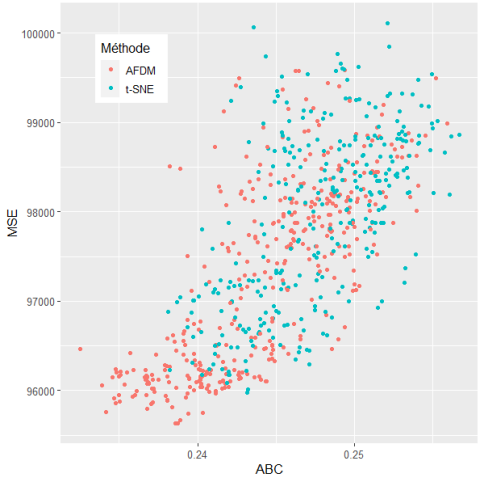


FIGURE 3.7 – Graphique du ABC et du MSE des modèles véhiculiers selon la méthode de réduction de dimension

Ce graphique permet de visualiser le MSE et le ABC des différents véhiculiers en fonction de la méthode de réduction de dimension utilisée. On remarque que les modèles issus d'une AFDM semble donner de meilleures performances en termes de MSE et de ABC (en bas à gauche). En effet, les 20 meilleurs véhiculiers en termes à la fois de MSE et de ABC semblent être construits par AFDM, on ne compte que 2 modèles construits sur un t-SNE dont le MSE est inférieur à 96 000.

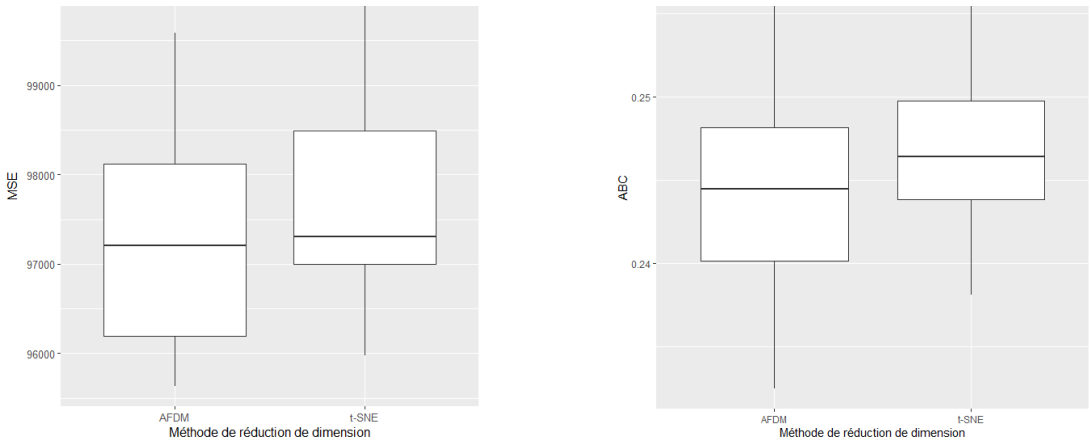


FIGURE 3.8 – Boxplots du MSE et du ABC des différents modèles selon la méthode de réduction de dimension

Les boxplots du MSE et du ABC par méthode de réduction de dimension permettent de voir plus en détails les différences de performances entre les deux algorithmes. Bien que les deux boxplots se chevauchent, la distribution du MSE est plus faible pour l'AFDM que pour le t-SNE. La meilleure performance des véhiculiers issus d'une AFDM semble d'autant plus claire pour le ABC : la médiane du ABC pour l'AFDM est quasiment égale au premier quartile du ABC du t-SNE.

On peut ensuite étudier plus de façon plus chiffrée les performances de ces deux méthodes.

	Mesure de performance	AFDM	t-SNE
Lissage par adjacence CART	MSE	43[1]	20
	ABC	49	14
	AIC	48	15
Lissage par adjacence EVTREE	MSE	40	23
	ABC	42	21
	AIC	47	16
Moyenne	MSE	66%	34%
	ABC	72%	28%
	AIC	75%	25%

TABLE 3.13 – Comparaison des véhiculiers un à un selon les deux algorithmes de réduction de dimension

Ici, quand on compare deux véhiculiers construits par lissage par adjacence avec les mêmes paramètres sur la base de données de validation, ceux issus de l'AFDM donnent de meilleurs véhiculiers que le t-SNE dans 66% des cas si l'on se réfère à la mesure du Mean Squared Error, 72% pour le ABC et dans 83% des cas pour l'AIC.

Mesure de comparaison	AFDM	t-SNE
Meilleur véhiculier en termes de		
– MSE (base de validation)	– 95 630.93	– 96 226.1
– ABC (base de validation)	– 0.233	– 0.238
– AIC (base d'entrainement)	– 442 466.3	– 442 528.3
Moyenne en termes de		
– MSE (base de validation)	– 97 234.3	– 97 707.6
– ABC (base de validation)	– 0.244	– 0.247
– AIC (base d'entrainement)	– 442 925.1	– 443 044.5

TABLE 3.14 – Comparaison des performances des véhiculiers issus de l'AFDM et du t-SNE

Pour ce qui est de la comparaison des performances moyennes des véhiculiers issus des deux méthodes de réduction de dimension, les trois mesures vont dans le même sens. En moyenne, les véhiculiers construits à partir d'une AFDM donnent de meilleurs résultats que ceux issus du t-SNE pour les trois

[1]Cela signifie que pour les lissages par adjacence avec un CART, quand on compare les véhiculiers ayant les mêmes paramètres (même paramètres de lissage et de classification), dans 43 cas sur 63, le véhiculier issu de l'AFDM est meilleur que le t-SNE.

mesures utilisées (le Mean Squared Error, le ABC et l'Akaike Information Criterion). De même, lorsqu'on compare les performances du meilleur véhiculier issu de l'AFDM et du t-SNE, le premier donne de meilleurs résultats.

Conclustion partielle : Toutes ces observations permettent de conclure que l'AFDM est plus une méthode plus performante que le t-SNE lorsqu'il s'agit de réduire l'espace de variables véhicule dans le cas de cette garantie Bris-de-Glace. Afin de confirmer cette affirmation, des analyses similaires seront faites par la suite sur une autre garantie (voir la section 3.4).

3.3.2 Performances des Evolutionary Trees face aux CART

A présent, il s'agit de comparer les pertinences relatives des deux algorithmes de classification : les arbres de régression et les Evolutionary Trees. Nous pouvons commencer par observer graphiquement leurs performances.

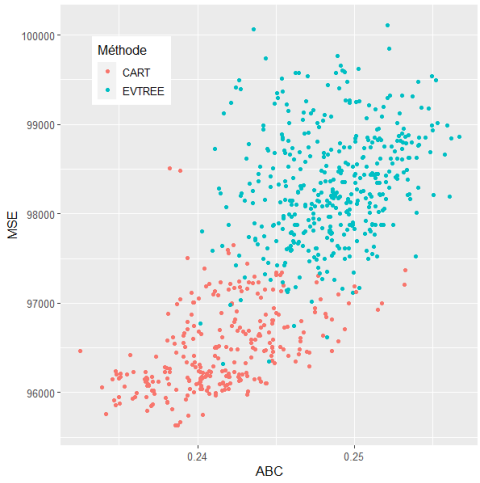


FIGURE 3.9 – Graphique du ABC et du MSE des modèles véhiculiers selon la méthode de classification

Ce graphique du MSE et du ABC des véhiculiers en fonction de la méthode de classification permet d'avoir une première idée de la performance relative du CART et de l'Evolutionary Tree. Le CART semble être beaucoup plus efficace que l'Evoutionary Tree puisque le nuage de point correspondant est concentré en bas à gauche du graphique et celui des Evolutionary Trees est bien plus en haut à droite. Les deux nuages de points sont pratiquement disjoints, ce qui semble indiquer que la meilleure performance du CART ne fait pas de doute.

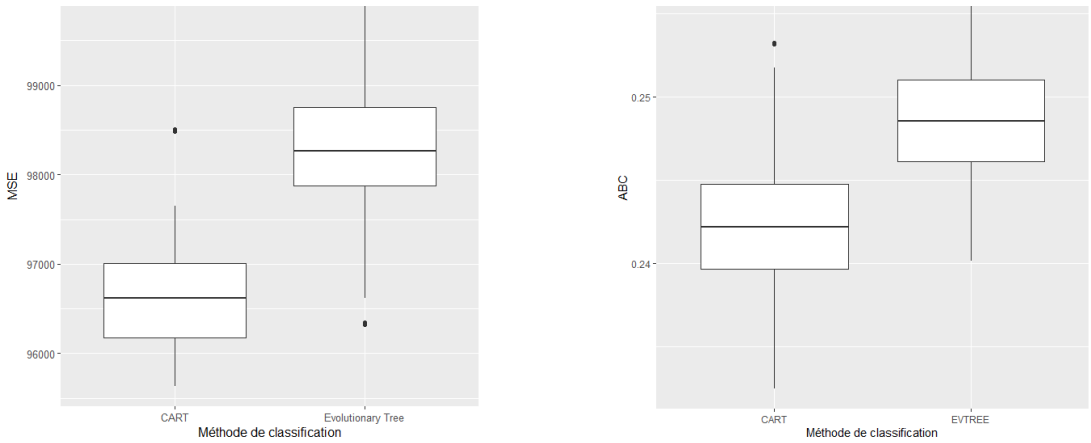


FIGURE 3.10 – Boxplots du MSE et du ABC des différents modèles selon la méthode de classification

L'analyse des boxplots du MSE et du ABC selon l'algorithme de classification confirme la meilleure performance du CART. Les deux boites à moustaches ne se chevauchent pas, très peu de modèles construits par Evolutionary Tree sont capables d'avoir des performances similaires à celles du CART. C'est particulièrement clair pour le MSE : alors que plus 75% des modèles avec un véhiculier construit par CART ont un MSE inférieur à 97 000. seuls quelques modèles véhiculier construits par Evolutionary Tree y parviennent. Ainsi le meilleur véhiculier construit avec un Evolutionary Tree n'est que 167-ième dans le classement conjoint du MSE et du ABC.

De même que pour les méthodes de réduction de dimension, la comparaison entre les deux algorithmes de classification peut se faire de la façon suivante :

- Déterminer le meilleur véhiculier pour chaque méthode de classification puis comparer les performances de ces deux modèles.
- Comparer les MSE, AIC et ABC moyens des modèles.
- Pour chaque combinaison de paramètres, comparer un à un tous les véhiculiers ayant les mêmes paramètres et regarder lequel des deux a les meilleures performances.

	Mesure de performance	CART	Evolutionary Tree
Lissage par distance AFDM	MSE	142 [2]	2
	ABC	133	11
	AIC	144	0
Lissage par adjacence AFDM	MSE	62	1
	ABC	52	11
	AIC	63	0
Lissage par distance t-SNE	MSE	143	1
	ABC	122	22
	AIC	144	0
Lissage par adjacence t-SNE	MSE	62	1
	ABC	45	18
	AIC	62	1
Moyenne	MSE	99%	1%
	ABC	85%	15%
	AIC	99.8%	0.2%

TABLE 3.15 – Comparaison des véhiculiers un à un selon les deux algorithmes de classification

Ce tableau permet de comparer à paramètres identiques les performances de deux véhiculiers selon leur méthode de classification. Ainsi on compare le MSE, ABC et AIC de deux véhiculiers construits d'après les mêmes méthodologies et suivant les mêmes paramètres. Cela permet ainsi de déterminer la qualité relative des deux méthodes de classification. Ici le résultat semble assez clair : les véhiculiers construits par un CART sont bien meilleurs que les véhiculiers construits par Evolutionary Tree. En moyenne, à paramètres identiques, les modèles construits par CART sont meilleurs que ceux construits par Evolutionary Tree dans 99% des cas d'après le MSE, 85% pour le ABC et quasiment 100% pour l'AIC.

[2]Cela signifie que pour les lissages par distance après une AFDM, quand on compare les véhiculiers ayant les mêmes paramètres (même paramètres de lissage et de classification), dans 142 cas parmi 144, le véhiculier construit par un CART est meilleur (en termes de MSE) que celui construit par Evolutionary Tree.

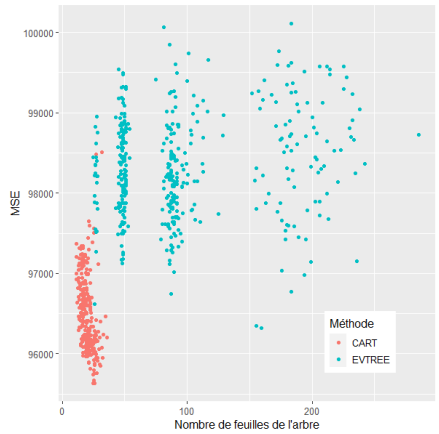
Mesure de comparaison	CART	EVTREE
Meilleur véhiculier en termes de		
– MSE (base de validation)	– 95 630.93	– 96 323.5
– ABC (base de validation)	– 0.232 525	– 0.240 186
– AIC (base d'entrainement)	– 442 466.3	– 442 907.13
Moyenne en termes de		
– MSE (base de validation)	– 96 625.4	– 98 316.4
– ABC (base de validation)	– 0.242 314	– 0.248 606
– AIC (base d'entrainement)	– 442 739.7	– 443 229.9

TABLE 3.16 – Comparaison des performances des véhiculiers issus d'un CART et d'un EVTREE

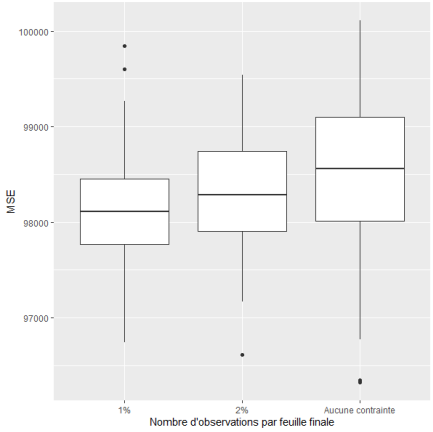
La comparaison des meilleurs véhiculiers selon la méthode de classification permet encore une fois de confirmer la meilleure performance du CART. Quelle que soit la mesure choisie, le meilleur véhiculier issu d'un CART est meilleur que le meilleur véhiculier issu d'un Evolutionary Tree. Pour ce qui est des performances moyennes, les chiffres sont particulièrement révélateurs : en moyenne les véhiculiers construits par CART ont des MSE 2% plus faibles que pour les Evolutionary Tree et 5% plus élevés pour le ABC.

Conclusion partielle : Il semble assez clair que les CART sont bien plus pertinents que les Evolutionary Trees pour classifier les résidus lissés. Toutes les statistiques étudiées (les performances moyennes, la comparaison des meilleurs véhiculiers et la comparaison de modèles à paramètres identiques) vont dans ce sens.

Comment expliquer la mauvaise performance relative des Evolutionary Trees par rapport aux CART ? On pourrait supposer que cette méthode plus complexe que les arbres de régression classiques ait tendance à surapprendre les données. Pour vérifier cela, on peut commencer par observer les performances en fonction du nombre de classes créées par les algorithmes de classification afin de savoir si l'Evolutionary Tree ne crée pas des arbres trop complexes.



MSE en fonction de la taille de l'arbre et de la méthode de classification



Boxplot des modèles construits par Evolutionary Tree en fonction des paramètres de classification

Ces graphiques permettent de confirmer les hypothèses précédentes : les véhiculiers construits par Evolutionary Trees ont tendance à contenir plus de classes que les CART (bien qu'il y ait les mêmes contraintes en termes de nombre d'observations par feuille). Il semblerait qu'au-delà de 50 classes, le MSE augmente de façon significative et il y a un risque de surapprentissage. Cependant, notons aussi que pour les véhiculiers construits par Evolutionary Tree qui ont un nombre de classes plus faible (autour de 40 classes, nombre de classe similaire aux CART), leur performance n'est pas véritablement meilleure que pour les autres véhiculiers. Les graphiques boxplot permettent de confirmer cette dernière hypothèse : la performance des véhiculiers Evolutionary Trees ne semble pas meilleure dans le cas où la contrainte du nombre d'observations par feuille est plus stricte : quand on impose un minimum de 1% d'observations par feuilles finales, les résultats sont meilleurs que lorsqu'on supprime cette contrainte mais aussi quand la contrainte est de 2% .

Conclusion partielle : D'après toutes les observations précédentes, on peut aussi conclure que le choix de l'algorithme de classification est déterminant. En effet, on remarque sur la Figure 3.9 que le choix de l'algorithme de classification est très important parce qu'il est très discriminant en termes de performances. On le comprend facilement étant donné que cette étape est la dernière étape de construction du véhiculier et permet d'attribuer à chaque véhicule une classe.

3.3.3 Pertinence de la variable véhiculier

A présent, il s'agit d'évaluer la pertinence du meilleur véhiculier construit par les 828 combinaisons de paramètres possibles. Rappelons ses caractéristiques :

- Réduction de dimension par AFDM
- Lissage par distance
- Paramètres de lissage : $d = 2$, $k = 1000$, $p = 2$
- Classification par CART
- Nombre d'observations par feuilles mimimum : 1%

Il s'agit d'abord de vérifier la pertinence de cette nouvelle variable. On peut commencer par évaluer son importance pour la prédiction des coûts de sinistre Bris-de-Glace en calculant l'AIC d'un modèle GLM à une seule variable. La variable la plus pertinente sera donc celle pour laquelle l'AIC sera le plus faible. En d'autres termes, on cherche à déterminer la première variable intégrée dans un modèle GLM à sélection de variable forward.

Variable	**AIC (Base d'entraînement)**
Véhiculier	552 266.4
Prix	553 749.6
Puissance	553 761.9
Marque	554 094.3
Segment	554 168.2
Poids-Lourd ou voiture	555 580.2
Code Mosaic	555 943.4
Nombre d'années sans sinistres en tort	555 957.0
Carburant	555 975.2
Age du conducteur	555 992.7
Classe Sociale	555 995.0
Présence garage	556 015.0
Zonier	556 017.2
Fractionnement	556 019.2
Présence gobox	556 069.3
Etat civil	556 074.1
Année	556 079.6
Ancienneté du contrat	556 082.9
Nombre de places	556 086.2
Nombre de conducteurs	556 090.0
Age du véhicule	556144.2
Masse	556 161.7
Cylindrée	556 163.4

TABLE 3.17 – Importance des variables

On remarque que la variable véhiculier est clairement la plus pertinente de toutes dans ce modèle. Elle est plus pertinente que toutes les variables véhicule prises individuellement, elle est donc capable d'agréger l'information de plusieurs variables véhicule à la fois et de sélectionner correctement l'information relative aux véhicules nécessaire à la prédiction des coûts des sinistres.

Une deuxième manière de mesurer la pertinence de cette variable véhiculier est de comparer les performances d'un modèle avec cette variable avec celle d'un modèle sans.

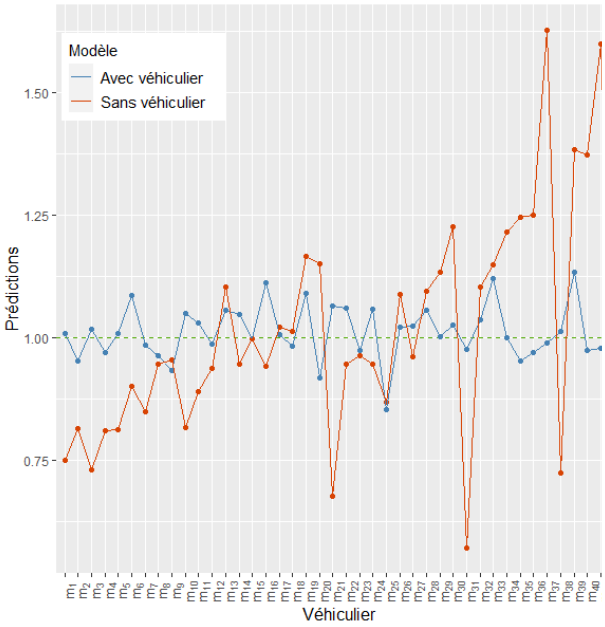


FIGURE 3.11 – Prédiction du modèle avec et sans véhiculier sur la base de données test

Ce graphique permet de comparer les ratios entre les coûts observés et les prédictions des modèles avec ou sans véhiculier (mais avec toutes les variables non véhicule) sur la base de données test. Plus ce quotient est proche de 1, meilleur est le modèle. On remarque que le véhiculier semble pertinent puisque les prédictions sont plus proches de 1 que dans le modèle sans véhiculier.

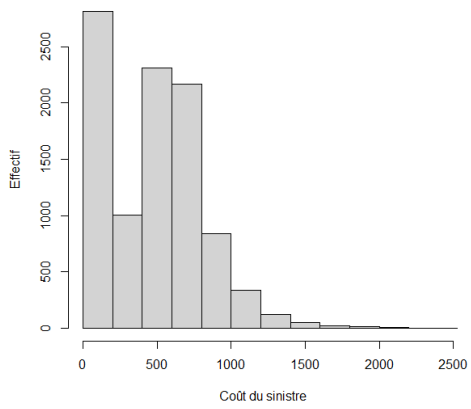
Conclusion partielle : On peut conclure à la pertinence de la variable véhiculier en soi : cette variable contient beaucoup d'information utile à la prédiction, permet de discriminer les coûts des sinistres et est construite de façon à être très pertinente (elle devient la plus pertinente de toutes). Cependant, afin de mesurer l'impact réel d'un véhiculier, il faut comparer ses performances non seulement avec un modèle sans cette variable mais surtout avec un modèle alternatif qui intègre l'information brute relative aux véhicules.

3.3.4 Comparaison du modèle véhiculier aux modèles optimaux

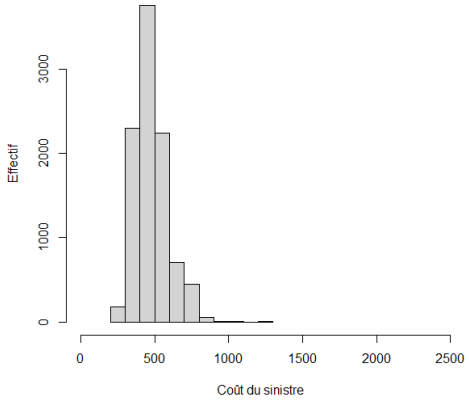
Il reste à présent à comparer les performances du meilleur véhiculier (déterminé par les résultats de prédiction sur la base de données de validation) avec les modèles benchmark. Afin que la comparaison soit pertinente, on relance le modèle GLM contenant le meilleur véhiculier une seconde fois mais sur les bases d'entraînement et de validation afin de comparer des modèles entraînés sur les mêmes données (pour rappel, les modèles benchmark sont entraînés sur les bases de données d'entraînement et de validation puisqu'il n'y a pas de sélection de modèle pour ceux-ci). Voici les performances du modèle véhiculier ainsi que celles des deux modèles benchmark :

Modèle	**Mean Squared Error** (Base de données test)	**ABC** (Base de données test)	**AIC** (Base de données d'entrainement et validation)
Modèle GLM fused lasso avec • toutes les variables non véhicule • toutes les variables véhicule	90 828.46	0.244 788	552 387.3
Modèle GLM fused lasso avec • toutes les variables non véhicule • 4 variables véhicule	91 002.94	0.246 008	552 473.8
Modèle GLM avec • toutes les variables non véhicule • le meilleur véhiculier	91 247.58	0.247 507	552 311.3

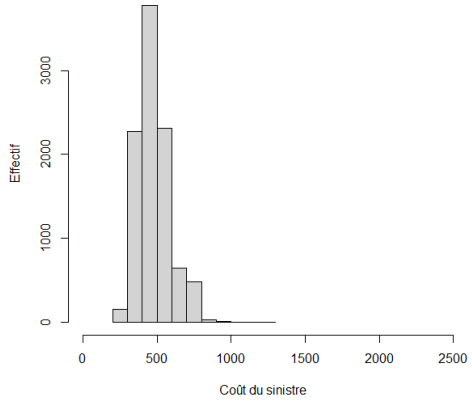
Bien que l'AIC du modèle avec véhiculier soit légèrement meilleur que celui des modèles fused lasso, il donne de moins bons résultats pour les mesures de prédictions (MSE et ABC), bien que la différence ne soit pas très grande. Comment expliquer la moindre performance du véhiculier ? L'analyse graphique des prédictions de sinistres sur la base de test permet d'avoir une première idée des prédictions de chacun des modèles.



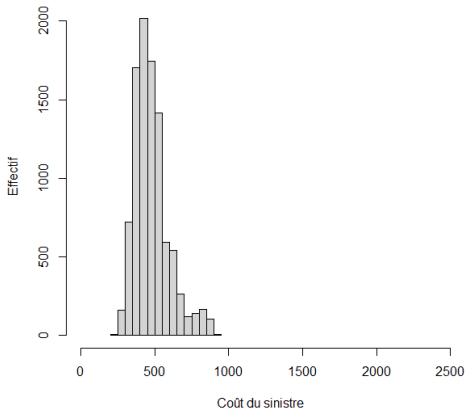
Coûts observés sur la base de test



Prédictions du modèle benchmark avec toutes les variables véhicule



Prédictions du modèle benchmark avec quatre variables véhicule



Prédictions du modèle avec le meilleur véhiculier

FIGURE 3.12 – Graphique des prédictions des différents modèles

On remarque tout d'abord que les trois modèles n'ont pas réussi à reproduire l'allure de l'histogramme des coûts observés sur la base de données test : celui-ci est bimodale avec un pic autour de 100 euros et un autre autour de 750 euros et il y a quelques sinistres dont le coût est supérieur à 1 500 euros. Au contraire, les trois modèles prédisent des coûts dont la distribution est unimodale avec un pic autour de 500 euros et aucun sinistre n'a de prédiction supérieure à 1 300 euros.

Cependant, parmi les trois modèles, celui avec véhiculier prédit des sinistres dont la dispersion est plus faible que les autres modèles : les coûts maximum prédits sont aux alentours de 1 000 euros contre 1 300 euros pour les modèles benchmark. En d'autres termes, le modèle avec véhiculier semble moins bien prédire les sinistres dont le coût est important et ce, malgré la mise en place de poids dans le CART afin de donner plus de poids aux sinistres importants.

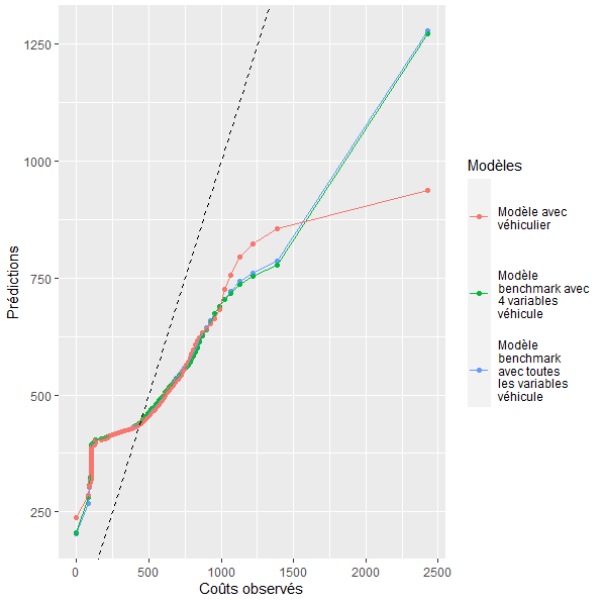


FIGURE 3.13 – Grapique des percentiles de chacun des modèles en fonction du coût observé correspondant

Ce graphique permet de comparer les prédictions des modèles benchmark avec celles du modèle véhiculier mais aussi de comparer ces trois modèles aux coûts observés sur la base de test. On remarque tout d'abord que tous les trois ont tendance à surestimer les sinistres dont le coût est faible et à sous-estimer ceux dont le coût est très grand. Cela confirme les observations précédentes : le défaut de ces modèles est de ne pas prédire correctement les sinistres les plus graves.

De plus, on remarque que la courbe des percentiles des modèles avec véhiculier est très proche de celle des autres modèles. Il paraît difficile de savoir en quoi ce modèle a de moins bonnes performances que les modèles benchmark. Ce modèle est même meilleur que les modèles avec toutes ou 4 variables véhicule pour prédire les sinistres entre 1000 et 1500 euros. On peut essayer de déterminer pour différents segments de coûts le MSE des différents modèles afin de déterminer pour quels sinistres le modèle véhiculier est moins efficace. Tout comme dans la partie 3.1, on calcule pour différents segments I (correspondant à l'intervalle entre deux déciles consécutifs) la valeur suivante :

$$\frac{1}{n}\sum_{i:y_i\in I}(y_i-\hat{y}_i)^2$$

Intervalle	$[0.q_{10}[$	$[q_{10},q_{20}[$	$[q_{20},q_{30}[$	$[q_{30},q_{40}[$	$[q_{40},q_{50}[$	$[q_{50},q_{60}[$	$[q_{60},q_{70}[$	$[q_{70},q_{80}[$	$[q_{80},q_{90}[$	$[q_{90},max[$
Modèle avec toutes les variables véhicule	14272.0	6597.7	8505.8	2705.1	895.3	1550.1	2895.4	4127.7	7749.4	27290.0
Modèle véhiculier	14245.5	6783.3	8528.6	2774.5	936.1	1682.3	3025.6	4284.7	8005.0	26849.2
Différence	-26.5	185.6	22.8	69.5	40.8	132.2	130.3	157.0	255.6	-440.7

TABLE 3.18 – MSE des modèles par décile des coûts observés

Le modèle véhiculier fait des prédictions plus pertinentes pour les sinistres les plus importants (sur les 10% des sinistres les plus graves) ainsi que sur les 10% des sinistres les moins graves. Mais pour les autres sinistres, le modèles véhiculier se montre moins pertinent que les modèles avec pénalisation fused lasso, notamment pour le second décile mais aussi entre les quantiles q_{50} et q_{90}.

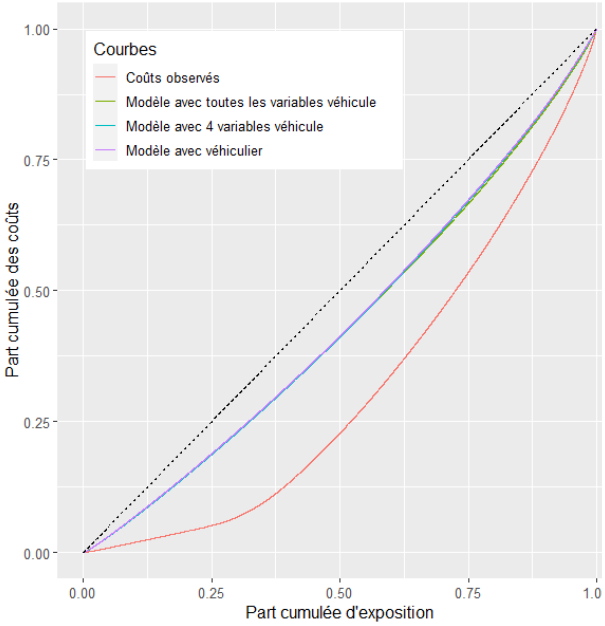


FIGURE 3.14 – Courbe de Lorenz

La courbe de Lorenz des trois modèles et des coûts observés confirme les observations précédentes : d'une part ces modèles surestiment les coûts dont le montant est faible (ils prédisent que les 25% des sinistres les moins graves représentent 19% du total des sinistres alors que ce n'est que 6% en réalité) et sous-estiment les sinistres les plus graves (ils prédisent que les 10% des sinistres les plus graves représentent 15% du total alors que c'est 24% en réalité). De plus, ces trois modèles ont des courbes de Lorenz très proches, mais celle du modèles benchmark avec toutes les variables véhicule semble un peu meilleure car plus basse que les autres notamment entre les quantiles à 60% et 80%, même si la différence reste faible.

Le temps de calcul est un autre aspect à ne pas oublier lorsqu'il s'agit de choisir ou non un véhiculier. Voici quelques détails du temps nécessaire pour mettre en place les différents modèles :

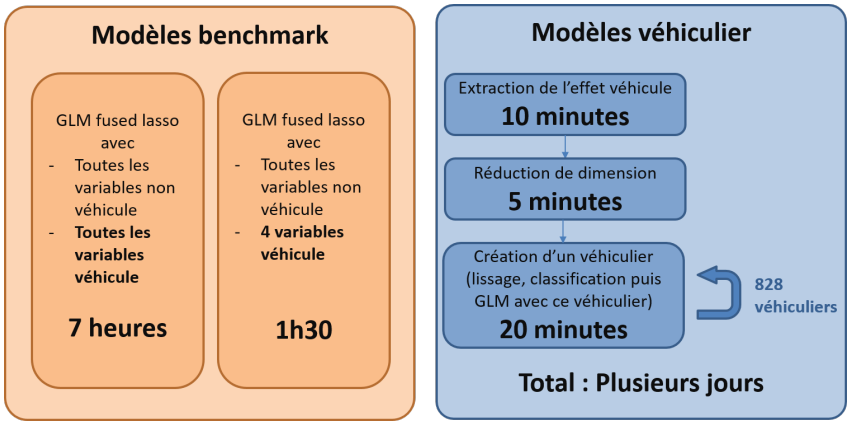


FIGURE 3.15 – Temps de calcul des différents modèles

Conclustion partielle : Le véhiculier nécessite beaucoup de temps de calcul en amont afin d'en sélectionner la meilleure version, ce qui au total, nécessite plus de temps que la mise en place des GLM benchmark, sans que les résultats n'en soient meilleurs. Cependant, selon l'utilisation qui est faite de ces différents modèles, un véhiculier pourrait être utile. On peut imaginer que dans le cas où il faut exécuter régulièrement un modèle GLM, le choix du véhiculier (bien que coûteux en amont) peut se révéler plus judicieux puisque le modèle GLM final (après avoir déterminé le véhiculier) est quant à lui très rapide à calculer.

3.4 Cas particulier d'une garantie avec peu d'observations

Afin de s'assurer de la pertinence des conclusions énoncées précédemment, il semblait intéressant de construire à nouveau un véhiculier sur une autre garantie. De plus, cette étape supplémentaire permettra de confirmer ou d'infirmer l'hypothèse faite au cours de ce mémoire selon laquelle, la construction d'un véhiculier pourrait s'avérer pertinente dans le cas d'une garantie ayant peu d'observations. En effet, cette situation se présente souvent dans le cas d'une jeune entreprise d'assurance ayant peu de clients et donc peu données mais aussi pour une compagnie plus ancienne qui lance un nouveau produit ; ou encore lors de l'étude d'une garantie dont la fréquence de sinistres est très faible. Utiliser un véhiculier pourrait être une bonne solution dans ce cas afin de veiller à ce que le nombre de paramètres d soit assez faible puisque le nombre d'observation n est lui aussi faible.

Dans le cas présent, le coût de la garantie VOL sera étudiée. Les observations seront de nouveau découpées en trois bases de données :

- Base de données d'entrainement : 2 957 observations
- Base de données de validation : 611 observations
- Base de données de test : 718 observations

Dans cette partie, nous ne reviendrons pas en détails sur la pertinence des différentes étapes de construction du véhiculier, celles-ci sont identiques au travail sur la garantie Bris-de-Glace. Le meilleur véhiculier d'après les performances sur la base de validation sera simplement sélectionné et comparé sur la base de test avec les modèles benchmark.

La méthodologie utilisée est identique : d'une part, les modèles benchmark sont construits, d'autre part la construction du véhiculier suit les mêmes étapes que pour la garantie Bris-de-Glace : extraction de l'effet véhicule grâce à un premier GLM excluant les variables véhicule, réduction de dimension, lissage puis classification des résidus.

3.4.1 Pertinence de la pénalisation fused lasso

L'étude de cette nouvelle garantie permet de vérifier l'affirmation de Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu et Keith Knight (dans leur article introduisant la pénalisation fused

lasso [20]) selon laquelle ce nouveau type de pénalisation est pertinent pour les cas où d le nombre de paramètres du modèles (c'est-à-dire le nombre de variables) est relativement grand comparé au nombre d'observations (noté n), voire le cas où $d > n$. Pour les modèles où l'on intègre la totalité des variables à disposition, il y a 284 variables (en comptant comme variable chaque modalité des variables qualitatives) et 2 957 observations seulement, ce qui est très peu.

Modèle	**Mean Squared Error** (Base de données test)	**ABC** (Base de données test)	**AIC** (Base de données d'entraînement)
Modèle GLM avec • toutes les variables non véhicule • toutes les variables véhicule • pénalisation fused lasso	9.548792×10^7	0.264 758	53 322.2
Modèle GLM avec • toutes les variables non véhicule • toutes les variables véhicule • pénalisation lasso	9.599243×10^7	0.390 350	54 423.2
Modèle GLM avec • toutes les variables non véhicule • toutes 4 variables véhicule • pénalisation fused lasso	9.734380×10^7	0.267 486	53 364.4
Modèle GLM avec • toutes les variables non véhicule • toutes 4 variables véhicule • pénalisation lasso	9.890363×10^7	0.392 321	55 962.3

Tout d'abord, avant d'analyser le tableau, notons que les deux modèles se distinguent clairement. Dans le cas du modèle avec toutes les variables véhicule, 34% des modalités des variables du modèles ont été regroupées avec d'autres et 15% des modalités ont été écartées (celles qui ont un coefficient égal à 0), alors que le modèle lasso écarte 80% des variables, ce qui est très important. On pourrait même penser qu'il y a un risque ici de sous-apprentissage pour le modèle lasso, ce modèle n'étant pas capable d'attribuer des coefficients aux variables, étant donnée le faible volume d'observations.

De plus, quelle que soit la mesure utilisée (MSE, ABC ou AIC), le modèle avec pénalisation fused lasso donne de meilleurs résultats que la pénalisation lasso (que ce soit avec toutes les variables véhicule ou bien seulement 4). Mais contrairement à ce qu'on aurait pu espérer, les différences de performances ne sont pas très grandes. La différence entre le Mean Squared Error des modèles lasso et fused lasso avec toutes les variables véhicule est de 2%, ce qui est de l'ordre de grandeur des différences observées sur la garantie Bris-de-Glace. Ce type de pénalisation demeure pertinent, mais ne semble pas plus efficace sur une petite base de données.

3.4.2 Comparaison des méthodes de réduction de dimension et de classification

Voilà les boxplots permettant de comparer les performances des deux méthodes de classification et de réduction de dimension sur cette garantie Vol.

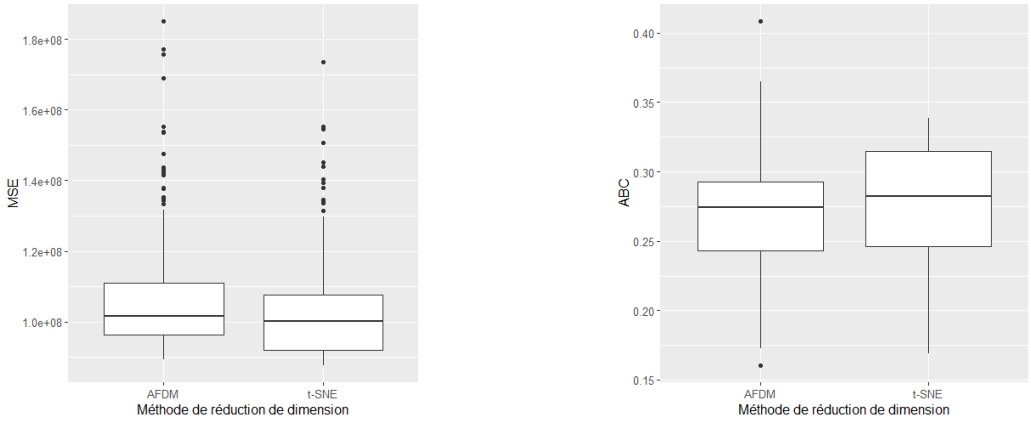


FIGURE 3.16 – Boxplot des MSE et ABC des modèles véhiculier selon la méthode de réduction de dimension

Pour ce qui est des méthodes de réduction de dimension, les résultats ne sont pas vraiment similaires à ceux observés pour la garantie Bris-de-Glace. Alors que pour cette première garantie, les résultats étaient en faveur de l'AFDM (que ce soit pour l'ABC ou le MSE), les boxplots de l'ABC sont plus similaires pour les deux méthodes de la garantie Vol, bien qu'ils restent favorables à l'AFDM. Mais pour le MSE c'est l'inverse, les modèles issus d'un t-SNE sont légèrement meilleurs.

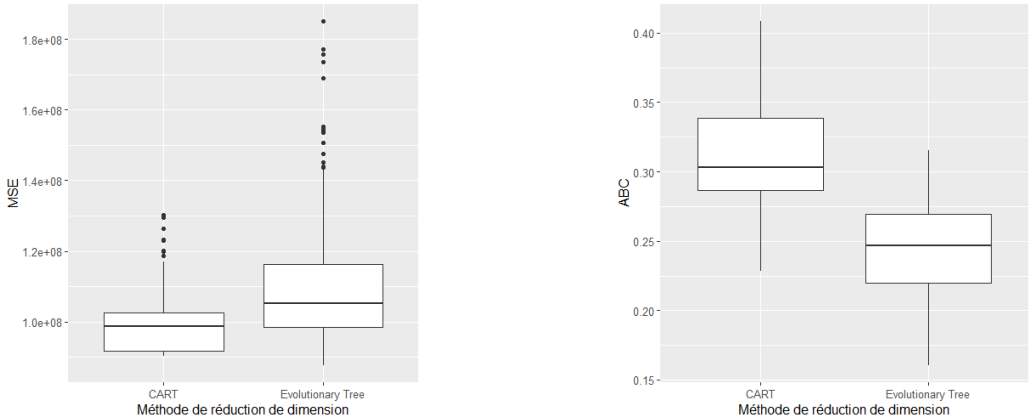


FIGURE 3.17 – Boxplot des MSE et ABC des modèles véhiculier selon la méthode de classification

Pour la classification, les résultats sont assez contradictoires. D'après le MSE, l'algorithme CART semble meilleur que l'Evolutionary Tree (comme pour la garantie Bris-de-Glace). Mais pour l'ABC c'est l'inverse : les véhiculiers issus d'un Evolutionary Tree semblent bien meilleurs que ceux construits par CART.

Conclusion partielle : Les comparaisons des méthodes de réduction de dimension et de classification

ne confirment pas vraiment les observations faites sur la garantie Bris-de-Glace. Alors que le CART était bien meilleur pour cette première garantie, les observations sur la garantie Vol semblent nuancer ces conclusions. De même pour le t-SNE, les observations sur la garantie Vol infirment en partie celles faites sur la garantie Bris-de-Glace.

3.4.3 Pertinence du véhiculier

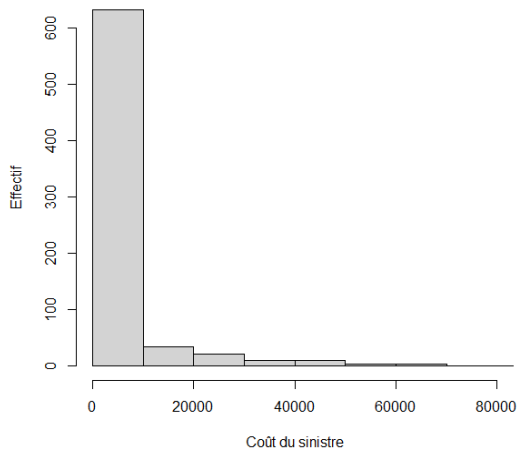
Le véhiculier sélectionné comme étant celui qui a les meilleurs performances sur la base de données de validation (en termes de classement du MSE et du ABC) a les caractéristiques suivantes :

- Méthode de réduction de dimension : t-SNE
- Type de lissage : Adjacence
- Paramètres de lissage : $\alpha = 0.5$, $p = 2$
- Méthode de classification : Evolutionary Tree
- Paramètre de classification : Pourcentage minimal d'observation par feuille = 1%

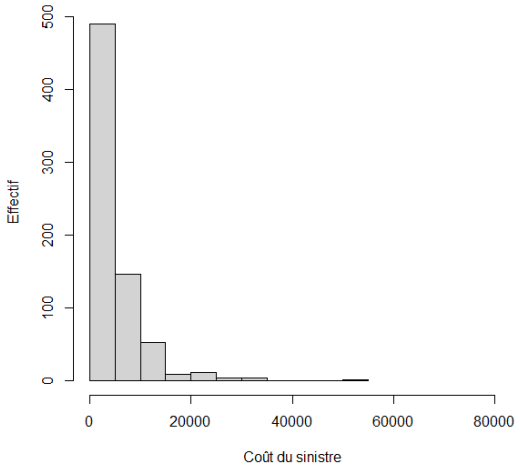
Voici les performances en termes de MSE, ABC et AIC des modèles benchmark fused lasso avec 4 ou toutes les variables véhicule ainsi que le modèle avec véhiculier.

Modèle	**Mean Squared Error** (Base de données test)	**ABC** (Base de données test)	**AIC** (Base de données d'entrainement et validation)
Modèle GLM fused lasso avec • toutes les variables non véhicule • toutes les variables véhicule	9.548792×10^7	0.264 758	53 322.2
Modèle GLM fused lasso avec • toutes les variables non véhicule • 4 variables véhicule	9.734380×10^7	0.267 486	53 364.4
Modèle GLM avec • toutes les variables non véhicule • le meilleur véhiculier	1.064427×10^8	0.288 011	53 399.7

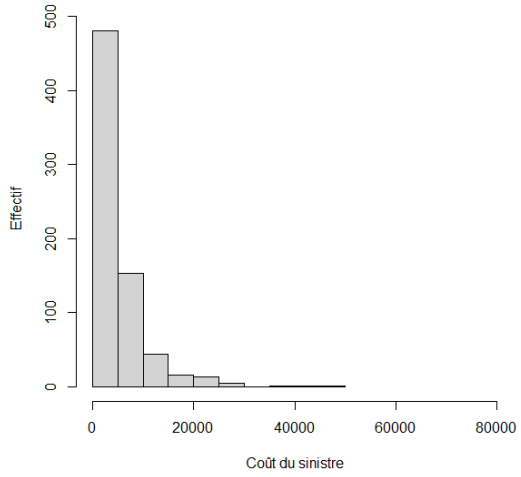
On remarque que le modèle avec véhiculier, quelle que soit la mesure choisie, a des performances moins bonnes que les deux modèles benchmark y compris celui avec seulement 4 variables véhicule. De plus, la différence de performance entre ces modèles est du même ordre de grandeur que pour la garantie Bris-de-Glace. La différence entre le MSE du modèle fused lasso avec toutes les variables véhicule et celui du modèle avec véhiculier est de 10% dans le cas du Vol contre 1% dans le cas du Bris-de-Glace. Ainsi, contrairement à ce qu'on pouvait espérer, le véhiculier ne semble pas plus efficace dans le cas d'une base de données dont le nombre d'observations est faible.



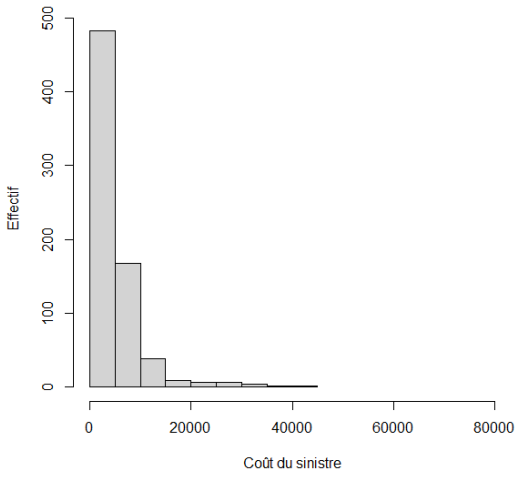
Coûts observés sur la base de test



Prédictions du modèle benchmark avec toutes les variables véhicule



Prédictions du modèle benchmark avec quatre variables véhicule



Prédictions du modèle benchmark avec le meilleur véhiculier

FIGURE 3.18 – Graphique des prédictions des différents modèles

Globalement, les deux modèles benchmark et le modèle avec véhiculier n'arrivent pas à prédire les sinistres dont le coût est le plus important, même si le modèle fused lasso avec toutes les variables véhicule y arrive légèrement mieux que le modèle avec seulement 4 variables véhicule, lui même meilleur que le modèle avec véhiculier. En effet, le modèle avec véhiculier prédit des coûts de sinistres plus concentrés autour de faibles valeurs alors que les autres modèles prédisent des distributions de coûts éclatées et plus proches des coûts observés. La variable véhiculier ne parvient pas à discriminer suffisamment les observations afin d'effectuer des prédictions sur une échelle de valeur plus grande.

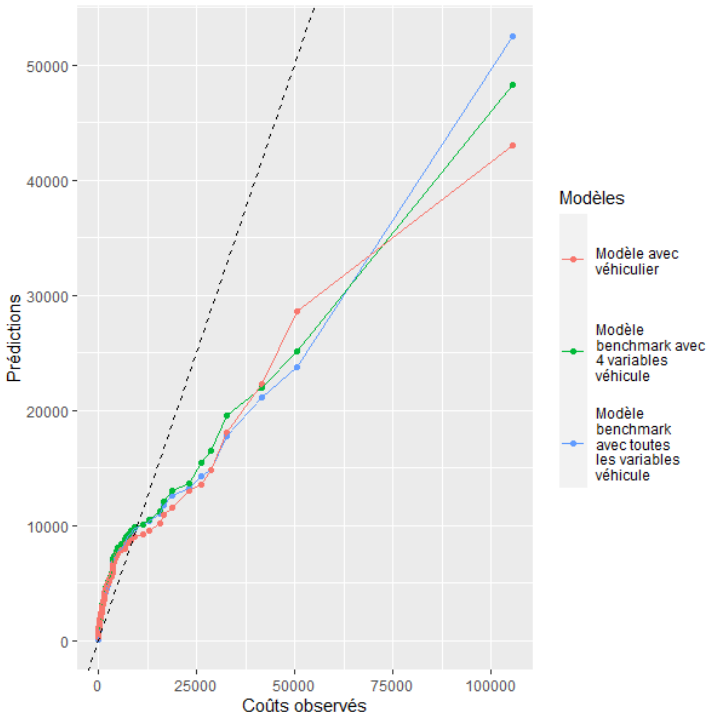


FIGURE 3.19 – Graphique des percentiles des prédictions des différents modèles

Ce graphique permet d'observer les percentiles des prédictions des trois modèles en fonction des sinistres observés sur la base de données test. On remarque d'abord que ces trois modèles sont assez proches : les courbes des quantiles sont très rapprochées. Ces trois modèles ont tendance à surestimer les sinistres ayant un coût observé inférieur à 10 000 euros, puis sous-estimer ceux dont le coût est supérieur. De plus, au-delà de 10 000 euros, les prédictions sont d'autant moins précises que le coût observé est grand (plus les sinistres sont graves plus les courbes des percentiles s'écartent de la courbe noire d'équation $y = x$), ce qui confirme les observations précédentes.

De plus, il faut remarquer que le modèle avec véhiculier a tendance à plus sous-estimer les sinistres graves que les modèles benchmark, comme observé précédemment. En effet, au-delà de 10 000 euros, la courbe des percentiles du modèle véhiculier est quasiment toujours en dessous de celle des modèles benchmark.

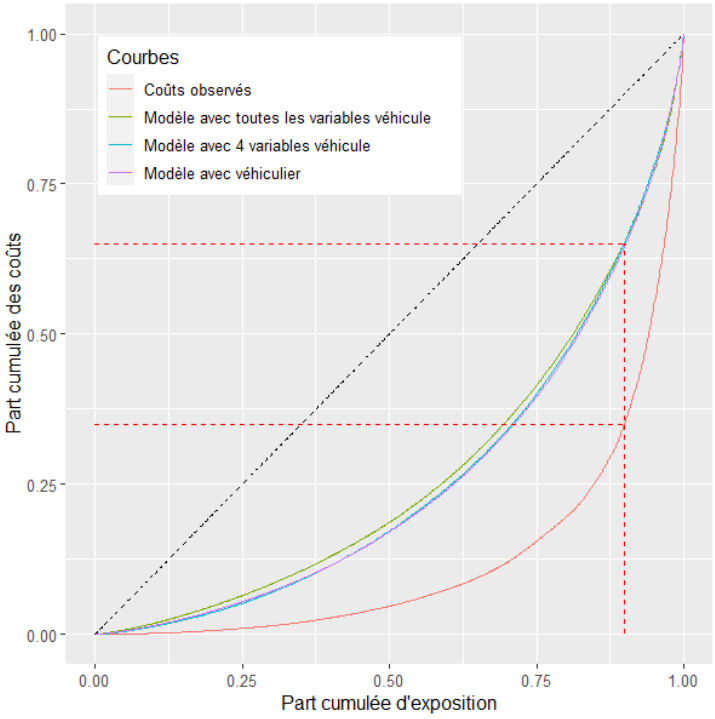


FIGURE 3.20 – Graphique des prédictions des différents modèles

Ce graphique permet de voir que malgré les observations précédentes, la courbe de Lorenz du modèle avec véhiculier (en vert) est très proche de celle des modèles benchmark (en bleu et violet) pour la partie droite de la courbe. Ainsi, bien que le modèle véhiculier sous-estime les sinistres les plus importants, il attribue la même part des coûts importants dans le montant total que les modèles benchmark. C'est sur les sinistres dont le coût est plus faible (les 75% d'exposition les plus faible) que la courbe de Lorenz du modèle véhiculier est plus haute que celle des autres modèles benchmark. Cela signifie qu'il surestime la part des sinistres dont le coût est faible par rapport aux modèles benchmark.

Comme le montre le graphique des quantiles, les trois modèles ont tendance a surestimer les sinistres dont le coût est faible et à sous-estimer ceux dont le coût est élevé. Ainsi la courbe de Lorenz montre que les trois modèles prédisent que la tranche des 10% des sinistres dont le coût est le plus élevé correspond à environ 35% du total des coûts des sinistres alors que pour les données observées, les 10% des sinistres les plus graves représentent 65% du coût des sinistres.

On peut toutefois nuancer les observations précédentes en observant la pertinence de la variable véhiculier en soi. Voici le tableau contenant l'AIC de modèles GLM à une seule variable.

Variable	AIC (Base d'entraînement)
Véhiculier	53 638.1
Prix	53 730.9
Puissance	53 830.8
Marque	53 940.2
Segment	54 111.4
Age du véhicule	54 173.0
Poids-Lourd ou voiture	54 201.4
Code Mosaic	54 281.1
Carburant	54 347.8
Nombre d'années sans sinistre en tort	54 350.7
Classe sociale	54 352.5
Age du conducteur	54 358.8
Ancienneté du contrat	54 364.7
Présence de gobox	54 400.2
Fractionnement	54 401.1
Année	54 405.4
Etat civil	54 410.0
Présence de garage	54 410.4
Nombre de places	54 413.8
Cylindrée	54 415.9
Nombre de conducteurs	54 419.4
Zonier	54 420.6
Masse	54 429.9

TABLE 3.19 – Importance des variables

On remarque que la variable véhiculier est celle qui permet de minimiser l'AIC dans un modèle à une seule variable. On peut en déduire que ce véhiculier est la variable la plus pertinente de toutes celles à disposition (y compris plus pertinent que chaque variable véhicule prise individuellement).

Conclusion partielle : Le véhiculier reste une variable très pertinente afin de prédire le coût des sinistres Vol, mais il demeure tout de même moins efficace que les quatre variables véhicule les plus pertinentes (c'est-à-dire le prix, le segment, la marque et la puissance du véhicule). Cela peut s'expliquer par le fait que la perte d'information lors du passage de variables *brutes* à un véhiculier est important. En effet, le graphique ci-dessous indique l'importance des variables (dont la méthodologie est détaillée en Annexe C) dans l'algorithme de classification utilisé pour le véhiculier optimal. 4 des 5 variables les plus importantes sont les 4 variables utilisée dans le modèle benchmark. Ainsi, ce sont d'une certaine manière les mêmes variables utilisées dans le modèle benchmark à 4 variables et celui avec véhiculier, et pourtant les résultats sont différents. On peut en conclure qu'il y a une perte d'information non négligeable lors de la construction du véhiculier.

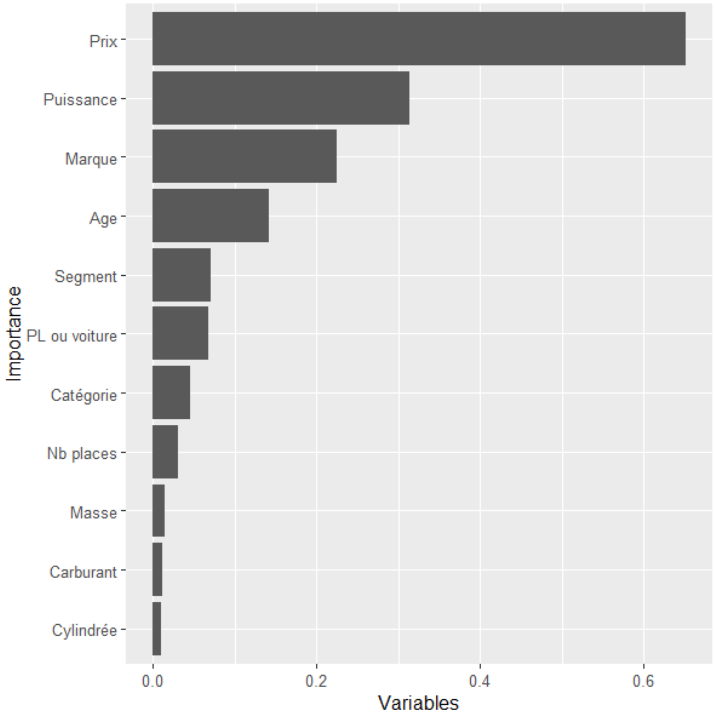


FIGURE 3.21 – Importance des variables dans la classification du véhiculier optimal

Conclusion

Les conclusions à tirer de ce mémoire sont multiples. Tout d'abord, notons que cette méthodologie permet de construire un véhiculier riche en information utile à la prédiction puisqu'il se montre plus pertinent que n'importe quelle autre variable. Mais force est de constater que cette variable synthétique n'est pas meilleure que les variables véhicules prises toutes ensemble. Ce n'est cependant pas surprenant, puisque toute synthèse de l'information conduit à en perdre une partie. Cette moins bonne performance ne permet pas de distinguer suffisamment les bons des mauvais risques. Cela posera problème pour un assureur lors de la tarification d'un produit, le risque étant de sous-tarifer les mauvais risques et de sur-tarifer les bons, ce qui fait courir un risque de pertes pour l'assureur mais aussi le risque d'attirer de plus en plus de mauvais risques au profit des bons et donc mettre en danger la solvabilité de l'assureur à long terme.

Une des explications de cette moins bonne performance du véhiculier pourrait résider dans l'hypothèse de base de ce modèle qui peut elle aussi être remise en question : l'idée selon laquelle les résidus du modèle sans variable véhicule correspondraient à l'effet véhicule. On peut aussi considérer qu'il s'agit non seulement de cela, mais aussi de bruit qu'il est difficile d'isoler, ce qui empêche par la suite de constituer des véhiculiers extrêmement performants.

Faut-il cependant écarter l'option du véhiculier ? Tout dépend de l'utilité qui en est faite et des enjeux à concilier pour une compagnie d'assurance. S'il s'agit de privilégier la capacité de prédiction à tout prix, alors le véhiculier n'est pas la meilleure solution. S'il s'agit de faire tourner régulièrement des modèles de tarification, alors le véhiculier peut s'avérer utile en termes de temps de calcul. Bien que sa conception soit longue, on peut considérer qu'un véhiculier bien construit sera robuste dans la durée et ne nécessitera pas de mise à jour trop régulière. Une fois cette construction effectuée, les modèles de tarification en deviendront bien moins coûteux en termes de temps de calcul, d'autant plus si le nombre de variables véhicule est grand à l'origine. La question est donc à étudier sur le long terme : si le véhiculier est robuste à long terme, la perte en temps de calcul sera compensée par le gain lors de prochains modèles, sachant que d'après nos observations, un modèle GLM avec un véhiculier déjà construit met 6 fois moins de temps à tourner qu'un modèle benchmark avec 4 variables véhicule et encore moins que le modèle avec toutes les variables véhicule.

On pourrait aussi supposer qu'une alternative à mi-chemin entre le modèle véhiculier et benchmark puisse être pertinente. On pourrait garder les variables véhicule les plus pertinentes de façon brute directement dans un modèle GLM et tenter de capter l'information utile des variables véhicules moins pertinentes au sein d'un véhiculier. Cela permettrait de s'assurer que l'on ne perd pas l'information portée par les variables les

plus pertinentes tout en réduisant la taille de la base de données pour les variables moins pertinentes.

Enfin, ce mémoire était aussi l'occasion de tester la pénalisation fused lasso, assez peu connue mais plutôt novatrice. Celle-ci se montre plus pertinente que la pénalisation lasso standard pour les deux garanties testées et permet de simplifier le modèle en regroupant de façon cohérente des modalités similaires. Cependant, étant donné le temps de calcul nécessaire, cette pénalisation semble pertinente si elle est précédée d'une première sélection de variables afin d'écarter celles qui ne semblent pas nécessaires.

Bibliographie

[1] Magali Ruimy (2017), *Elaboration d'un véhiculier en assurance automobile*, Mémoire ISFA.

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

[3] Teuvo Kohonen (2001), *Self-Organizing Maps*, 3th ed., No. 30 in Springer Series in Information Sciences, 2001, Berlin : Springer.

[4] Zou Hui, Trevor Hastie (2005), *Regularization and Variable Selection via the Elastic Net*, Journal of the Royal Statistical Society, Series B, Volume 67.

[5] Andrey Tikhonov (1963), *Solution of Incorrectly Formulated Problems and the Regularization Method*, Soviet Mathematics.

[6] Robert Tibshirani (1996), *Regression Shrinkage and Selection Via the Lasso*, Journal of the Royal Statistical Society, Series B.

[7] Brigitte Escofier, Jérôme Pagès (2008), *Analyses Factorielles Simples et Multiples*, Objectifs, méthodes et interprétation, 4e édition.

[8] Sander Devriendt, Katrien Antonio, Tom Reynkens, and Roel Verbelen (2020), *Sparse Regression with Multi-type Regularized Feature Modeling*, Insurance : Mathematics and Economics, Volume 96.

[9] Boris Delaunay (1934), *Sur la sphère vide*, Bulletin de l'Académie des sciences de l'URSS.

[10] Laurens van der Maaten, Geoffrey Hinton (2008), *Visualizing Data using t-SNE*, Journal of Machine Learning Research 9.

[11] Geoffrey Hinton, Sam Roweis (2002), *Stochastic Neighbor Embedding*, 15th International Conference on Neural Information Processing Systems.

[12] Matthieu Quilfen (2018), *Classification des Véhicules en Assurance Automobile*, Mémoire ISFA.

[13] Jean Lejay (2016), *Non-standard geographical smoothing approach*, Mémoire d'actuariat.

[14] Nynke C. Krol (2013), *Penalized logistic regression : a quadratic difference penalty*, Master Thesis at Universiteit Leiden.

[15] Leo Breiman, Jerome Friedman, Richard A. Olshen, Charles J. Stone (1984), *Classification and Regression Trees*, Wadsworth, Belmont.

[16] Thomas Grubinger, Achim Zeileis, Karl-Peter Pfeiffer (2014), *evtree : Evolutionary Learning of Globally Optimal Classification and Regression Trees in R*, Journal of Statistical Software Vol. 61.

[17] Tom Michoel (2016), *Natural coordinate descent algorithm for L1-penalised regression in generalised linear models*, Computational Statistics and Data Analysis 97.

[18] Anita Rahayu Purhadi Sutikno, Dedy Dwi Prastyo (2020), *Multivariate Gamma Regression : Parameter Estimation, Hypothesis Testing, and Its Application*, Symmetry 12.

[19] Michel Denuit, Dominik Sznajder, Julien Trufin (2019), *Model selection based on Lorenz and concentration curves, Gini indices and convex order*, Insurance : Mathematics and Economics, Vol. 89.

[20] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, Keith Knight (2005), *Sparsity and smoothness via the fused lasso*, Journal of the Statistical Royal Society, Series B, Vol. 67.

Annexes

A Compléments sur les modèles linéaires généralisés

On note :

$$\log \mathcal{L}_i = \log(c_\phi(y_i)) + \left(\frac{y_i\theta - a(\theta)}{\phi}\right)$$

$$\forall i \in \{1, \dots, n\} \quad \frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \mu_i} = \frac{\partial \mu_i}{\partial \eta_i} \times \frac{y_i - \mu_i}{V(Y_i)} X_{ij}$$

Les $\beta_1, \dots, \beta_d$ solutions du modèle linéaire généralisé vérifient donc :

$$\forall j \in \{1, \dots, n\} \quad \sum_{i=1}^{d} \frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \sum_{i=1}^{d} \frac{\partial \mu_i}{\partial \eta_i} \times \frac{y_i - \mu_i}{V(Y_i)} X_{ij} = 0$$

Comme la plupart du temps, ces équations ne peuvent pas être résolues numériquement, il existe plusieurs algorithmes permettant de résoudre ce problème :

- La méthode appelée *iteratively reweighted least squares* (IRLS) qui s'appuie sur la méthode des moindres carrés
- La méthode de descente de gradient et ses différentes variantes (Méthode de Newton, Descente de gradient stochastique, Gradient Boosting...)

GLM Gamma

La fonction de lien canonique est donc la fonction inverse et la fonction de perte à minimiser est donc la suivante :

$$\mathcal{L}(y_1, ..., y_n, \theta, k) = \prod_{i=1}^{n} f_Y(y_i, \theta, k) = \prod_{i=1}^{n} \frac{y_i^{k-1} e^{-\frac{y_i}{\theta}}}{\Gamma(k)\theta^k}$$

$$\mathcal{L}(y_1, ..., y_n, \theta, k) = \prod_{i=1}^{n} f_Y(y_i, \theta, k) = \prod_{i=1}^{n} \frac{y_i^{k-1}}{\Gamma(k)} \exp(-\frac{y_i}{\lambda} - k\log(\lambda))$$

avec ici

$$c_\phi(y) = \frac{y^{k-1}}{\Gamma(k)}; \quad \theta = -\frac{1}{\lambda}; \quad a(\theta) = -k\log(-\theta); \quad \phi = 1$$

B Mesures de performances des modèles

Dans le cadre de ce mémoire, trois mesures de performances des modèles GLM sont utilisées : le Mean Squared Error (MSE), l'Area Between Curves (ABC) et l'Akaike Information Criterion (AIC).

B.1 Mean Squared Error

Le Mean Squared Error, comme son nom l'indique, est une mesure de performance correspondant à la moyenne des erreurs d'un modèle. Notons $Y = (y_1, \ldots, y_n)$ la variable cible et $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)$ les prédictions du modèle. On définit le Mean Squared Error de la façon suivante :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ainsi, plus le MSE est faible, plus son erreur l'est et meilleur est le modèle.

B.2 Area Between Curves

Avant de définir l'Area Between Curves (ABC), il faut commencer par définir l'indice de Gini et la courbe de Lorenz. La courbe de Lorenz d'une variable aléatoire X de loi de probabilité p, est notée L_p et est définie de la façon suivante :

$$\begin{aligned} L_p :&[0,1] \to [0,1] \\ &x \mapsto \frac{\int_0^x q(t)dt}{\int_0^1 q(t)dt} \end{aligned}$$

avec q la fonction quantile. En d'autres termes, dans le cadre d'une étude de coûts de sinistres, la courbe de Lorenz consiste à exprimer la part correspondant aux $x\%$ des coûts les plus faibles dans le total des coûts observés.

L'indice de Gini est une mesure qui permet notamment de donner des indications sur la distribution d'une variable. On peut définir l'indice de Gini d'après la courbe de Lorenz de la façon suivante :

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

avec A et B les aires décrite par le schéma suivant :

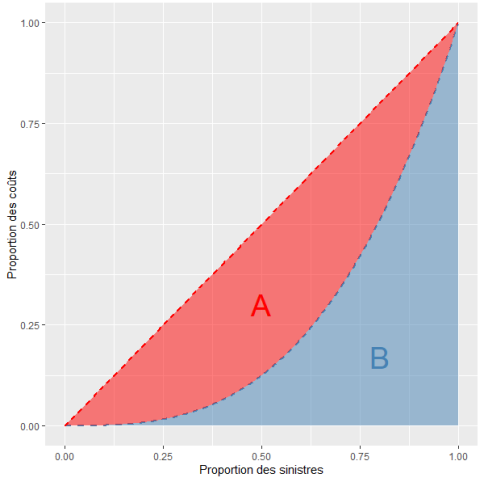


FIGURE 22 – Indice de Gini

A correspond à l'aire entre la droite d'équation $y = x$ et la courbe de Lorenz et B l'aire sous la courbe de Lorenz. Ainsi plus l'aire A est grande, plus B est faible et plus l'indice de Gini est proche de 1. De plus, si A est grand, la courbe de Lorenz s'éloigne de la droite $y = x$ et plus la distribution des coûts est *inégalitaire* dans le sens où les sinistres graves représentent une grande part du coût total des sinistres.

Or cet indice de Gini ne permet pas d'indiquer autre chose que la structure d'une distribution et donc ne permet pas de mesurer la performance d'un modèle. L'idée est de comparer la courbe de Lorenz (et donc l'indice de Gini) d'un modèle avec celle des observations réelles. Dans le cas de ce mémoire, il s'agit de comparer la courbe de Lorenz des prédictions d'un modèles avec les coûts observés sur une base de données test.

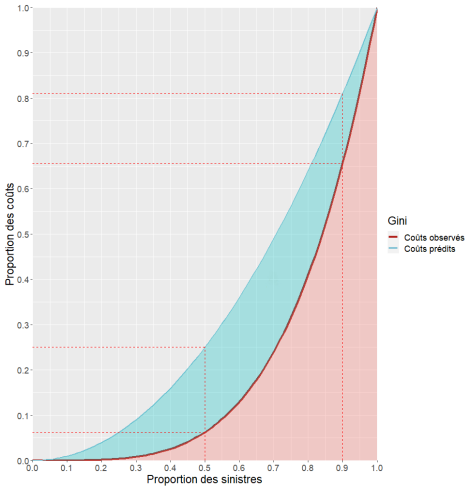


FIGURE 23 – Courbe de Lorenz d'un modèle prédictif et des observations réelles

Comparer les courbes de Lorenz d'un modèle avec les observations réelles permet de comparer leurs distributions. Plus la courbe de Lorenz des prédictions se rapproche de celle des observations, meilleur est le modèle puisqu'il arrive à en copier la distribution. Ainsi, l'objectif est de rendre l'aire bleue du

graphique ci-dessus la plus petite possible. Ainsi dans l'exemple précédent, 50% des coûts observés des sinistres les plus faibles représentent environ 6% du coût total des sinistres. Or dans le modèle, ces 50% représentent 25% du coût total prédit. De même, les 10% des sinistres observés dont le coût est le plus important représentent 34% du coût total des sinistres observés alors que d'après le modèle cela ne représente que 11%.

Une solution consisterait donc à s'intéresser non pas à l'indice de Gini tel quel, mais plutôt à l'aire de l'aire notée C dans le schéma ci-dessous.

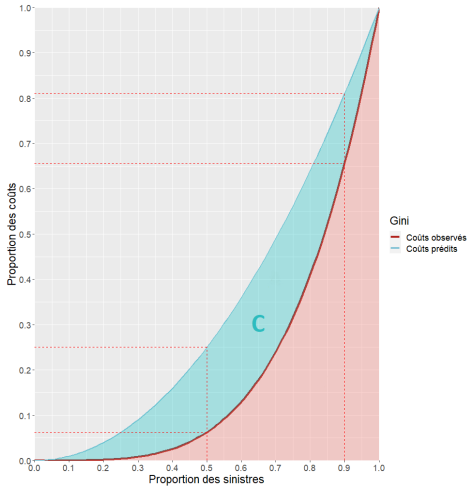


FIGURE 24 – Courbe de Lorenz d'un modèle prédictif et des observations réelles

En effet, plus cette aire C est petite, meilleur est le modèle. Notons cependant, qu'il ne suffit pas d'avoir des indices de Gini identiques pour que les courbes de Lorenz se superposent et que l'aire C soit nulle.

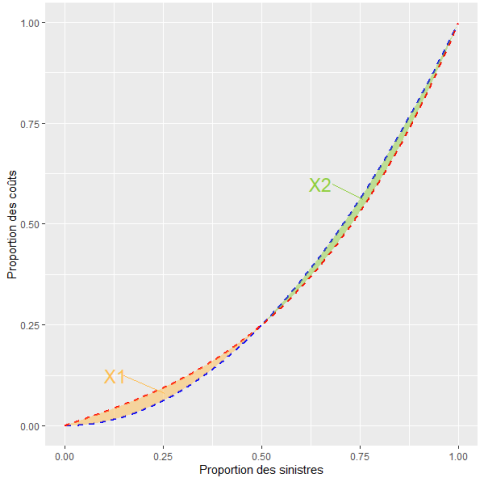


FIGURE 25 – Courbe de Lorenz d'un modèle prédictif et des observations réelles

Dans ce schéma, les aires $X1$ et $X2$ sont identiques et se compensent ainsi les deux modèles ont le même indice de Gini mais pas la même courbe de Lorenz. Ainsi, on pourrait calculer un autre indicateur,

l'Area Under Curve (ABC) défini de la manière suivante :

$$ABC = \int_0^1 |L(t) - L_*(t)|dt$$

avec L la courbe de Lorenz du modèle en question et L_* celle des observations. Ainsi, on obtiendra une nouvelle mesure qui varie entre 0 (modèle parfait pour lequel la courbe de Lorenz se superpose à celle des observations) et 0.5 (cas extrême où une des deux courbes de Lorenz correspond à un indice de Gini de 0 et l'autre 1).

B.3 Akaike Information Criterion

Contrairement aux deux mesures précédentes, l'Akaike Information Criterion est une mesure de performance calculée que sur la base d'entraînement du modèle. Et contrairement aux mesures précédentes, celle-ci prend en compte la complexité du modèle. L'AIC est défini de la façon suivante

$$AIC = 2k - 2\log(\mathcal{L})$$

avec k le nombre de paramètres à estimer et $\mathcal{L}$ la vraisemblance du modèle en question. Ainsi, plus le nombre de paramètre est faible ou bien la vraisemblance grande, meilleur est le modèle et plus l'AIC est faible.

C Variable Importance Plot

Lors de la construction d'un arbre, on peut se demander quelles sont celles qui permettent d'améliorer l'ajustement de l'arbre. Une solution consiste à observer dans l'arbre les variables les plus utilisées et le plus haut dans l'arbre lors de la segmentation. Or, pour un nœud donné, seule la variable et le seuil maximisant l'impureté (dans le cadre de ce mémoire) sont choisis, les autres variables qui réduisent elles aussi l'impureté (certes moins) ne sont pas visibles. La notion d'importance des variables, développée par Breiman ([15], permet de tenir compte des variables cachées.

Pour un noeud donné t d'un arbre, on définit des substituts : pour chacune des autres variables ne minimisant pas l'impureté de ce noeud, donc non choisies, on détermine le seuil qui minimise l'impureté. Pour une variable X_i donnée prenant des valeurs dans l'intervalle V_i, on définit la division de substitution $\delta_i(t)$ comme étant le seuil permettant de se rapprocher au plus près de la segmentation optimale notée $\delta_*(t)$. On considérera que les divisions $\delta_i(t)$ et $\delta_*(t)$ sont d'autant plus proches qu'elles répartissent les mêmes observations dans les feuilles filles. Notons t_i^g et t_i^d les feuilles filles issues de la division $\delta_i(t)$ et t_*^g et t_*^d celles de la division $\delta_*(t)$. On définit donc la probabilité $p(\delta_i(t), \delta_*(t))$ que la division $\delta_i(t)$ prédise la même chose que $\delta_*(t)$ de la façon suivante :

$$p(\delta_i(t), \delta_*(t)) = p_g(\delta_i(t), \delta_*(t)) + p_d(\delta_i(t), \delta_*(t))$$

avec $p_g(\delta_i(t), \delta_*(t))$ (respectivement $p_d(\delta_i(t), \delta_*(t))$) la probabilité que $\delta_i(t)$ et $\delta_*(t)$ envoient les mêmes

observations dans le noeud fils gauche (respectivement droit). Ces deux probabilités peuvent être estimées de la manière suivante :

$$p_g(\delta_i(t), \delta_*(t)) = \frac{\#\{x_j : x_j \in t_*^g \cap t_i^g\}}{\#\{x_j : x_j \in t\}}; \quad p_d(\delta_i(t), \delta_*(t)) = \frac{\#\{x_j : x_j \in t_*^d \cap t_i^d\}}{\#\{x_j : x_j \in t\}}$$

Notons ensuite pour un noeud donné t et une variable donnée X_i, $\tilde{\delta}_i(t)$ définie par :

$$\tilde{\delta}_i(t) = \underset{\delta_i(t) \in V_i}{\operatorname{argmax}}\, p(\delta_i(t), \delta_*(t))$$

Enfin l'importance de la variable X_i, notée $VI(X_i)$ correspond à :

$$VI(X_i) = \sum_{t \in T \setminus \tilde{T}} (\Delta R(\tilde{\delta}_i(t), t))$$

$$\Delta R(\tilde{\delta}_i(t), t) = R(t) - R(\tilde{t}_i^g) - R(\tilde{t}_i^d)$$

avec $T \setminus \tilde{T}$ l'ensemble des nœuds non terminaux, $\tilde{t}_i^g$ et $\tilde{t}_i^d$ les nœuds fils issus de $\tilde{\delta}_i(t)$

$$R(t) = \sum_{i : X_i \in t} (Y_i - \bar{Y}(t))^2$$

$\bar{Y}(t)$ correspondant à la prédiction de l'arbre pour le nœud t.

D V de Cramer

Le V de Cramer est une mesure statistique permettant d'évaluer la proximité entre deux variables. Cette méthode est plus pratique que le test du Khi-deux ordinaire dont il est issu puisqu'il varie entre 0 et 1 (0 correspondant à une faible proximité et 1 une très forte). En effet, ce-dernier est très sensible à la taille de l'échantillon et n'est pas interprétable tel quel. Le V de Cramer permet de pallier à ce problème en normalisant cette statistique du Khi-deux. Soient X et Y deux variables dont on étudie la proximité ; n_X et n_Y étant le nombre de modalités respectif de chaque variable. La statistique du V de Cramer est définie de la façon suivante :

$$V = \sqrt{\frac{\mathcal{X}^2}{n(\min(n_X, n_Y) - 1)}}$$

avec n le nombre d'observations des variables X et Y, et $\mathcal{X}^2$ correspondant à une mesure du Khi-deux entre ces deux variables, c'est-à-dire

$$\mathcal{X}^2 = \sum_{i \in V_X ; j \in V_Y} \frac{\left(n_{i,j} - \frac{n_{X_i} n_{Y_j}}{n}\right)^2}{\frac{n_{X_i} n_{Y_j}}{n}}$$

avec V_X et V_Y l'ensemble des modalités prises par X et Y respectivement ; $n_{X_i,Y_j}, n_{X_i}, n_{Y_j}$ le nombre d'apparition respectif de (X_i, Y_j), X_i et Y_j.

E Analyse en composantes principales

L'Analyse des Composantes principales (ACP) est une méthode très répandue de réduction de dimension pour des variables continues. Elle peut avoir deux utilités selon le point de vue adopté :

- Elle étudie le lien entres deux variables et détermine celles dont la corrélation est positive. L'ACP permet aussi de trouver des variables synthétiques (appelées composantes principales) permettant de résumer l'information contenue par l'ensemble des variables et de respecter la structure du nuage de points d'origine.
- Elle permet aussi de savoir quelles observations se ressemblent, de déterminer des groupes homogènes d'individus.

Supposons que l'on ait n observations de d variables quantitatives $X = (X_1, \ldots, X_d)$. On notera x_{ij} la i-ème observation de la variable j. Afin d'étudier de façon identique ces variables, on les retraite en les centrant et les réduisant afin de donner la même importance à chaque variable :

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

avec

$$\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}; \quad \sigma_j = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}$$

On note $Y_j = (y_{1j}, \ldots, y_{nj})$, la j-ième colonne de la matrice contenant les observations centrées-réduites.

Proximité entre deux observations : On définit la mesure de la ressemblance entre deux points i et i' par la distance euclidienne usuelle définie de la façon suivante :

$$d^2(i, i') = \sum_{k=1}^{d}(y_{ik} - y_{i'k})^2$$

Proximité entre deux variables : Afin de mesurer la proximité entre deux variables i et j, nous utilisons le coefficient de corrélation linéaire r, défini de la façon suivante pour le couple de variable (i, j) :

$$\begin{aligned} r(i,j) &= \frac{\text{covariance}(i,j)}{\sqrt{Var(i)Var(j)}} \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_{ik} - \bar{x}_i}{\sigma_i}\right)\left(\frac{x_{jk} - \bar{x}_j}{\sigma_j}\right) \end{aligned}$$

Dans le cadre de ce mémoire, c'est le point de vue des variables qui nous intéresse : il s'agit de trouver des variables synthétiques permettant de représenter au mieux le nuage de points dans un espace réduit. Dans ce cas, on représente l'ensemble des variables dans un espace vectoriel de dimension n noté R^n : chaque variable a n coordonnées correspondant aux différentes valeurs prises par l'individu pour la

variable. Les coordonnées de la variable i dans R^n sont donc Y_i.

On mesure ensuite le cosinus de l'angle entre deux variables i et j dans cet espace R^n, c'est-à-dire l'angle formé par $\vec{Oi}$ et $\vec{Oj}$:

$$\cos(\theta_{ij}) = \frac{\langle Y_i, Y_j \rangle}{\|Y_i\| \|Y_j\|}$$

Ici, les données étant centrées et réduites, on obtient :

$$\begin{aligned} \cos(\theta_{ij}) &= \frac{\sum_{k=1}^{n} y_{ki} y_{kj}}{\sqrt{\sum_{k=1}^{n} y_{ki}^2} \sqrt{\sum_{k=1}^{n} y_{kj}^2}} \\ &= r(i,j) \end{aligned}$$

Il s'agit donc de chercher un ensemble de variables synthétiques notés $(v_s)_{s\in\{1,...,S\}}$ correspondant à des directions dans R^n. Elles sont choisies de façon à maximiser l'inertie par rapport à l'origine de la projection du nuage de points des variables sur v_s et sont contraintes de façon à ce que chaque nouvelle composante soit orthogonale aux précédentes.

Les variables étant centrées et réduites, leur projection sur la première composante v_1 est égale à leur coefficient de corrélation avec elle. Donc v_1 est la direction qui maximise :

$$\sum_{k=1}^{d} OH_k^2$$

Puis la seconde composante principale v_2 sera la direction orthogonale à v_1 qui maximise de nouveau l'inertie. Et ainsi de suite pour les autres composantes principales.

F Analyse des correspondances multiples

L'Analyse des correspondances multiples est une méthode de réduction de dimension analogue à l'Analyse en Composantes Principales dédiée aux variables qualitatives. Elle peut être étudiée de deux façons différentes : du point de vue des individus (on analyse la ressemblance entre les observations) ou bien de celui des variables (on se concentre sur la proximité entre variables).

Supposons que l'on ait n observations de d variables $X = (X_1, ..., X_d)$. Notons x_{ij} la i-ème observation de la variable j. La première étape de l'ACM est de construire le tableau disjonctif complet (TDC) de cette base de données. Il s'agit de créer une colonne par modalité de chaque variable qualitative puis de la valoriser à 1 si l'individu i possède cette modalité et 0 sinon. Ainsi, on obtient une table avec le même nombre de lignes que précédemment, mais le nombre de colonne sera différent. Si K_i correspond au nombre de modalités de la variable i, alors, le tableau disjonctif complet aura $K = K_1 + ... + K_d$ colonnes. On notera $(x'_{ij})_{i=1...n;j=1...K})$ les observations de ce TDC et n_i le nombre d'observations de

la modalité i.

L'objectif de l'ACM est de mesurer la variabilité des individus, c'est-à-dire le degré de ressemblance et de dissemblance entre les différentes observations. Pour cela, on doit extraire les principales dimensions de la variabilité des individus et d'extraire des variables synthétiques combinant les variables X.

Tous les individus ont un poids identique c'est-à-dire $\frac{1}{n}$. Une modalité possédée par beaucoup individus (voire tous) ne caractérisera pas beaucoup un individu. Au contraire, si un individu possède une modalité peu fréquente, alors celle-ci le caractérisera beaucoup plus. On divise donc chaque valeur du tableau disjonctif complet x'_{ij} par le proportion d'observation de la modalité p_j pour tenir compte de la fréquence d'apparition de chaque modalité et on note $y_{ij} = \frac{x'_{ij}}{p_j} - 1$. La moyenne des $\frac{x'_{ij}}{p_j}$ étant égale à 1, on obtient des observations centrées.

De même que pour l'ACP, on peut observer le nuage des modalités des variables dans l'espace des observation noté R^n.

La variance de la modalité k est définie par :

$$Var(k) = d^2(k, O) = \frac{1}{n}\sum_{i=1}^{n} y_{ik}^2 = \frac{1}{p_k} - 1$$

avec O l'origine de l'espace R^n des individus. Ainsi plus la modalité k est rare, plus elle est loin de l'origine. L'inertie de la modalité k est définie de la façon suivante :

$$Inertie(k) = \frac{p_k}{d} d^2(k, O) = \frac{1 - p_k}{d}$$

Enfin, la distance entre deux modalités k et k' vaut :

$$d^2(k, k') = \sum_{i=1}^{n} (x'_{ik} - x'_{ik'})^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_{k'} p_k}$$

avec $p_{kk'}$ la proportion des observations qui prennent à la fois k et k'.

Notons $F_s(i)$ la projection de la ligne i sur l'axe de rang s de R^n et $G_s(j)$ la projection de la colonne j sur l'axe de rang s de R^n. L'objectif de l'ACM est de trouver des axes orthogonaux sur lesquels projeter le nuage de points tout en maximisant son inertie dans le nouvel espace. Il s'agit de maximiser, pour l'axe s :

$$\frac{1}{d}\sum_{j=1}^{d} \eta^2(F_s, j)$$

avec $\eta^2(F_s, j)$ le rapport de corrélation au carré entre une variable j et le facteur F_s,

$$\eta^2(F_s, j) = \frac{\sum_{k=1}^{K_j} \frac{n_k}{n} F_s(G_k))^2}{\lambda_s}$$

G Tableau des coordonnées et des contributions des variables véhicule à l'AFDM

Variable	Coordonnées			Contribution			Cosinus		
	Dim.1	Dim.2	Dim.3	Dim.1	Dim.2	Dim.3	Dim.1	Dim.2	Dim.3
Prix	0.14	0.7	0.0028	3.8	20	0.13	0.019	0.48	8e-06
Cylindrée	1.1e-06	8.6e-06	0.00063	3e-05	0.00025	0.029	1.2e-12	7.3e-11	4e-07
Puissance	0.12	0.71	0.0005	3.3	20	0.023	0.014	0.5	2.5e-07
Masse	5.5e-05	3.4e-05	0.0003	0.0015	0.00098	0.013	3.1e-09	1.2e-09	8.8e-08
Catégorie	0.78	0.14	9.2e-05	22	4.1	0.0042	0.3	0.01	4.3e-09
Nombre places	0.63	0.17	0.38	17	4.9	17	0.057	0.0042	0.021
Marque	0.3	0.5	0.42	8.4	14	19	0.0018	0.0049	0.0034
Segment	0.77	0.64	0.68	21	18	31	0.04	0.028	0.031
Carburant	0.021	0.034	0.22	0.58	0.99	9.8	5.6e-05	0.00015	0.0058
Age du véhicule	0.0012	0.06	0.024	0.032	1.7	1.1	6.9e-08	0.00018	3e-05
PL ou auto	0.85	0.52	0.48	23	15	22	0.08	0.03	0.026

H Stochastic Neighbor Embedding

La méthode *Stochastic Neighbor Embedding* est une méthode de visualisation de données et de réduction de dimension développée par Hinton et Roweis [11]. L'idée principale est de convertir la distance euclidienne entre deux points en termes de probabilité. Notons $X = (x_1, \ldots, x_n)$ les observations dans l'espace de départ (de grande dimension d) et $Y = (y_1, \ldots, y_n)$ les observations correspondantes dans l'espace réduit d'arrivée (de faible dimension). On note $p_{j|i}$ la probabilité que le point x_i choisisse pour voisin le point x_j, en supposant que cet événement suive une loi gaussienne centrée en x_i et de variance σ_i^2. Ainsi, plus le point x_j se trouvera proche de x_i plus cette probabilité est grande. De même, on définit $q_{j|i}$ la probabilité conditionnelle que le point y_i choisisse le point y_j comme voisin dans l'espace d'arrivée, d'après une loi gaussienne centrée en y_i et de variance $\frac{1}{\sqrt{2}}$ par simplicité. On peut donc écrire

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

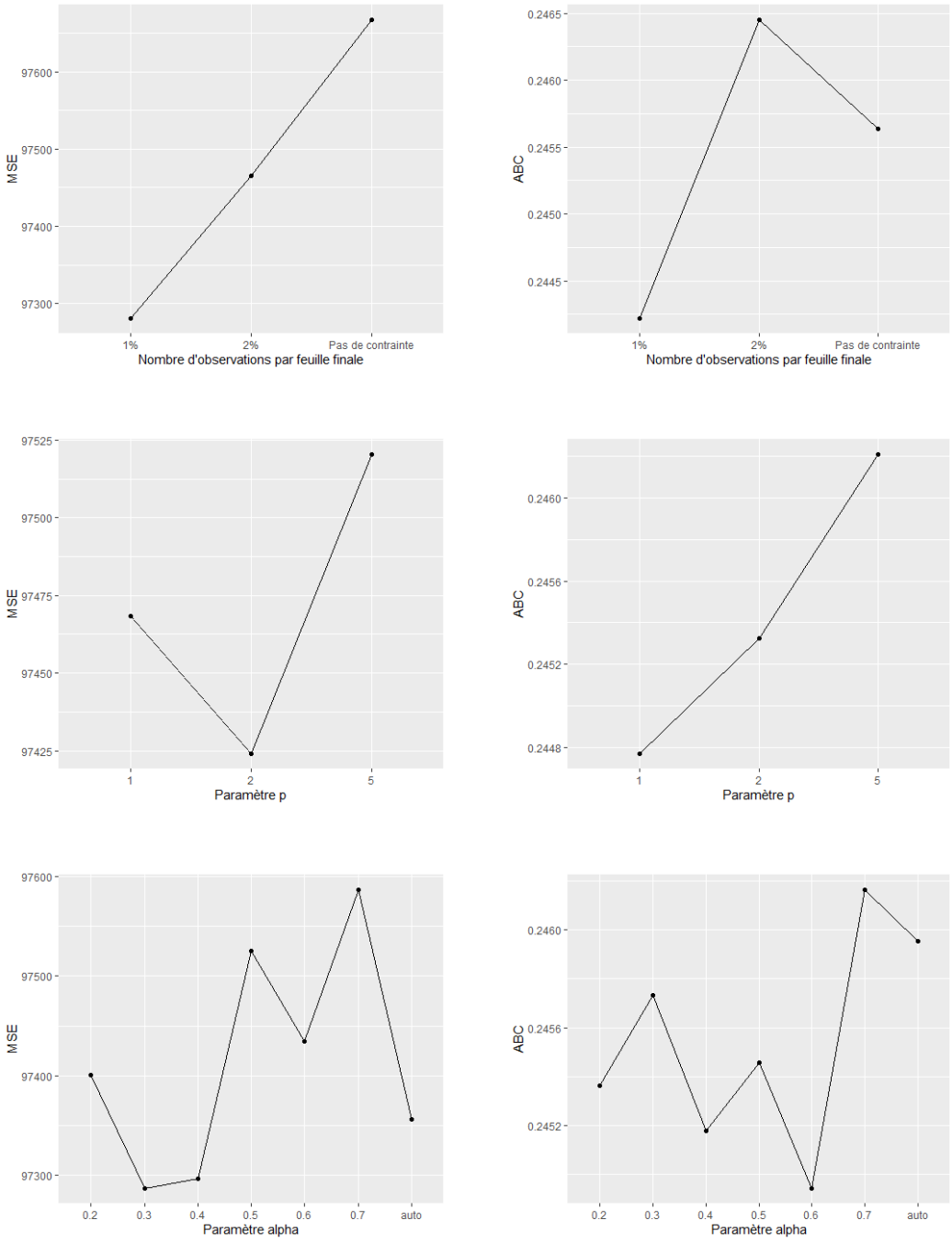
$$q_{j|i} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq i} \exp(-||y_i - y_k||^2)}$$

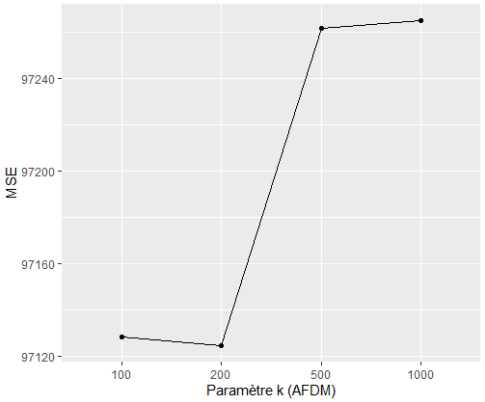
Ainsi, s'agissant de garder une structure identique entre l'espace de départ et d'arrivée, l'objectif est donc d'obtenir une égalité entre $p_{j|i}$ et $q_{j|i}$. La mesure à minimiser est la divergence de Kullback-Leibler entre $p_{j|i}$ et $q_{j|i}$. La fonction de coût à minimiser (souvent par descente de gradient) est donc la suivante :

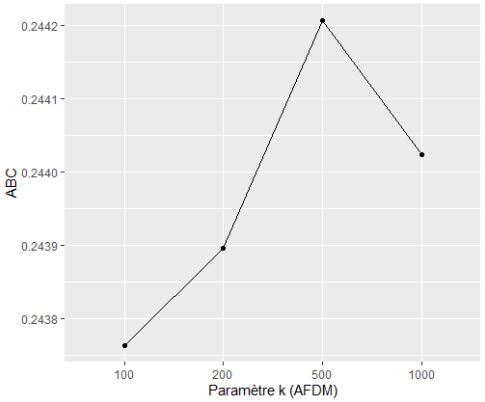
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

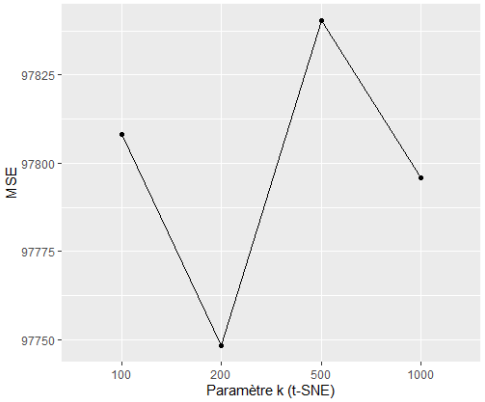
avec P_i (respectivement Q_i) la distribution sachant x_i (resp. y_i) sur l'ensemble des données de départ (resp. d'arrivée).

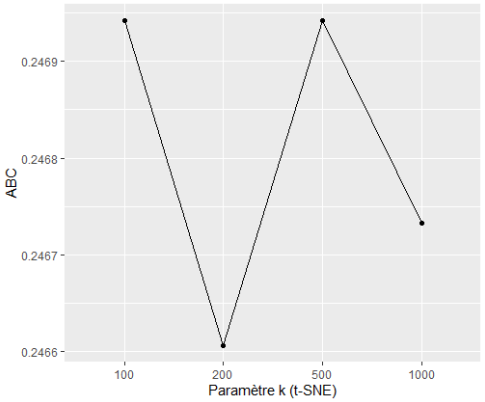
I Performances du véhiculier Bris-de-Glace en fonction des différents paramètres

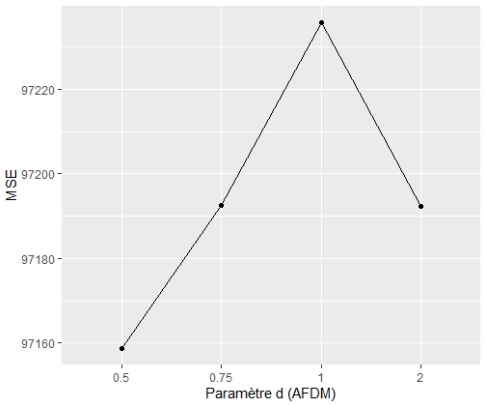


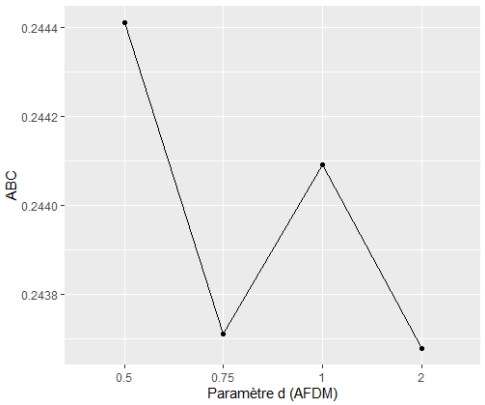

MSE
97240
97200
97160
97120
100
200
500
1000
Paramètre k (AFDM)

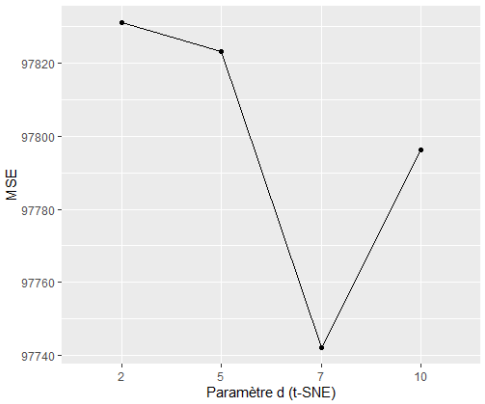

ABC
0.2442
0.2441
0.2440
0.2439
0.2438
100
200
500
1000
Paramètre k (AFDM)

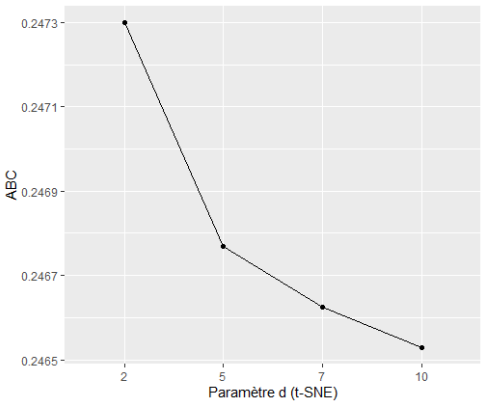

MSE
97825
97800
97775
97750
100
200
500
1000
Paramètre k (t-SNE)


ABC
0.2469
0.2468
0.2467
0.2466
100
200
500
1000
Paramètre k (t-SNE)


MSE
97220
97200
97180
97160
0.5
0.75
1
2
Paramètre d (AFDM)


ABC
0.2444
0.2442
0.2440
0.2438
0.5
0.75
1
2
Paramètre d (AFDM)


MSE
97820
97800
97780
97760
97740
2
5
7
10
Paramètre d (t-SNE)


ABC
0.2473
0.2471
0.2469
0.2467
0.2465
2
5
7
10
Paramètre d (t-SNE)