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

Par : **Matthieu Quilfen**

Titre : **Classification des Véhicules en Assurance Automobile**

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

Membres présents du jury de l'IA *Signature*

Entreprise : *DIRECT ASSURANCE*
Nom : Isabelle Antoine
Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise
Nom : Catherine Khounlivong
Signature :

Invité
Nom :
Signature :

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

Signature du responsable entreprise

Secrétariat :
Mme Christine DRIGUZZI

Signature du candidat

Bibliothèque :
Mme Patricia BARTOLO

Résumé

Mots clés : Assurance non vie, marché du direct, sous-périls, classification de véhicules, analyse de données, lissage spatial.

Les assureurs automobiles font face à une concurrence très rude et se doivent de proposer un tarif adapté afin d'éviter l'anti-sélection. Les modèles de tarification nécessitent d'être mis à jour régulièrement et l'estimation du risque doit se faire de la façon la plus correcte possible.

La Responsabilité Civile Matérielle (RCM) est l'une des garanties principales des contrats automobiles et représente un volume conséquent dans la mesure où elle est obligatoire. Deux typologies de sinistres sont identifiables au sein de cette garantie : les sinistres relevant de la convention IRSA et ceux indemnisés au coût réel. Le présent mémoire porte sur la revue de la RCM et la sophistication des modèles, d'une part avec une segmentation de périls, et d'autre part avec la création d'une classification de véhicule.

Dans un premier temps la prime pure RCM sera segmentée en sous-périls. Deux modèles de fréquence (IRSA et non IRSA) et un modèle de coût moyen (non IRSA) seront créés, et l'amélioration de cette nouvelle structure sera prouvée.

Dans un second temps, la méthodologie de création d'un véhiculier sera exposée. Elle reprend la théorie de classification géographique et tente de récupérer un signal expliqué par les variables véhicules dans les résidus d'un modèle GLM [1]. La grande problématique de la classification de véhicules est d'obtenir une carte, i.e. une représentation spatiale des véhicules et de définir une relation de voisinage entre chacun d'entre eux. L'Analyse Factorielle de Données Mixtes et la Triangulation de Delaunay permettent de répondre à ce besoin. Plusieurs véhiculiers seront créés sur le modèle de fréquence IRSA en testant différentes méthodes de classification. Il s'agira alors de choisir celui permettant d'améliorer le plus les performances du modèle étudié.

1. Modèle Linéaire Généralisé

Abstract

Key Words : Non-life insurance, direct market, peril segmentation, vehicle classification, data mining, spatial smoothing.

Car insurance is a very competitive market and insurers have to provide a pricing adapted to each policy in order to avoid anti-selection. Models have to be updated regularly and the risk estimation has be be the finest possible.

The Third Party Liability Cover for Material Damages is one of the main covers in car insurance and the exposure is quite high as it is mandatory in the French market. There are two types of claims : those coming under the IRSA convention with fixed indemnity amounts and those paid with real cost. This study will cover the update and the sophistication of TPL Material models with peril segmentation on one hand, and with the creation of a vehicle classification on the other hand.

First, the TPL Material pure premium will be segmented in IRSA and non IRSA perils. Two frequency models and one severity model will be created (as the IRSA fee is constant, there is no need to build a model for IRSA average cost). It will then be possible to measure the improvement of the new pricing structure.

The methodology of vehicle classification creation will be discussed in a second part. It relies on geographical classification theory and tries to pick up a signal explained by vehicle variables in GLM residuals. The problem is that a map of vehicle is needed, i.e. a spatial representation, and a definition of neighbourhood between vehicles. It is possible to do so with multivariate analysis with mixed quantitative and the Delaunay triangulation. Various vehicle classification will be created on the IRSA Frequency model testing multiple clustering methods. The classification that improves the most the initial model will be retained.

Note de Synthèse

Le présent mémoire porte sur la revue des modèles de la garantie Responsabilité Civile Matérielle en assurance automobile de particuliers. Il a pour objectif de sophistiquer la modélisation à travers une segmentation des périls et d'innover en créant une classification de véhicules à l'aide d'outils d'analyse de données et un lissage effectué sur les résidus d'un modèle linéaire généralisé de prime pure.

Contexte et objectifs

Les assureurs automobiles font face à une concurrence très rude à cause de stratégies commerciales développées par certains, à l'émergence de nouveaux acteurs ou aux bancassureurs qui bénéficient déjà d'un portefeuille de clients. Le marché du direct est d'autant plus soumis à cette concurrence qu'il mue de façon plus rapide que le marché traditionnel. L'assureur doit alors innover pour conserver ses parts de marché, notamment lorsqu'il s'agit de sélection des risques à travers sa segmentation tarifaire. Une tarification ajustée permet de s'affranchir du phénomène d'anti-sélection. Ceci implique une revue récurrente des modèles, ainsi que le test de nouvelles structures tarifaires ou de variables de segmentation.

L'objectif de ce mémoire est la revue des modèles de Responsabilité Civile Matérielle avec des données récentes, ainsi que d'innover, d'une part en effectuant une sous-segmentation des périls, d'autre part en créant une classification de véhicules.

Segmentation des périls

La modélisation de la prime pure Responsabilité Civile Matérielle est habituellement faite (au sein d'AXA) à l'aide des Modèles Linéaires Généralisés. Un modèle de fréquence est construit en prenant en compte la totalité des sinistres de la garantie. Le coût moyen est une constante calculée directement à partir de la charge et du nombre de sinistres, puis un coefficient d'ajustement permet de prendre en compte les effets de marché ou l'inflation. Ceci s'explique par la présence de la convention IRSA.

Les sinistres peuvent être traités de deux façons en fonction de leur typologie et de l'assureur du tiers. Certains sont régis par la convention IRSA afin d'accélérer l'indemnisation de l'assuré : l'assureur non responsable rembourse son assuré au coût réel et peut effectuer un recours auprès de l'assureur responsable. Ce recours est un forfait fixe réévalué chaque année. Ce type de sinistre représente 75% du total des sinistres RCM. Lorsque la convention ne s'applique pas, e.g. pour des sinistres dommages aux biens, des sinistres internationaux ou lorsque l'assureur tiers ne fait pas partie de la convention IRSA, le sinistre est payé au coût réel par l'assureur responsable. Il en résulte une distribution des coûts moyens polluée de montants forfaitaires, et il n'est pas possible de faire un GLM traditionnel.

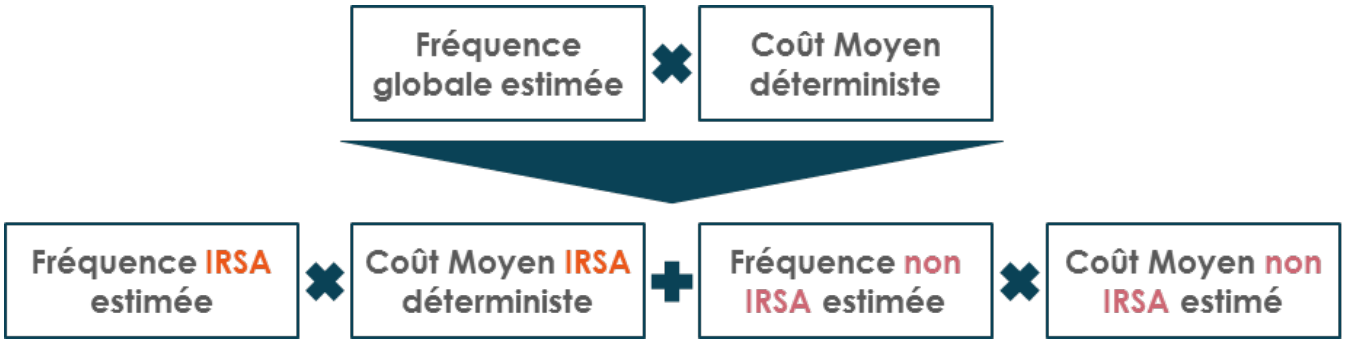


FIGURE 1 – Processus de segmentation des périls

Afin de proposer une segmentation du coût moyen, l'idée est de séparer les sinistres en deux types (IRSA et non IRSA) : il s'agit de la segmentation des périls. Un modèle de propension aurait été envisageable, cependant une étude de la fréquence révèle que les profils sinistrés IRSA sont bien différents des profils sinistrés non IRSA. Il est alors possible de décomposer la prime pure en deux modèles GLM de fréquence supposée suivre une loi de poisson, un modèle GLM de sévérité supposée suivre une loi Gamma, et une constante pour le coût moyen des sinistres IRSA.

Création de la base de données

La création de la base de données nécessite une étude préalable du périmètre de l'étude. Le modèle de fréquence actuel a été calibré à l'aide de données de 2012 à 2014. Il conviendra de travailler sur des données plus récentes, reflétant les changements de marché et de mix du portefeuille. Le déroulé, i.e. le vieillissement, des sinistres IRSA étant suffisamment court, le périmètre choisi est l'ensemble des années 2014 à 2016, soit trois ans, permettant de vérifier les hypothèses de robustesse temporelle lors de la création des modèles. Quant aux sinistres non IRSA, le déroulé est plus long, i.e. les sinistres mettent plus de temps à se stabiliser. Ceci peut s'expliquer par l'attente nécessaire pour recevoir des procès-verbaux de la part des forces de l'ordre dans le cas de sinistres dommages au bien ou carambolages. Le périmètre choisi est alors de 2013 à 2016 pour le modèle de fréquence et de coût moyen.

La base est constituées d'images de risques issues d'une fusion de bases d'informations contrats, personnes, sinistres, géographiques et enfin véhicules. Chaque ligne est considérée comme un risque couvert à part entière avec une période d'exposition. Une attention particulière a été donnée au traitement de la base véhicule dans la mesure où l'étude se concentre sur la création d'un véhiculier.

La base est échantillonnée de façon aléatoire en 80% (échantillon de modélisation) nécessaires à la création des modèles et 20% (échantillon de validation) permettant de mesurer la qualité des modèles créés. Un quart de l'échantillon de modélisation servira également à la construction du zonier, et aura l'appellation échantillon de détermination du lissage.

Résultats de la segmentation des périls

Les trois modèles sont créés à l'aide des GLM et les variables sont ajoutées l'une après l'autre tout en respectant quatre tests : l'importance des relativités, le test de robustesse temporel, les tests statistiques et le jugement.

Le modèle de prime pure IRSA contient 24 variables et le modèle de prime pure non IRSA (fréquence x sévérité) en contient 32. Ces chiffres sont à comparer aux 26 variables du modèle initial. La comparaison entre le modèle de prime pure initial et le modèle final se fait à l'aide du coefficient de GINI. Plus ce dernier est élevé, plus le modèle est adapté aux données. La nouvelle décomposition de la prime pure avec la segmentation des périls obtient un meilleur résultat que la modélisation précédente, en notant une forte amélioration de l'ordre de 49% sur l'échantillon de modélisation et de 36% sur l'échantillon de validation.

Les objectifs de mise à jour des modèles avec des données plus récentes et la nouvelle approche de modélisation avec une segmentation des périls sont atteints. Il est alors possible de se concentrer sur la partie innovation avec la création d'une classification de véhicules.

Création de classifications de véhicules

Dans une problématique de segmentation du tarif et d'innovation, la classification de véhicule peut s'avérer un atout dans un modèle de tarification automobile. La théorie de la classification géographique a largement été reprise pour être transposée dans le cas des véhicules.

L'hypothèse principale des zoniers est que les résidus des modèles linéaires généralisés (autrement dit la part non expliquée du modèle) contiennent, outre du bruit, une information liée à des données géographiques, appelée variation résiduelle géographique. Dans le cas du véhiculier, il serait alors possible de récupérer une variation résiduelle dépendant des variables véhicules.

Deux types de classifications sont alors envisageables : une effectuée sur la variation résiduelle lissée dans le but de l'introduire directement dans le modèle GLM de la même façon que les autres variables ; une autre effectuée sur les prédictions du modèle combinées à la variation résiduelle, permettant ainsi de simplifier le modèle initial et de s'affranchir des variables véhicules sensées être reflétées dans le véhiculier. Le modèle incluant un véhiculier sur les prédictions combinées à la variation résiduelle sera donc plus simple en termes de nombre de variables que le modèle incluant un véhiculier sur la variation résiduelle seulement.

La grande problématique de la classification de véhicules est d'obtenir une carte, i.e. une représentation spatiale des véhicules et de définir une relation de voisinage entre chacun d'entre eux. Pour y répondre, nous avons utilisé un outil d'analyse de données : l'Analyse Factorielle de Données Mixtes. A mi-chemin entre Analyse en Composante Principale et Analyse par Correspondance Multiple, c'est une méthode factorielle de réduction de dimension applicable sur un jeu de données contenant à la fois des données qualitatives et des données quantitatives. Appliquée à la base véhicules, le résultat est un nuage de point en trois dimensions cohérent, chaque point représentant un véhicule.

Une fois le nuage de véhicules constitué, il est nécessaire de définir la relation de voisinage entre les points. Une première approche a été de considérer que tous les points situés dans un cube de longueur définie étaient considérés comme voisins. Elle n'a cependant pas été retenue dans la mesure où la longueur du cube, déterminante pour créer les liaisons, n'a pas été implémentée en vue d'un calibrage automatique. L'approche retenue ne doit donc pas comprendre de paramètres humains (du moins dans cette étape) dans un but de simplification, d'automatisation et de déploiement de la classification de véhicules sur d'autres garanties et d'autres portefeuilles. La triangulation de Delaunay permet de satisfaire cette condition. C'est une méthode commune de discrétisation

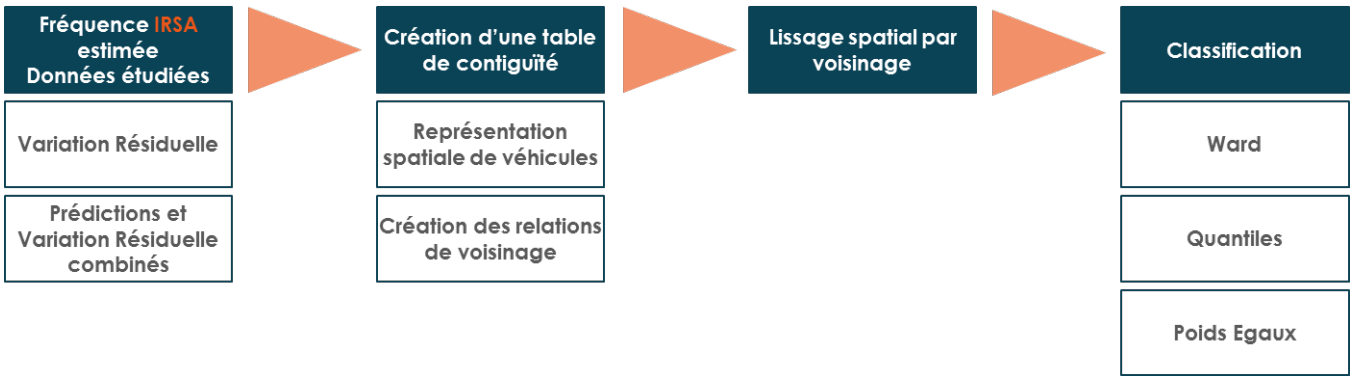


FIGURE 2 – Processus de création de véhiculiers

spatiale d'un milieu continu, ou maillage, dont de nombreux algorithmes ont déjà été codés et sont disponibles facilement sur internet pour différents environnements informatiques. Quelques retraitements automatiques sont alors effectués pour obtenir une vraie cohérence des voisinages.

Dès que la table de contiguïté est créée (à l'aide de la représentation spatiale des véhicules et leurs relations de voisinage), il est possible d'appliquer la méthodologie des zoniers, c'est-à-dire le lissage spatial de la variation résiduelle, puis la classification. Trois types de classifications sont étudiés : les méthodes de Ward, des quantiles et des poids égaux. Nous obtenons alors six véhiculiers : trois sur la variation résiduelle et trois sur les prédictions et la variation résiduelle combinées.

La classification a été créée sur le modèle de fréquence IRSA dans la mesure où cette typologie de sinistre représente les trois quarts des sinistres Responsabilité Civile Matérielle. On peut s'attendre alors à une segmentation plus importante du risque.

Résultats des véhiculiers

La première approche d'introduction des véhiculiers dans les GLM a été la méthode traditionnelle de test et d'ajout de variable dans un modèle GLM. Pour les véhiculiers créés sur la variation résiduelle, le véhiculier est directement introduit dans le modèle sur l'échantillon de modélisation, puis les critères de qualité du modèle ont été vérifiés sur l'échantillon de validation. Cette méthode s'est révélée inadéquate et présentait un fort surapprentissage sur l'échantillon de modélisation. Ceci s'explique du fait que les classifications ont été calibrées sur les trois quarts de cet échantillon.

Une deuxième approche a été utilisée pour éviter ce phénomène. Le calibrage des coefficients du véhiculier se fait sur l'échantillon de détermination du lissage, qui n'a pas été impliqué dans la création des véhiculiers. Ces coefficients ont été ensuite forcés sur l'échantillon de modélisation tandis que les coefficients de toutes les autres variables ont été recalibrés. Les résultats sur l'échantillon de validation sont meilleurs et il n'y a pas de sur-apprentissage.

Les mêmes méthodologies ont été appliquées aux les véhiculiers combinant prédictions et variation résiduelle. Les variables véhicules sont retirées du modèle dans la mesure où leur signal est déjà présent dans le véhiculier. Puis la classification est ajoutée au modèle. Le phénomène de sur-apprentissage a également été constaté en suivant cette méthode. Les coefficients du véhiculier ont donc été calibrés sur l'échantillon de détermination du lissage, puis fixés dans sur l'échantillon de modélisation tout en laissant les coefficients des autres variables se recalibrer. Les résultats sont également meilleurs qu'avec la première approche.

Le choix de la classification optimale se fait à l'aide des indicateurs statistiques usuels des GLM : l'AIC, le BIC et la Déviance. Comme précisé ci-dessus, la première approche entrainait un sur-apprentissage. La seconde méthode a alors été retenue. Les véhiculiers sont alors démarqués en fonction de leur coefficient de GINI. Ce dernier permet de sélectionner la classification de WARD. Enfin, les résultats entre le véhiculier de Ward sur la variation résiduelle seulement obtient de meilleurs résultats en termes d'Expected Deviance Ratio (EDR), même si les résultats du véhiculier de Ward sur les prédictions et la variation résiduelle combinées sont très proches.

Points d'amélioration

La segmentation des périls peut s'avérer intéressante pour d'autres garanties, comme en vol, où les risques de vol total et de vol partiel sont habituellement groupés tout en ayant des caractéristiques différentes. La méthodologie de classification de véhicule peut également être reproduite sur diverses garanties. A ce jour, les résultats sur la garantie vol n'ont montré qu'une amélioration minime.

La principale limite de la classification des véhicules est la maintenance. En effet, il faut mettre à jour le véhiculier de façon régulière. Par ailleurs, il est nécessaire de développer un algorithme permettant de rattacher de nouveaux véhicules commercialisés depuis l'étude à une classe donnée.

Il existe aujourd'hui de nombreuses méthodes de machine learning permettant de créer des classifications de façon plus automatisées, comme les *random forests*. Cependant, les résultats de ces classifications sont souvent difficiles à interpréter. Il pourrait être intéressant de comparer ces nouvelles méthodes au processus développé dans ce mémoire.

Synthesis Note

The following study carries on updating Third Party Liability - Material Damages in car insurance, for the retail line. The objective is to sophisticate the modelling through peril segmentation and innovate creating a vehicle classification using data mining tools and spatial smoothing on pure premium GLM residuals.

Context and objectives

Car insurance is a very competitive market due to various commercial strategies, new competitors or banks that provide insurance products to their own portfolio. The direct market is even more competitive as it changes much faster than the traditional market. The insurer has to innovate in order to keep its clients, especially concerning risk selection through tariff segmentation. A pricing adapted to each policy avoids anti-selection. Models have to be updated regularly and the risk estimation has be be the finest possible. New segmentation variables also have to be tested.

The objective of this study is the update to Third Party Liability - Material Damages cover with recent data, to innovate through new pricing structure and to create a vehicle classification.

Peril segmentation

Pure premium modelling is usually done using Generalized Linear Models. A frequency model is build on the total number of claims of this cover and the severity is a constant directly calculated through the Incurred charge and the number of claims. An adjustment coefficient will then take into account market effects or inflation. This type of structure is due to the IRSA convention.

Claims can be handled in two way depending on their type and on the third person insurer. Some are regulated by the IRSA convention in order to speed up the compensation of the victim : the non responsible insurer pays its client based on the amount of the damage invoice and can issue a recourse to the insurer of the responsible driver. This recourse is a fixed amount reevaluated every year. This type of claims represents 75% of the total amount of claims. When the convention is not relevant, i.e. for property damage claims, international claims or when the third party insurer is not part of the IRSA convention, the claim will be paid with the real cost. The average cost distribution is then polluted fixed amounts, on which it is not possible to apply a traditional GLM.

In order to provide a segmentation of the average cost, claims will be split up between two types (IRSA and non IRSA) : it is the segmentation of perils. A propensity model could have been studied but a frequency study shows that the profils of IRSA claims generators are not the same as non IRSA claim ones.,It is hence possible to split the pure premium in two frequency models following a Poisson law, a severity model following a Gamma law, and a constant for the IRSA average cost.

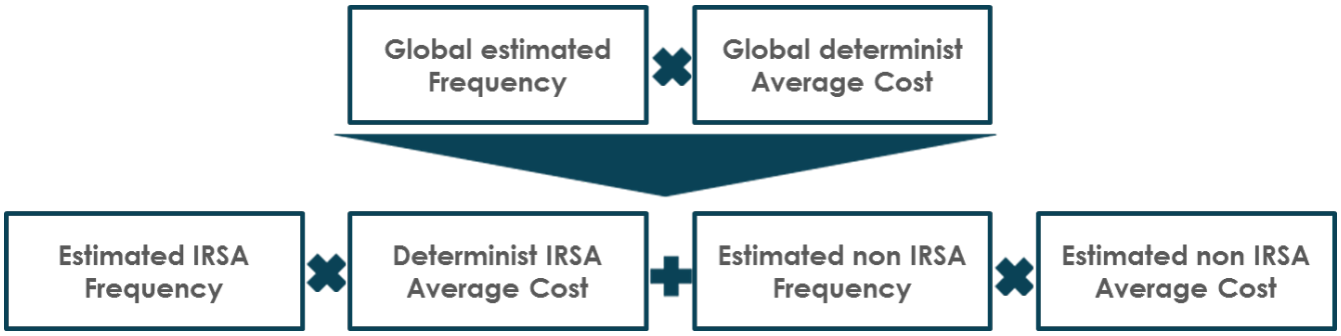


FIGURE 3 – Peril segmentation structure

Database creation

The database creation needs a prior study of the perimeter that has to be taken into account. The former frequency model was calibrated on data from 2012 to 2014. The new models will be calibrated on more recent data taht already take into account market changes or portfolio mix changes. IRSA claims are very short tail risk and the perimeter will be 2014-2016, i.e. three years in order to check time consistency. It takes more time for non IRSA claims to stabilise, mostly because the insurer needs to wait and receive official report. Hence the perimeter will be 2013 to 2016 for the frequency and severity models.

The database is presented as a number of images, resulting of contract, person, claims, geographical and vehicle databases merge. Each line is considered as a specific covered risk with an exposition. The vehicle database has been retreated with care as it is part of the main objectives of this study.

The database is then randomly sampled in a modelling set (80%) and a validation set (20%) used to measure the quality of the models. A quarter of the modelling set will also be used to created the vehicle classification and will be called smoothing set.

Peril segmentation results

The three models are created with GLM and variables are added one after another respecting four tests : relativity importance, time consistency test, statistical tests and judgement.

The pure premium IRSA model contains 24 variables whereas the non IRSA pure premium model contrains 32 variables. This has to be compared to the former model which had only 26 variables. The comparison between the two final pure premium structures is made according to the GINI coefficient. The new modelling with the peril segmentation has better results than the former models, with an increase of 49% of the GINI coefficient on the validation set and of 36% on the validation set.

The update of the models with more recent data and the new modelling structure objectives are fulfilled. It is then possible to concentrate on the innovation part with the creation of a vehicle classification.

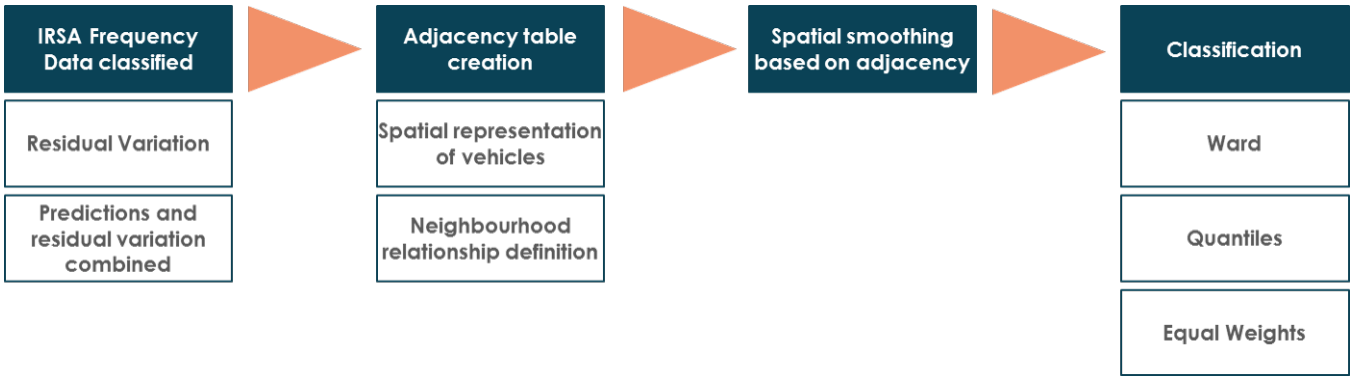


FIGURE 4 – Vehicle classification process

Vehicle classification creation

Vehicle classification can be an advantage in a pricing segmentation and innovation context in car insurance. It mainly relies on geographical classification theory, applied on vehicles.

The main hypothesis is that GLM residuals contains, other than noise, a signal that can be explained by vehicle variables, called residual variation.

Two types of classification are possible : one made on the smoothed residual variation to be introduced directly in GLM models like the other variables. The other one made on the vehicle variables predictions combined with the residual variation. The latter can be introduced in the model while removing all the other vehicle variables. The model is then much simpler than the former as the number of variables can decrease considerably.

The main problem with vehicle classification is that there is no spatial representation of vehicles and no neighbourhood relation between them. It is possible to do so with multivariate analysis with mixed quantitative, which does a factorial analysis for quantitative and qualitative variables at the same time. Applied to the vehicle database, the result is a cloud of vehicles in three dimension, each point representing a vehicle.

Once the cloud of vehicle is created, it is necessary to define the neighbourhood relationship between each point. The first approach was to consider as neighbours all the points within a cube of a give length. This approach was not kept as the length of the cube has to be determined manually, and this step have not been implemented to become automatic. The method had to avoid any human intervention in order to be as simple and automatic as possible, and be deployed on other covers or portfolios. Delaunay Triangulation meets these criteria. It is a method of spatial discretization and multiple algorithms have already been coded in R and are available on the internet. Some fine tuning is then made to have a coherence in neighbourhoods.

Once the adjacency table is created, it is possible to apply the geographical classification, i.e. the spatial smoothing on the residual variation, then the classification. Three types of classification will be studied : Ward method, Equal Weights methods and Quantiles method. There will be 6 vehicle classification : three on the residual variation and three on the predictions combined with the residual variation.

The classification will be made on the IRSA frequency model as these types of claims represent 75% of TPL Material claims. We can predict a segmentation of risk more important than on other covers.

Vehicle classification results

The first approach of the vehicle classification inclusion in GLM models is the traditional method : for classifications created on residual variation only, they are directly included in the model on the modelling set and quality criteria are checked on the validation set. This method ended up inappropriate as a strong overfitting appeared on the modelling set. This may be due to the classification creation method : is has been calibrated on three quarters of the modelling set.

A second approach was thought to avoid this phenomenon. The vehicle classification coefficients are calibrated on the smoothing set which was not included in the classification process (except to set the smoothing level). These coefficients were then offsetted in the modelling set while all other variables were re-calibrated. Results on the validation are much better and there is no overfitting.

The same approaches were taken with then classification made on the prediction combined to the residual variation. The vehicle variables are removed from the model as their signal is already taken into account in the vehicle classification. Then the classification is added to the model. The overfitting on the modelling set could also be observed. The vehicle classification coefficients were then set on the smoothing set, then offsetted on the modelling set while letting the other variables re-calibrate. The results were also better than with the first approach.

Usual statistic indicators were used to choose the best vehicle classification : AIC, BIC and Deviance. As described above, the first approach of variable introduction in the model lead to an overfitting. The second approach was kept. The classifications were evaluated according to their GINI coefficient. The classification with the best GINI coefficient was the one using the Ward method. The results of the Ward classification on the residual variation were slightly better than the Ward classification on the predictions combined to the residual variation in terms of Expected Deviance Ratio. The best classification retained was then the Ward classification on the residuals variation only.

Improvements

The peril segmentation is a very interesting change in a cover modelling structure and can be spread to theft cover, in which total or partial theft represent different risks but are usually modelled together. The vehicle classification can also be applied to other covers, although it only showed a minimal improvement on the theft cover.

One of the main drawbacks of the vehicle classification is the maintenance. The classification has to be updated quite regularly, and an algorithm has to be developed in order to price new vehicles commercialised after this study and not taken into account in the classification.

The best alternative to this method could be machine learning algorithms such as random forest. The process is much quicker and automatised, although the results are quite hard to understand. It could be interesting to challenge this study methodology with these machine learning methods.

Remerciements

Je tiens en premier lieu à remercier l'équipe de la Direction Technique de Direct Assurance pour son accueil tout au long mes 4 années passées à travailler à leur coté. Je remercie chaleureusement **Catherine Khounlivong**, mon ancienne responsable et tutrice d'entreprise, auprès de qui j'ai beaucoup appris. Merci pour sa disponibilité, son temps, ainsi que toute son expertise et son expérience. Je remercie **Isabelle Antoine**, Directrice Technique, pour m'avoir fait confiance et permis d'évoluer au sein de sa direction.

J'adresse mes sincères remerciements à **Denis Clot**, mon tuteur pédagogique, pour son encadrement et la pertinence de ses conseils au cours de mon travail de mémoire. Je remercie également **Anne Eyraud Loisel** de me permettre de soumettre ce mémoire plusieurs années après la fin de ma formation, ainsi que l'ensemble du corps éducatif de l'ISFA pour m'avoir accueilli tout au long de ma formation et pour leurs connaissances transmises.

Je souhaite exprimer ma reconnaissance à **Salma Said**, **Claire Mouminoux**, **Antoine Guillot** et **Florence Chatelle** pour leur temps, leur écoute, leur soutien, leurs conseils avisés et leur relecture.

Sommaire

Introduction

L'assurance automobile représente une grosse partie du marché de l'assurance des biens et des dommages, majoritairement dû au caractère obligatoire de certaines garanties telles la Responsabilité Civile. Le chiffre d'affaire généré par cette branche est conséquent, ce qui pousse les professionnels de l'assurance à proposer des produits d'assurance automobiles. En plus d'un cadre réglementaire très strict, les assureurs automobiles font face à une concurrence très rude à cause de stratégies commerciales développées par certains, à l'émergence de nouveaux acteurs ou aux bancassureurs qui bénéficient déjà d'un portefeuille de clients. Le marché du direct est d'autant plus soumis à cette concurrence qu'il mue de façon plus rapide que le marché traditionnel.

L'assureur doit alors innover pour conserver ses parts de marché, notamment lorsqu'il s'agit de sélection des risques à travers sa segmentation tarifaire. Une tarification ajustée permet de s'affranchir du phénomène d'anti-sélection. Ceci implique une revue récurrente des modèles, ainsi que le test de nouvelles structures tarifaires ou de variables de segmentation, comme l'âge par trimestres.

Une grande partie des dernières variables tarifaires testées ces dernières années étaient de type géographique : intégration de variables externes directement dans les modèles, création de classification géographiques dîtes « zoniers ». Le but de ces études était de capter un signal géographique, non saisi par les variables intégrées dans le modèle, dans les résidus de ce dernier. On peut appliquer ce raisonnement aux caractéristiques véhicule. Est-il possible de récupérer un signal dans les résidus expliqué par des variables véhicules ?

Le premier chapitre expose le cadre assurantiel dans lequel s'inscrit l'étude. Le fonctionnement de l'assurance automobile y sera présenté dans ses aspects généraux et réglementaires, ainsi que le produit, la garantie et le portefeuille sur lesquels l'étude est conduite. Cette étape permettra de mettre en évidence les trois objectifs : une revue des modèles de tarification, une segmentation des périls, et la création d'une classification de véhicules. La modélisation du risque en assurance ainsi que la création de la base de données seront également évoquées. Ceci permettra au lecteur d'avoir une vue d'ensemble de l'environnement de l'étude et de connaître le but de l'étude.

Le deuxième chapitre fera un état de l'art de la classification de véhicules en présentant la classification actuelle. Une nouvelle méthodologie, inspirée de la théorie des zoniers, sera alors décrite, tant au niveau théorique que l'enchaînement des différentes étapes permettant de calculer cette nouvelle variable. Les différentes sections couvertes iront de la représentation spatiale des véhicules à l'aide d'analyse de données, à des techniques de lissage spatial et de méthodes de classification. Le lecteur aura ainsi connaissance de tous les outils utilisés dans la création du véhiculier.

Le troisième et dernier chapitre se concentre sur l'application des résultats dans le but de répondre aux trois problématiques de l'étude. Ainsi l'analyse préparatoire du périmètre de l'étude,

les résultats sur la revue des modèles et la segmentation des périls seront exposés, suivis par l'application de la nouvelle méthodologie de classification permettant de créer de nouvelles variables de segmentation. Ces classifications seront alors intégrées aux modèles, qui seront comparés aux modèles initiaux afin de mesurer l'apport du travail effectué.

Première partie

Cadre de l'étude

Chapitre 1

Contexte et objectifs

Dans ce chapitre, nous allons tout d'abord étudier le contexte de l'étude : l'assurance automobile, ses grands chiffres, ainsi que Direct Assurance et son produit. Nous nous focaliserons ensuite sur la garantie responsabilité civile matérielle, garantie choisie pour l'étude grâce à son volume de données important, son fonctionnement et l'existence de la convention IRSA. Ensuite il conviendra de réaliser une étude de portefeuille afin de comprendre le profil des assurés de Direct Assurance. Ces sections permettront de définir et de justifier les objectifs de cette étude.

1 L'assurance automobile et l'assurance directe

Généralités

L'assurance de biens et de responsabilité est une branche de l'assurance représentant environ 26% des 209 milliards d'euros de cotisations d'assurance en 2016. Le tiers de ces cotisations concerne les produits d'assurances pour les particuliers, et 56% de ce tiers concerne l'assurance automobile (hors flotte), soit 18.7 milliards d'euros, ce qui fait de l'automobile la principale activité des sociétés d'assurance de dommage. Ce marché n'est pas prêt de s'essouffler dans la mesure où l'assurance responsabilité civile automobile est obligatoire, et où le parc automobile français ne cesse d'augmenter. Le taux d'équipement des ménages est supérieur à 80% en France.

L'assurance automobile a de multiples modes de distribution : le réseau traditionnel (mutuelle, agent, courtier, les bancassureurs) ou direct. Ce dernier peut se faire directement sur internet, via le site d'un assureur comme Direct Assurance, ou via les comparateurs en assurance, comme LeLynx.com ou LesFurets.com. Certains assureurs proposent de simuler un tarif et de commencer la souscription sur internet, puis de la terminer en agence. Le mode de distribution du direct ne représentait que 2.6% de l'ensemble de l'assurance automobile en 2016. Ce type de distribution est néanmoins en croissance, mais n'est pas comparable à celui du Royaume-Uni, où les ventes directes représentent plus de 70% du marché.

Le ratio combiné du marché de l'assurance automobile, correspondant à la somme des sinistres et des frais engagés (frais d'acquisition, de structure, de réassurance, etc.) sur la somme des primes acquises, est supérieur à 100% depuis plusieurs années (source FFA). Ceci veut dire que pour 100€ de prime récoltée, l'assureur dépense plus que ce qu'il a obtenu (en moyenne sur le marché). Il faut donc faire particulièrement attention à tarifer correctement les profils souscrits afin de ne pas être déficitaire.

Au vu de la concurrence grandissante et des problématiques de profitabilité, l'assurance directe présente beaucoup d'atouts et d'avenir.

Direct Assurance

Direct Assurance était jusqu'à fin 2017 une entité de la Business Unit Axa Global Direct qui regroupe les sociétés d'assurance directe du Groupe Axa en Belgique, Italie, Espagne, Pologne, Royaume-Uni, Corée du Sud, Japon et Chine. Elle a en 2016 plus de 1400 employés et affiche un chiffre d'affaire de 450 millions d'euros.

La société propose des produits d'assurance automobile, multirisque habitation et emprunteur sous la marque Direct Assurance, ainsi qu'une assurance automobile connectée sous la marque YouDrive. Elle se positionne en tant que leader de l'assurance du direct en France avec plus de 1.000.000 clients.

Le produit automobile de Direct Assurance se décline en 4 formules : Tiers Mini, Tiers Essentielle, Tiers Maxi et Tous Risques. Le détail des garanties incluses dans chaque formule est disponible dans la table 1.1. Les garanties, définies telles que sur le site de Direct Assurance et dans les conditions générales, sont les suivantes :

Dommages causés à autrui, ou plus communément appelée Responsabilité Civile. Il s'agit d'une couverture pour les dommages matériels et corporels causés par le véhicule assuré à une personne tierce. Le fonctionnement de cette garantie sera décrit dans la section suivante.

Défense pénale et recours. Les intérêts de l'assuré sont défendus lorsque ce dernier est poursuivi devant les juridictions répressives (Tribunal de police) à la suite d'un accident impliquant le véhicule assuré. Direct Assurance intervient également auprès de la compagnie d'assurance automobile du responsable identifié d'un accident afin d'obtenir la réparation des dommages subis par le véhicule assuré et ses occupants.

Protection juridique automobile. Les intérêts de l'assuré sont défendus par des juristes dans le cas par exemple de poursuites judiciaires ou administratives hors les accidents, à la suite de l'utilisation du véhicule (infraction du code de la route).

Assistance, consiste en une aide d'urgence immédiate en cas de panne, accident, vol, maladie au cours d'un voyage avec le véhicule assuré. Il comprend le remorquage du véhicule avec une franchise kilométrique de 50km par rapport au domicile de l'assuré, et la prise en charge du conducteur et de ses passagers pour la poursuite du voyage ou du rapatriement.

La **garantie personnelle du conducteur**, permettant d'assurer le conducteur du véhicule dès lors que le taux d'atteinte permanente à l'intégrité physique et psychique est supérieure à 10%, lorsqu'il est victime et responsable d'un accident. La limite de l'indemnisation est de 400.000€.

Tempêtes, évènements climatiques exceptionnels, attentats, catastrophes naturelles et technologiques. Une catastrophe naturelle doit forcément faire l'objet d'un arrêté dans le Journal Officiel, tandis que la tempête n'a pas besoin d'être reconnue. La tempête pourra concerner les épisodes de grêle par exemple.

Bris de glace. Remplacement ou réparation du pare-brise, lunette arrière ou vitres latérales, soumis à une franchise pour les remplacements.

Incendie et vol. Concernant l'incendie, il s'agit de la couverture des dégâts causés par le feu, hors les actes de vandalismes qui seront alors couverts par la garantie Dommage Tous Accidents.

Garantie	Tiers Mini	Tiers Essentiel	Tiers Maxi	Tous Risques
Dommages causés à autrui (Responsabilité Civile)	x	x	x	x
Défense Pénale et Recours	x	x	x	x
Protection Juridique Automobile	x	x	x	x
Assistance 24h/24, 7j/7	x	x	x	x
Garantie Personnelle du Conducteur à 400.000€	x	x	x	x
Tempête et évènements climatiques exceptionnels		x	x	x
Attentats, Catastrophes naturelles et technologiques		x	x	x
Bris de glace		x	x	x
Incendie (hors vandalisme)			x	x
Vol			x	x
Dommages Tous Accidents (y compris vandalisme)				x

TABLE 1.1 – Garanties proposées par Direct Assurance par Formule

Les deux garanties sont soumises à une franchise variable en fonction du prix du véhicule et du montant des réparations.

Dommage tous accidents, concerne l'indemnisation de tous les dommages causés au véhicule assuré en cas de sinistre, qu'il soit issu d'une chute d'objet, d'un acte de vandalisme, d'un accident responsable avec un autre véhicule, d'un accident où seul le véhicule assuré est impliqué tel un accident de parking. Cette garantie est également soumise à une franchise variable en fonction du prix du véhicule et du montant des réparations.

Direct Assurance propose au delà de ces formules 3 packs. Le Pack Protection permet d'étendre la couverture de la garantie personnelle du conducteur de 400.000€ à 800.000€. Le Pack Tranquillité étend cette limite à 1.500.000€ et inclut le prêt de véhicule durant la période des réparations. Enfin, le Pack Sérénité comprend le Pack Tranquillité et supprime la franchise kilométrique de la garantie assistance initiale (de 50km).

La charge sinistre de Direct Assurance est d'environ 300 millions d'euros. La répartition par garantie n'est pas tout à fait la même que celle du marché, comme le montrent les figures 1.1 et 1.2. La part des sinistres de la garantie responsabilité civile est bien plus importante que sur le marché (65% chez Direct Assurance vs 50% sur le marché). Ceci peut s'expliquer par le mix du portefeuille, représentant en partie les profils ciblés par la stratégie de la compagnie. En effet, la proportion de jeunes dans le portefeuille est très importante, et ces profils sont générateurs d'une sur-sinistralité et ont tendance à ne prendre que la formule Tiers Mini.

La proportion de la charge en incendie et vol reste comparable, tandis que celle du bris de glace et la garantie dommage tous accidents est inférieure chez Direct Assurance. Ce vase communiquant, au profit de la responsabilité civile, peut également s'expliquer par le montant des franchises (certains assureurs choisissent de supprimer les franchises tout en augmentant leur prix pour compenser la perte due aux sinistres).

Une bonne tarification et segmentation de la garantie responsabilité civile est donc primordiale pour Direct Assurance. Il conviendra d'expliquer dans les sections suivantes le fonctionnement de la garantie ainsi qu'une étude de portefeuille, permettant de définir les objectifs de cette étude.

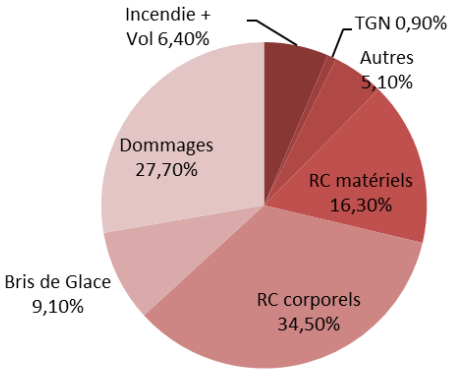


FIGURE 1.1 – Répartition de la charge des sinistres 2015 sur le marché

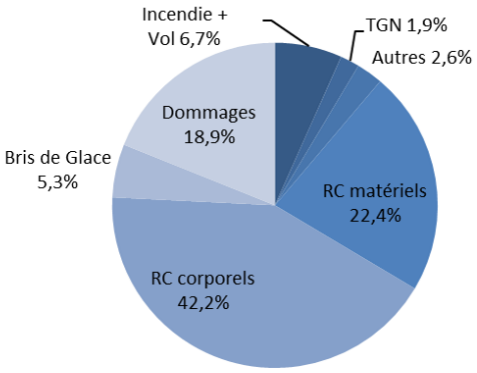


FIGURE 1.2 – Répartition de la charge des sinistres 2015 Direct Assurance

2 La garantie Responsabilité Civile Matérielle

Dans cette section nous allons définir la garantie responsabilité civile matérielle, son fonctionnement, ainsi que la convention Indemnisation directe de l'assuré et Recours entre Sociétés d'Assurance (IRSA), qui sera au centre de cette étude.

Assurer son véhicule est obligatoire. Au delà de l'obligation légale, les assureurs proposent des couvertures optionnelles à la souscription en fonction des besoins spécifiques de chaque assuré, comme évoqué dans la sous-section précédente. Mais seule la garantie responsabilité civile demeure obligatoire. Elle permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un de ses passagers :

— blessures ou décès d'un piéton, d'un cycliste, d'un passager, ou d'un occupant d'un autre véhicule
— dommages aux autres voitures, deux-roues, immeubles...

Cette garantie responsabilité civile couvre les conducteurs autorisés ou non autorisés. Cependant, après avoir indemnisé les victimes, l'assureur peut disposer d'un recours à l'encontre des conducteurs non autorisés.

Le non-respect de cette obligation d'assurance représente un délit, et la personne n'ayant pas assuré son véhicule peut recevoir une amende ainsi qu'une suspension de permis et une mise en fourrière de son véhicule.

Fonctionnement

Afin de bien comprendre le fonctionnement de la garantie responsabilité civile matérielle, ainsi que sa gestion au sein de Direct Assurance, il conviendra de considérer l'exemple suivant.

Monsieur X, assuré chez Direct Assurance, et Madame Y, assurée chez un concurrent, ont eu un accident. Le préjudice subi par Monsieur X s'élève à 1200€ et celui de Madame Y à 4300€. L'assureur concurrent ne fait partie de la convention IRSA, définie dans la prochaine sous-section. Autrement dit, la gestion du sinistre se fait dans le cadre de la loi Badinter.

Supposons que Monsieur X soit responsable à 100% du sinistre. Les règles de gestion seront les suivantes :

— Si Monsieur X est assuré en Tous Risques, son préjudice sera remboursé à hauteur de 1200€ au titre de la garantie DTA, moins la franchise
— Ayant connaissance d'un véhicule endommagé, Direct Assurance va affecter une provision sur la garantie RCM
— Le concurrent devra évaluer le montant du préjudice subi par son assurée et émettre un recours égal au total de la facture à Direct Assurance
— Direct Assurance pourra alors régler l'assureur du lésé du montant du recours
— Une fois le recours encaissé par l'assureur concurrent, ce dernier pourra rembourser son client.

Supposons maintenant que Monsieur X ne soit pas responsable du sinistre. Les règles de gestion seront les suivantes :

— Direct Assurance doit évaluer le montant du préjudice subi par son assuré et émettre un recours égal au total de la facture à la compagnie concurrente, sur la garantie Recours Matériel
— Si Monsieur X est assuré en Tous Risques, il pourra être indemnisé en avance au titre d'une avance sur recours. Le cas échéant, il devra attendre l'encaissement du recours par Direct Assurance avant d'être indemnisé
— Une fois le recours reçu, Direct Assurance peut indemniser son client à hauteur de 1200€

Ce processus peut prendre longtemps dans la mesure où chaque acteur doit attendre d'encaisser le recours adverse avant d'indemniser son assuré, outre politique spécifique de compagnies.

Convention IRSA

Depuis 1968, les assureurs français ont mis en place des systèmes conventionnels qui ont permis d'accélérer et d'améliorer l'indemnisation des victimes d'accidents de la circulation. Les premières relevaient du domaine matériel, avec notamment l'obligation pour l'assureur du lésé, dit assureur direct, d'indemniser les dommages matériels subis par son assuré dans la mesure de son droit à réparation, déterminé selon les règles du droit commun. Aujourd'hui connue sous le nom de Convention IRSA, pour Indemnisation directe de l'assuré et Recours entre Sociétés d'Assurance, elle n'est pas obligatoire, mais toutes les compagnies d'assurance peuvent y adhérer.

Le principe fondamental de la convention est l'indemnisation directe de l'assuré, défini de la façon suivante : *Quels que soient la typologie de l'accident de la circulation, la nature et le montant des dommages, les sociétés adhérentes s'obligent, préalablement à l'exercice de leurs recours, à indemniser elles-mêmes leurs assurés, dans la mesure de leur droit à réparation, déterminé selon les règles du droit commun.*

Son champ d'application est celui des *accidents de la circulation, y compris les opérations de chargement et de déchargement des véhicules, survenus, en France (métropolitaine et DOM) et dans la principauté de Monaco, impliquant au moins deux véhicules terrestres soumis à l'obligation d'assurance assurés auprès de sociétés adhérentes. Pour les accidents survenus hors « France et Monaco », la Convention s'applique si ne sont impliqués que des véhicules immatriculés dans ces territoires. En présence d'une fraude, les dispositions de la Convention ne sont pas applicables.*

Concrètement, lors d'un accident, l'assureur direct doit procéder à l'évaluation des dommages matériels de son assuré non responsable. Si le montant de la facture hors taxe dépasse un certain plafond, aujourd'hui égal à 6500€, l'indemnisation et le recours se feront au coût réel. Le cas échéant, l'assureur direct indemnisera son assuré du montant de la facture et exercera un recours forfaitaire envers l'assureur du responsable.

Pour plus de détails sur la convention, notamment sur le fonctionnement lors de sinistres avec des objets inertes ou lors de carambolages, i.e. des sinistres impliquant 3 véhicules ou plus, dont les règles sont bien plus complexes, le lecteur est invité à se référer au document officiel de la Convention IRSA [8].

Forfaits

Les forfaits des recours IRSA évoluent chaque année de 3% à 5%, prenant en compte l'inflation du coût des pièces de réparation et de main d'oeuvre. Le forfait était de 1236€ en 2012, puis est passé progressivement à 1242€ en 2013, 1276€ en 2014, 1308€ en 2015, 1354€ en 2016 et est arrivé en 2017 à un coût de 1420€.

Il faut noter que chacun de ses montants peut être réduit de moitié dans le cas de sinistres avec une responsabilité partagée à 50%. Ainsi pour un sinistre survenu en 2015 avec une responsabilité à 50%, le recours que pourra exercer l'assureur envers la compagnie concurrente ne sera que de 654€.

Ces montants représentent bien souvent le coût d'ouverture des sinistres 2 véhicules chez la plupart des assureurs.

3 Etude du portefeuille

Dans cette section, nous allons présenter quelques graphiques nous permettant de comprendre le profil des assurés de Direct Assurance et regarder les évolutions de mix, c'est à dire de proportion par modalité de variables, depuis les 6 dernières années. De plus, une différence sera faite entre le mix en affaire nouvelle et le mix portefeuille (i.e. les affaires nouvelles et les termes). En effet, une étude du mix portefeuille seul ne permettrait pas de rendre compte des profils cibles, les proportions ne rendant pas compte de la politique tarifaire de l'année, de la modification des règles de souscription, et embarquant l'impact des résiliations de façon trop importante.

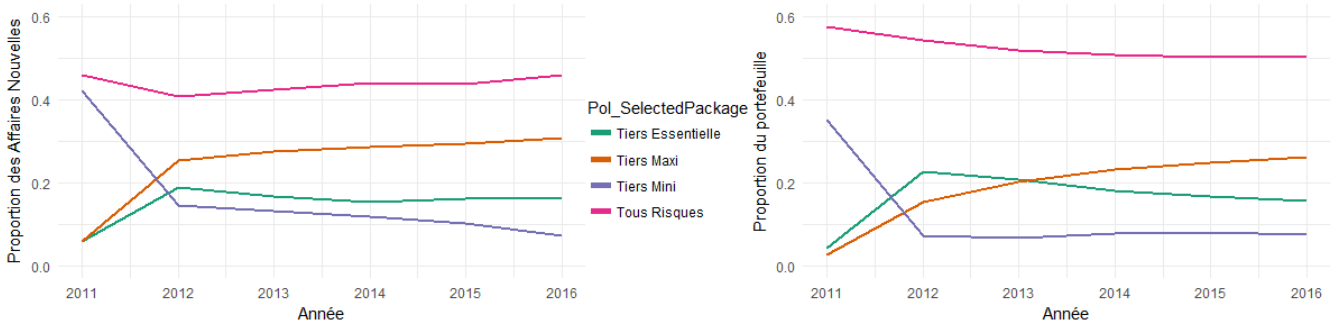


FIGURE 1.3 – Proportions des Formules dans le temps

L'évolution du mix en fonction de la Formule est présentée sur la figure 1.3. Les chiffres de l'année 2011 peuvent sembler étranges par rapport aux années suivantes au vu du faible taux de formules Tiers Essentielle et Tiers Maxi. Ceci s'explique par le lancement de la nouvelle offre auto courant 2011, donnant la possibilité au client de choisir entre les quatre formules actuelles. Avant ce lancement seuls la Tiers Mini et la Tous Risques, connues sous les noms de formule de Base

et Complète, étaient vendues. Ces formules ont migré au fur et à mesure sur les nouvelles. Nous pouvons voir que le taux de Tous Risques avoisine les 45% en affaires nouvelles, tandis que ce taux est proche de 70% en renouvellement (soit environ 50% en portefeuille). Ceci est un effet des résiliations, décrit dans le paragraphe précédent : les clients ayant choisi la formule Tous Risques ont moins tendance à résilier leur contrat que les autres. Le mix des formules en affaires nouvelles reste relativement stable (hormis une légère baisse de la formule Tiers Mini) comparé au mix en portefeuille, qui montre une baisse des Tous Risques et de la Tiers Essentielle au profit de la Tiers Maxi. Le taux d'inclusion par formule a un impact potentiellement important sur les modèles car le niveau de risque n'est pas le même pour chacune d'entre elles. Un assuré ayant choisi la Tous Risques n'aura pas le même comportement au volant qu'un assuré en Tiers Mini. Un profil Tiers Mini est bien plus risqué pour une garantie équivalente.

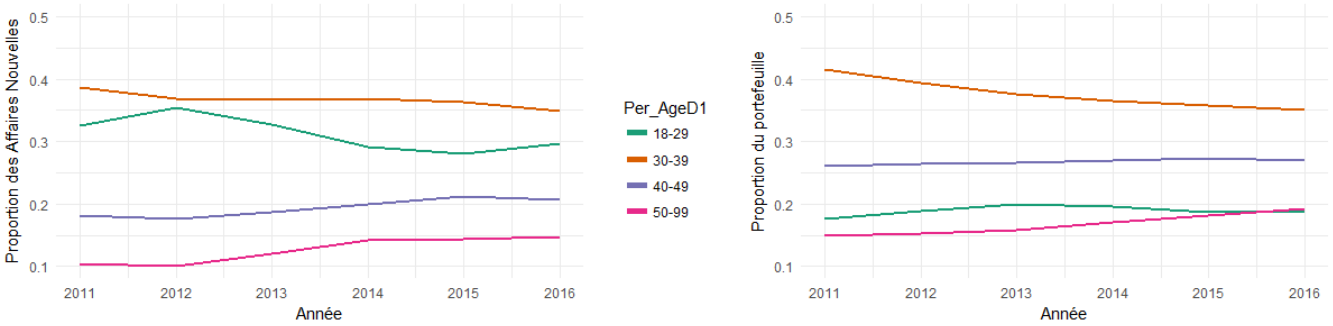


FIGURE 1.4 – Proportions de l'âge des assurés dans le temps

Le portefeuille de Direct Assurance est un portefeuille relativement jeune. En effet, les 18-39 ans représentent environ 70% des profils en affaire nouvelle et 55% en portefeuille (cf. figure 1.4) : ces profils ont plus tendance à résilier leur contrat que les clients plus âgés. Nous pouvons remarquer qu'il y a eu un changement de mix assez important depuis 2012 et la proportion des plus de 40 ans a augmenté en portefeuille, dû à un apport plus important de cette population en affaires nouvelles. En lien avec l'âge des conducteurs assurés, les clients sont plutôt urbains et actifs, et cela implique également une part de vieux véhicules assez importante (cf. figure A.1 en Annexe). Attention cependant, la proportion des véhicules anciens a également fortement augmenté (12 ans et plus) depuis 2013 alors que celle des jeunes n'a pas autant évolué : ceci prouve qu'il y a tout de même eu un changement de profil au sein du portefeuille. Il serait intéressant de faire une analyse de mix entre les variables âge du conducteur principal et âge du véhicule mais ceci n'est pas l'objectif de ce mémoire.

Nous pouvons également étudier le profil des véhicules en fonction du groupe SRA (cf. définition section 4.2), visibles sur la figure 1.5. Nous pouvons constater que la proportion des véhicules de groupe élevé (31 et plus) a fortement augmenté autant en portefeuille ainsi qu'en affaires nouvelles. Ceci découle d'une modification des règles de souscription, un assouplissement des critères pour lesquels un client est accepté ou refusé. Le volume de ces véhicules est ainsi plus important et permettra de créer des modèles plus robustes que ceux utilisant des données de 2011 et 2012. La même réflexion peut se faire sur la classe de prix SRA (cf. définition section 4.2). Par ailleurs, comme évoqué dans le paragraphe précédent, l'âge des véhicules a augmenté graduellement, passant de 8.42 ans en 2011 à 9.4 ans en 2017.

Enfin, il est intéressant de s'attarder sur la variable canal d'origine de l'assuré (téléphone, comparateurs ou site internet) disponible sur la figure 1.6. La majorité des clients souscrivent par téléphone

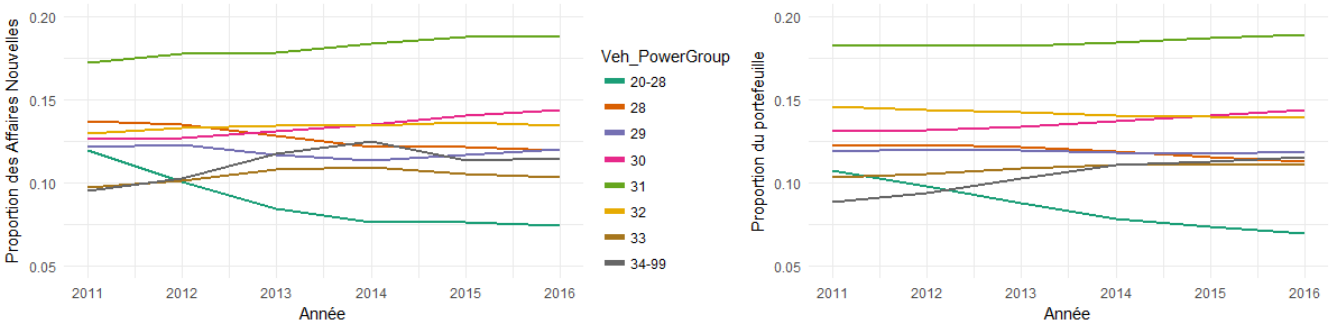


FIGURE 1.5 – Proportions des Groupes SRA dans le temps

(45% en 2016 en affaire nouvelle). Cette variable se répercute sur la sinistralité : de la même façon que la formule, les clients souscrivant au téléphone sont moins risqués que ceux ayant souscrit sur les agrégateurs d'assurance. Elle est d'autant plus importante qu'elle est directement impactée par des processus opérationnels ou des décisions stratégiques, comme l'ouverture d'un nouveau centre d'appel, l'ajout d'un agrégateur permettant la vente du produit ou le développement du site internet afin qu'il soit plus intuitif.

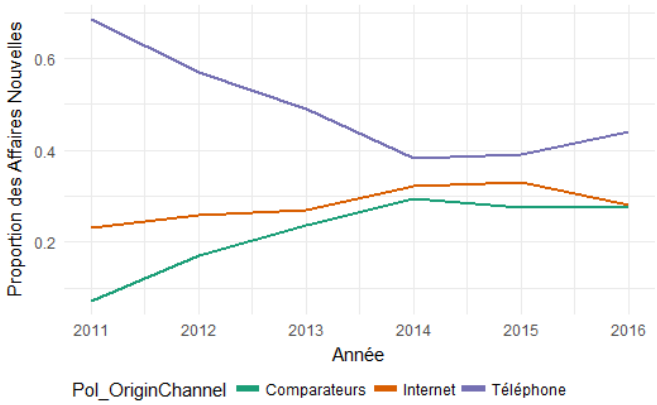


FIGURE 1.6 – Proportions du canal de distribution dans le temps

Le portefeuille de Direct Assurance décrit ci-dessus a donc subi beaucoup d'évolutions au cours des 6 dernières années, dues à des effets de marché ou à des décisions stratégiques tarifaires (modifications de tarif, modification des règles de souscription) ou opérationnelles. Ceci est une information déterminante permettant de clarifier l'objectif de cette étude.

4 Objectifs

Dans cette section, nous allons exposer les différents objectifs de ce mémoire. En effet, trois axes sont à développer.

Objectif 1 : L'objectif premier de ce mémoire est de faire une mise à jour ou une **revue des modèles linéaires généralisés**. Ces derniers n'ont pas été revus depuis juin 2015. Ils utilisaient

des données allant de 2011 à 2013 et une vision de la sinistralité au 30/06/2015 (l'état des lieux du modèle actuel est développé dans la section 2.3).

Or nous avons montré dans la section 1.3 qu'il y a eu un changement de mix important du portefeuille lors de ces dernières années avec inclusion de nouveaux profils d'assurés pouvant impacter le caractère prédictif des modèles. De plus, une multitude de décisions stratégiques ou opérationnelles ainsi que le lancement d'une nouvelle offre auto peuvent également rendre les modèles moins performants. Par ailleurs, les modèles sont généralement revus ou réajustés tous les ans, avec une refonte plus en détail tous les deux ans.

Objectif 2 : Ainsi, l'ensemble des modèles de chaque garantie nécessite une mise à jour. Nous avons vu dans la section 1.2 le fonctionnement de la garantie Responsabilité Civile Matérielle et l'existence de la convention IRSA permettant l'échange de forfaits entre les assureurs inscrits pour accélérer le processus de remboursement des assurés. Il existe donc deux types de sinistres pour une même garantie : les sinistres au coût réel et les sinistres au forfait.

La structure actuelle du modèle, exposée dans la section 2.3, ne propose qu'une modélisation de la fréquence globale à laquelle on ne multiplie qu'un coût moyen fixe au vu de la sur-représentation des forfaits et de facto l'impossibilité de faire une modélisation du coût moyen à part (cf. figure 2.2). Nous pouvons profiter de la refonte des modèles du premier objectif pour tenter de **segmenter les périls** et proposer une modélisation plus complète et adaptée de la garantie RCM.

Objectif 3 : Cette révision des modèles nous donne également la possibilité de tester de nouvelles variables afin d'améliorer la prédiction de la sinistralité. Ces dernières peuvent être le calcul d'un score d'orientation en garages partenaires dans lesquels le coût moyen des sinistres est généralement moins élevé, le calcul de l'âge d'obtention du permis de conduire (les variables habituellement intégrées dans les modèles sont l'âge du conducteur et l'ancienneté de permis de conduire) ou l'âge par trimestre (cf. CONDEMINE [2014] [7], ainsi que deux études de l'Institut Belge pour la Sécurité Routière, SLOOTMANS et *al.* [2011] [20] et DUPONT [2012] [10]).

Une grande partie de ces variables testées sont des variables géographiques externes : les assureurs souhaitent capter une sur-sinistralité en fonction du lieu d'habitation ou de travail de l'assuré, auparavant non captée dans les modèles. Les contraintes informatiques le forcent souvent à créer une classification géographique ou zonier [1].

Direct Assurance ayant déjà développé ce genre de méthodologie pour la garantie RCM, nous allons tester ces variables dans le modèle afin de vérifier si ces zoniers ont conservé leur caractère prédictif. Cependant nous souhaitons aller au delà d'un signal géographique non capté dans les modèles et rechercher un signal véhicule. En d'autre mots, nous souhaitons **appliquer la méthodologie des zoniers à des véhicules dans le but de créer un "véhiculier"**.

La classification géographique nécessitant une carte (de France par exemple), cette idée va introduire une grande problématique dans ce mémoire : comment créer une carte de véhicules ? Comme cette approche est innovante, nous souhaitons nous baser sur une garantie comportant suffisamment de données afin d'avoir des résultats robustes. C'est pour cela que la garantie RCM apparaît comme la garantie de prédilection : étant obligatoire, la totalité du portefeuille auto de Direct Assurance en est bénéficiaire. Cette méthode une fois construite pourra être appliquée à d'autres garanties.

1. Une multitude de documents et de mémoires sont disponibles à ce sujet (cf. bibliographie ou http ://www.ressources-actuarielles.net/)

Chapitre 2

Modélisation de la prime pure

Ce chapitre se consacre à la modélisation de la prime pure pour la garantie responsabilité civile matérielle, choisie dans le chapitre précédent. Nous allons dans un premier temps définir les indicateurs suivis par les assureurs. Nous exposerons ensuite les grands principes des modèles linéaires généralisés permettant de prédire une partie de ces indicateurs. Enfin, nous exposerons la situation actuelle de la modélisation de la garantie RCM, ainsi que les notions de contraintes et une proposition de segmentation des périls. Nous disposerons alors de tous les éléments théoriques, pratiques et méthodologiques pour modéliser la prime pure.

1 Indicateurs de sinistralité

L'assureur a à sa disposition plusieurs indicateurs lui permettant de suivre l'évolution de la sinistralité de son portefeuille. Ces indicateurs sont à la base de la tarification automobile ainsi que de la stratégie de l'entreprise.

Exposition
Cet indicateur correspond à la période durant laquelle un assuré est couvert par l'assurance, i.e. son exposition au risque. L'exposition d'une police étant restée une année entière vaudra 1. Elle est calculée en fonction de la date de début et de fin du contrat d'assurance à l'aide de l'expression suivante :

$$\text{Exposition} = \frac{\text{Date de fin} - \text{Date de début}}{365.25}$$

Fréquence
La fréquence correspond au nombre de sinistres qui se sont produits sur une certaine période.

$$\text{Fréquence} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

Elle est généralement utilisée pour différencier le risque entre plusieurs modalités d'une variable sur une même période. Exemple : la fréquence Responsabilité Civile Matérielle des conducteurs ayant obtenu leur permis avec le programme de conduite accompagnée (A) était de 4.3% au total de l'année 2015, tandis que celle des conducteurs n'ayant pas suivi le programme (B) était de 4.6%. On peut conclure que les polices (B) ont une tendance supérieure à générer des accidents que les polices (A).

Coût Moyen
Il s'agit du coût moyen des sinistres, obtenu par la formule :

$$\text{Coût moyen} = \frac{\text{Charge sinistre}}{\text{Nombre de sinistres}}$$

En conservant le même exemple que pour la fréquence, le coût moyen de la garantie Responsabilité Civile Matérielle des polices (A) était de 1480€ tandis que celle des polices (B) était de 1560€. Les sinistres des conducteurs (B) sont donc plus coûteux que ceux des conducteurs (A).

Prime Pure
La prime pure correspond à la charge moyenne générée par police sur une période donnée. C'est cet indicateur que l'assureur cherche à modéliser pour créer sa tarification.

$$\text{Prime pure} = \text{Fréquence} \quad \text{x} \quad \text{Coût moyen} = \frac{\text{Charge sinistre}}{\text{Exposition totale}}$$

Les trois indicateurs évoqués ci-dessus sont les plus importants dans la création d'un modèle de tarification. Les méthodes d'estimations seront développées dans la section 2.2.

Loss Ratio
Cet indicateur permet de calculer la rentabilité de l'assureur. Lorsque le Loss Ratio est inférieur ou égal à 1 pour un segment de portefeuille donné, alors ce segment est considéré comme rentable.

$$\text{Loss Ratio} = \frac{\text{Charge sinistres}}{\text{Primes acquises}}$$

Attention, il faut distinguer la prime acquise de la prime pure. En effet, si la prime pure permet de calculer le coût des sinistres par police sur une période donnée, la prime acquise correspond à la cotisation payée par l'assuré. Elle est issue de la Prime Pure, à laquelle s'ajoutent les frais (par exemple frais de gestion, salaires, loyer des bâtiments), les taxes dans le cas du TTC (différentes selon les garanties), les contraintes légales (prise en compte du facteur de bonus/malus, exclusion de la variable du genre masculin/féminin, remise effectuée aux personnes ayant fait de la conduite accompagnée), ainsi que de la stratégie commerciale, le tout multiplié par l'exposition.

Prime Moyenne Acquise
La prime moyenne correspond à la prime moyenne réglée par un assuré à son assureur sur une période donnée. Il ne s'agit pas d'un indicateur de sinistralité, mais comparé à la prime pure il permet de cibler les ajustements de stratégie commerciale sur certains segments.

$$\text{Prime Moyenne} = \frac{\text{Primes Acquises}}{\text{Exposition totale}}$$

L'ensemble de ces indicateurs permet à l'assureur de suivre l'évolution de son portefeuille et de sa sinistralité.

2 Modèles Linéaires Généralisés

L'introduction des modèles linéaires en assurance non-vie a permis de sophistiquer la tarification de produits tels que les assurances automobile. Cette méthode proposait une estimation d'une

variable réponse $Y = \mu + \epsilon$, avec μ sa moyenne et ϵ une variable aléatoire, avec pour hypothèse la possibilité d'écrire μ comme combinaison linéaire de variables explicatives X_i, i.e. il existe β_0, ..., β_n paramètres à estimer tels que $\mu = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$; et la normalité de ϵ centrée et de variance σ^2 constante. Les variables X_i sont des variables dont dispose l'assureur, telles l'âge du conducteur ou la marque de son véhicule. Cette méthode représente cependant un fort inconvénient. En effet, la condition de normalité sur ϵ est difficile à satisfaire, et la variable à expliquer Y ne suit pas toujours une loi normale.

Définition

Les Modèles Linéaires Généralisés (GLM) sont une extension des modèles linéaires. Il s'agit de la technique de modélisation la plus répandue en assurance non-vie. Cette méthode permet à la variable réponse Y de suivre une autre loi que la loi normale, et de s'affranchir du critère de variance constante. Elle est caractérisée par trois composantes :

— Une composante aléatoire Y
— Une composante déterministe η
— Une fonction lien

La composante aléatoire est la variable à expliquer (ou variable réponse) Y, dont la loi appartient à la loi exponentielle, i.e. dont la densité peut s'exprimer sous la forme

$$f(y) = \exp[\frac{y\theta - v(\theta)}{u(\phi)} + w(y, \phi)]$$

Où θ est appelé paramètre naturel de la famille exponentielle, ϕ le paramètre de dispersion, u une fonction réelle non nulle, v une fonction réelle deux fois dérivable et w une fonction définie sur $\mathbb{R}^2$.

On peut alors montrer que sous ces conditions l'espérance et la variance de Y peuvent s'écrire sous la forme : $\mu = \mathbb{E}[Y] = v'(\theta)$ et $\mathbb{V}[Y] = v''(\theta)\phi$

Il est d'usage d'utiliser la loi de Poisson pour modéliser la fréquence. La sévérité, i.e. le coût moyen des sinistres est généralement modélisé par une loi Gamma. Le modèle binomial négatif est également utilisé mais il représente bien souvent un cas d'école en assurance automobile. Le lecteur intéressé est invité à se reporter à l'ouvrage VASECHKO et *al.* [2009] [22].

La composante déterministe η est également appelée prédicteur linéaire. Soient $X = (X_1, \dots X_n)$ les variables explicatives du modèle. On le définit par $\eta(X) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$.

La troisième composante d'un modèle linéaire généralisé est la fonction lien, décrivant la relation entre la composante aléatoire Y et la composante déterministe, le prédicteur linéaire η. Notons $\mu = \mathbb{E}[Y]$ l'espérance de la variable à expliquer et g la fonction lien supposée monotone et différentiable. La fonction lien qui associe la moyenne μ au paramètre naturel θ de la famille exponentielle est appelé fonction lien canonique.

Le choix de la fonction lien va dépendre de la structure du modèle, additive ou multiplicative. Lorsque l'assureur choisi de modéliser la prime pure via deux modèles, de fréquence et de coût moyen, il est préférable d'opter pour un modèle multiplicatif. En effet, Prime Pure = Fréquence $*$ Coût Moyen. C'est le cas lorsque la fonction lien est la fonction logarithme.

Critère de sélection de variables

L'une des étapes clé dans la création ou la revue d'un modèle linéaire généralisé est la sélection des variables à introduire dans le modèle. ANDERSON et *al.* [2007] [1] proposent un certain nombre

de critères pratiques à respecter lors de la sélection de variables.

Relativité et écarts types : Une variable ordinale devra présenter une tendance avec des relativités permettant de segmenter suffisamment ses modalités. Si la variable est qualitative, seul le critère de relativité importe. Il faut de plus que les intervalles de confiance ne soient pas trop larges.

Test de robustesse : La tendance et les relativités ci-dessus doivent être robustes, c'est-à-dire que les relativités issues de l'interaction de la variable avec le temps (l'année de survenance) doivent conserver une certaine cohérence. Si une année particulière présente un caractère atypique dans sa sinistralité, par exemple une sur-fréquence dans une région donnée due à des conditions climatiques très défavorables), alors les paramètres estimés auront tendance à pénaliser cette région par rapport à d'autres à cause d'un phénomène ponctuel, et ne seront pas adaptés pour une estimation future du risque.

Test statistique : Le pouvoir prédictif d'une variable et la qualité d'ajustement d'un modèle sont mesurés à l'aide de tests tels le test du Khi2. Il s'agit alors d'étudier le compromis entre la précision du modèle incluant la variable et la complexité introduite due à un nombre plus élevé de paramètres à estimer. En définissant l'hypothèse nulle comme l'hypothèse selon laquelle les modèles sont globalement équivalents, on calcule la statistique du Khi2. L'hypothèse nulle sera rejetée si la statistique est inférieure à 5%, i.e. les modèles ont le même degré de complexité. Il conviendra alors de conserver la variable à l'étude et utiliser le modèle avec le plus de paramètres.

Jugement : Outre les trois tests précédents, il est important lors de la sélection des variables d'y ajouter un critère basé sur l'opinion du créateur du modèle. Ce dernier critère permet de construire un modèle cohérent et interprétable au-delà de critères statistiques et pratiques.

Critères de qualité d'un modèle

Il est possible d'obtenir plusieurs modèles. La qualité d'ajustement d'un modèle sur la base des valeurs observées et les valeurs prédites est évaluée selon différents critères décrits ci-dessous, et seront une aide à la détermination du meilleur modèle.

Déviance : La déviance est un indicateur statistique permettant de mesurer l'adéquation du modèle. Le modèle est comparé avec le modèle saturé qui a la particularité d'avoir autant de paramètres que d'observations et fournit ainsi une description parfaite des données. La déviance est alors donnée par la formule :

$$D = -2(L - L_{sat})$$

avec L la log-vraissemblance du modèle et L_{sat} la log-vraissemblance du modèle saturé.

Cette statistique est positive et plus sa valeur est faible, plus le modèle sera de qualité. Il est possible de montrer que D suit asymptotiquement une loi du χ^2 à $n-p$ degrés de liberté, avec n est le nombre d'observations de la base de données et p le nombre de paramètres estimés par le modèle. Ceci permet de construire un test d'acceptation du modèle.

Un indicateur utile est l'Expected Deviance Ratio (EDR), correspondant à l'amélioration entre la déviance d'un modèle sans aucune variable sélectionnée et la déviance du modèle étudié. L'amélioration de la déviance peut ainsi être quantifiée, plus la baisse de la déviance sera importante, plus le modèle sera performant.

Intervalle de confiance de Wald : Les intervalles de confiance de Wald encadrent les paramètres estimés β afin de tester leur nullité. Lorsque l'un de ces paramètres a un intervalle de confiance de Wald contenant 0, le test de nullité du paramètre n'est pas rejeté et la contribution de la modalité en question dans le modèle peut être remise en doute. Lorsque les intervalles de confiance sont larges pour de nombreuses modalités d'une variable, la significativité de la variable elle-même peut être remise en cause. La méthode de Wald repose sur l'approximation normale des paramètres estimés β, et un intervalle de confiance de seuil $1-\alpha$ est donné par la formule suivante.

$$[\hat{\beta}_j \pm z_{\alpha/2}\sqrt{v_{jj}}]$$

Test de type III : Un autre indice de qualité d'une variable explicative dans un modèle est le test de type III. Pour une variable donnée, le test de type III consiste à comparer la déviance du modèle sans inclusion de la variable concernée avec la déviance du modèle y compris cette variable. En comparant le quantile avec une statistique dite du Chi2 de Wald, le test de type III teste simultanément la nullité de tous les paramètres de la variable. La statistique est asymptotiquement distribuées selon une loi de Chi2_k, avec k le nombre de degré de liberté, i.e. le nombre de paramètres associés à la variable explicative étudiée. La variable sera d'autant plus explicative que sa p-valeur sera petite, l'hypothèse de nullité du coefficient étant rejetée si la p-valeur est inférieure à 5%. Pour rappel, la p-valeur d'un test est le seuil au-dessus duquel l'hypothèse est testée. Plus la p-value est petite, plus l'hypothèse de nullité simultanée de tous les paramètres de la variable est rejetée, reflétant ainsi un gage de significativité de cette variable. Il sera alors possible de classer les p-values des tests de type III par ordre croissant afin d'avoir un classement des variables selon leur contribution au modèle.

AIC/BIC : Le critère d'information d'Akaike (AIC) est une mesure de la qualité d'un modèle statistique. Lorsqu'une variable est ajoutée dans un modèle, cette dernière rajoute un paramètre et permet d'augmenter la vraissemblance. L'AIC pénalise les modèles en fonction du nombre de paramètres afin d'éviter le sur-ajustement. Il peut s'écrire de la façon suivante :

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

avec k le nombre de paramètres du modèle, calculé comme le nombre de modalités de toutes les variables plus 1 moins le nombre de variables, et $\mathcal{L}$ le maximum de la fonction de vraissemblance du modèle.

Ce critère repose sur un compromis entre la qualité de l'ajustement et la complexité du modèle en pénalisant les modèles possédant un nombre de paramètres élevés. Le meilleur modèle sera celui ayant un critère d'AIC le plus faible.

Le critère d'information bayésien (BIC) est l'indicateur dérivé de l'AIC dont l'utilisation est la plus populaire. Il permet de pénaliser le modèle sur la taille de l'échantillon, et non pas seulement sur le nombre de paramètres comme l'AIC. Il peut s'écrire de la façon suivante :

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

avec n le nombre d'observations dans l'échantillon étudié et k le nombre de paramètres. Le modèle le plus performant sera celui qui minimise le BIC.

Coefficient de GINI : Le coefficient de Gini est une mesure statistique permettant de mesurer la robustesse d'un modèle statistique. Il s'agit d'un nombre variant entre 0 et 1 : lorsqu'il est égal

à 0 cela signifie l'égalité parfaite, dans laquelle toutes les classes auraient le même risque, i.e. le modèle ne discrimine personne. Lorsqu'il est égal à 1, le modèle est extrêmement discriminant et cela signifie l'inégalité parfaite. L'inégalité est d'autant plus forte que le coefficient est élevé.

Il est possible de visualiser le coefficient de Gini à l'aide de la courbe de Lorentz (cf. figure 2.1). Il est d'usage en assurance de représenter en abscisse les proportions cumulées de l'exposition tandis que sont représentées en ordonnées les pourcentages cumulés des valeurs prédites par le modèle.

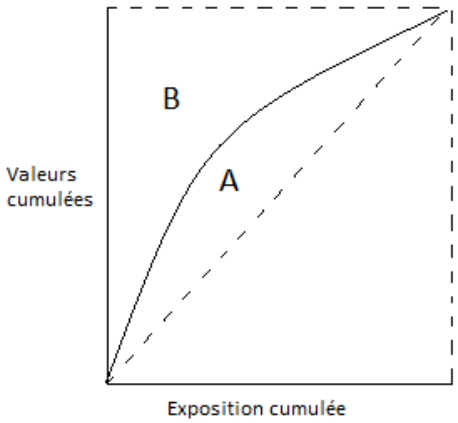


FIGURE 2.1 – Courbe de Lorentz

Lors de la situation la plus égalitaire, la courbe est confondue avec la première bissectrice. Dans les cas alternatifs, le coefficient de Gini se calcule comme étant l'aire entre la courbe et la première bissectrice. Plus l'aire est grande, plus la classification du modèle est bonne. Le modèle ayant le meilleur GINI sera alors considéré comme le meilleur.

$$GINI = \frac{A}{A+B} = \frac{A}{\frac{1}{2}} = 2A = 1 - 2B$$

3 Modélisation de la Responsabilité Civile Matérielle

Dans cette sous-section, nous allons faire l'état des lieux de la modélisation actuelle de la garantie RCM, puis nous décrirons la notion de modèle contraint par opposition à un modèle non-contraint. Enfin nous justifierons notre volonté de segmenter la modélisation de la garantie RCM en deux sous garanties, soit une segmentation des périls.

Etat des lieux

Le modèle de la garantie Responsabilité Civile matérielle actuellement utilisé en tarification chez Direct Assurance a été construit en juin 2015. Plus précisément, la méthode d'estimation de la prime pure d'AXA Global Direct France est une modélisation de la fréquence. La fréquence résultante sera multipliée par une moyenne des coûts moyens (cf. définition de la prime pure dans la section 2.1).

En effet, la modélisation du coût moyen est rendue difficile à cause de la présence des forfaits IRSA décrits dans la section 1.2. La formule de la prime pure RCM est donc définie de la façon suivante :

$$\text{Ancienne Prime Pure} = \text{Fréquence} * \text{Coût Moyen}$$

Le modèle a été construit sur les années de survenance de sinistre allant de 2011 à 2013 et nécessite donc une revue.

Notion de contrainte

Le modèle RCM actuellement utilisé est un modèle contraint, par opposition à un modèle non contraint. Un modèle non contraint permet d'estimer au plus juste la prime pure en s'affranchissant des contraintes informatiques et légales.

Une contrainte informatique est une impossibilité d'intégrer une variable dans les bases, ou demande un développement coûteux. C'est le cas par exemple de la création d'une variable précisant l'âge du conducteur principal en trimestres, facilement calculable par l'actuaire à l'aide de la date de naissance du conducteur. Il est alors nécessaire de calculer la valeur ajoutée de cette variable pour en justifier son développement dans les bases. Il en est de même lorsqu'un actuaire souhaite implémenter un zonier ou classification géographique après une étude approfondie. Dans ce mémoire, nous allons tenter de créer une classification de véhicule nécessitant des développements de plate-forme afin de pouvoir l'utiliser en tarification.

Une contrainte légale est une impossibilité ou une obligation de différencier le tarif en fonction de différentes variables précisées dans le code des assurances, telles le bonus-malus, le sexe de l'assuré et le mode d'obtention du permis de conduire.

Le coefficient de bonus-malus, autrement appelé coefficient de réduction-majoration (CRM), est une variable issue d'une méthode de pondération de l'appréciation du risque en fonction de la sinistralité passée. Définie dans le code des assurances (c'est une variable légale), elle permet de gratifier les conducteurs sans sinistres et de majorer ceux ayant engendré des accidents responsables. Un conducteur sans antécédents d'assurance aura un CRM égal à 100%. Chaque année passée sans sinistre, son CRM sera réduit de 5% ($\text{CRM}_{N+1} = \text{CRM}_N * (1 - 5\%)$) minoré à 50%. Le cas échéant, l'assuré aura une hausse de son CRM de 25% en cas d'accident responsable. Lors de la souscription d'un nouveau client, l'assureur peut vérifier l'exactitude de la déclaration d'un client à l'aide du relevé d'information (RI) fourni par l'AGIRA (Association pour la Gestion des Informations sur le Risque Automobile).

La Cour de justice de l'Union Européenne a également introduit une nouvelle contrainte légale suite à une décision du 21 décembre 2012. Cette dernière stipule que les assureurs n'ont plus le droit de distinguer les hommes des femmes dans leur modèle de tarification, même si ce critère permettait une segmentation statistiquement robuste et importante.

Enfin, le mode d'obtention du permis de conduire fait partie des critères légaux en tarification automobile. La conduite accompagnée (AAC pour Anticipation de l'Apprentissage à la Conduite) permet une réduction légale de la cotisation des conducteurs novices (i.e. moins de 3 ans de permis). La surprime généralement demandée aux conducteurs novices, pouvant atteindre 100% de la cotisation initiale, sera de 50% maximum chez les conducteurs ayant suivi le programme AAC la première année. Elle sera une nouvelle fois réduite la seconde année si le conducteur n'a eu aucun sinistre responsable, et sera supprimée au terme de la 2ème année d'assurance.

La prime pure de la garantie Responsabilité Civile matérielle est donc issue de modèles y compris ces contraintes.

Segmentation des périls

Le modèle actuel ne fait pas de modélisation du coût moyen de la responsabilité civile matérielle. En effet, l'existence des conventions entre assureur rend la modélisation difficile. Il suffit de regarder la distribution des coûts moyens visibles sur la figure 2.2 pour s'en convaincre : il y a une sur-représentation des montants forfaitaires, autour de 1300€, ainsi que des demi-forfaits pour les sinistres à responsabilité partagée. Nous notons également un second pic à 2010€. Ce montant correspond au forfait d'ouverture des sinistres impliquant plus de 3 véhicules communément appelés carambolages. Il est donc nécessaire de retirer les sinistres forfaitaires si nous désirons faire un modèle propre pour le coût moyen.

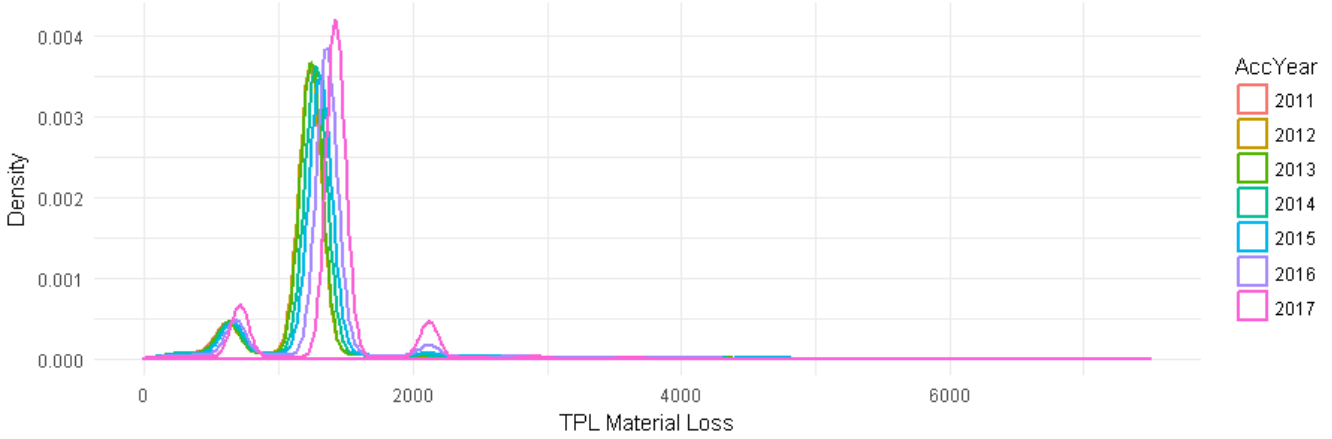


FIGURE 2.2 – Distribution du coût moyen des sinistres RCM par an

Nous allons alors étudier la distribution des sinistres non conventionnés, i.e. au coût réel. La plupart de ces sinistres sont des accidents de droit commun, impliquant des sinistrés dont l'assureur ne fait pas partie de la convention IDA, ou des sinistres dont la facture s'élève à un montant supérieur à 6500€. Cette distribution est visible sur la figure 2.3. Pour les années de survenance 2016 et 2017, nous retrouvons la forte proportion des forfaits d'ouverture des sinistres carambolages à 2010€ ainsi qu'une sur-représentation des montants de 1500€ correspondant au forfait d'ouverture des sinistres de droit commun. Ceci est normal, le développement de ces sinistres étant généralement plus long dans la mesure où la procédure de remboursement n'est pas automatisée et où il est souvent nécessaire d'attendre un procès verbal des forces de l'ordre. Pour les années précédentes nous pouvons néanmoins constater que la distribution du coût moyen semble bien suivre une loi gamma. Il serait alors possible de faire une modélisation à part pour ce genre de sinistres.

Le coût moyen du modèle initial peut alors être remplacé par un modèle de propension, avec un coût moyen des sinistres IRSA déterministe (dans la mesure où nous connaissons le montant des forfaits de l'année courante et de la prévision pour l'année suivante) et une modélisation des coûts moyens non conventionnés. Une alternative à ce modèle de propension serait la modélisation de deux fréquences distinctes : celle des sinistres IRSA et celle des sinistres au coût réel. Ceci ne serait utile que si les deux types de sinistres ont des profils de risques différents, i.e. les accidents

IRSA sont causés par une certaine catégorie de personnes et différente de celle des accidents non conventionnés.

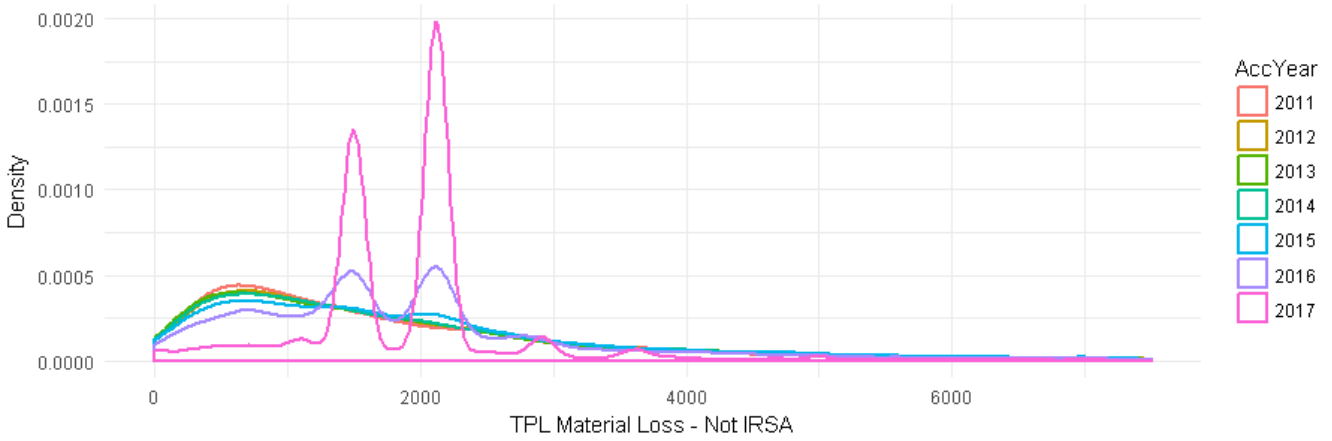


FIGURE 2.3 – Distribution du coût moyen des sinistres RCM hors IRSA par an

Nous pouvons vérifier cette hypothèse en traçant la fréquence des deux types de sinistres par variable. Les figures 2.4 et 2.5 représentent ces fréquences des années de survenance de 2013 à 2015 en fonction de l'âge et de la classe de prix (la fréquence non IRSA est multipliée par un coefficient afin de pouvoir être comparable à la fréquence IRSA). Nous avons également représenté l'intervalle de confiance à l'aide de la méthode de bootstrap en gris foncé. Il semble que les très jeunes ont une tendance à avoir des sinistres non conventionnés plus importante. On peut penser qu'il s'agit de sinistres plus graves ou bien plus d'accident dommage aux biens (choc contre un poteau, il faut alors rembourser le poteau à la ville). Cette tendance s'inverse après 35 ans. De plus, la fréquence des sinistres IRSA pour les jeunes ne fait pas partie de l'intervalle de confiance de l'autre. Les profils générant l'une ou l'autre des fréquences semblent donc bien distincts.

Nous pouvons retrouver ce résultat avec le même graphique en fonction de la classe de prix. Outre les extrêmes où il n'y a pas suffisamment de volume (rendant l'intervalle de confiance peu fiable), les véhicules de faible coût ont plus tendance à générer des sinistres conventionnés, tandis que les véhicules chers (de classe supérieure à M, 13 sur le graphique) génèrent des sinistres au coût

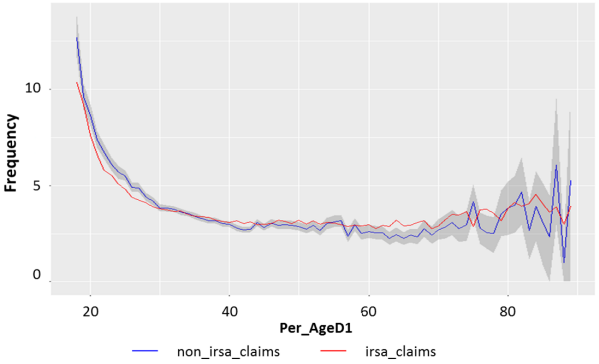


FIGURE 2.4 – Fréquences RCM selon l'âge

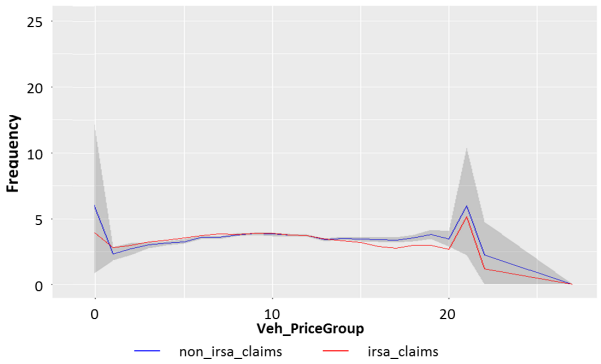


FIGURE 2.5 – Fréquences RCM selon le groupe SRA

réel. On peut remarquer que pour les classes élevées, la fréquence observée des sinistres IRSA reste assez différente de l'intervalle de confiance des sinistres conventionnels.

Nous prenons donc l'hypothèse que les deux types de sinistres sont générés par différents profils de personnes et nous allons faire deux modélisations de fréquence séparées. Ainsi nous aurons deux modèles de fréquence, un modèle de coût moyen, et un coût moyen déterministe. Le résultat sera deux primes pures, une pour les sinistres IRSA et une pour les sinistres conventionnels. Il s'agit là d'une segmentation des périls.

Chapitre 3

Création de la base de données

Dans ce chapitre, il conviendra de décrire et d'analyser les différentes sources de données ainsi que les traitements effectués dans le but de créer une base consolidée permettant d'effectuer l'étude.

1 Traitement des bases de données

Définitions générales

Direct Assurance dispose de plusieurs bases de données sur lesquelles il faut effectuer un travail préalable dans le but d'obtenir une base finale sur laquelle la modélisation sera faite. Elles sont pour la plupart organisées en image. Chaque image correspond à une ligne, et représente une vision de contrat ou de sinistre sur une période donnée. Elle est donc définie avec une date de début d'image et une date de fin. Chaque fois qu'une variable est modifiée, une nouvelle image est créée.

Par exemple, pour les sinistres :

— La première image du sinistre est celle contenant les informations lors de la déclaration du sinistre. La date de début d'image D est donc la date de déclaration, pouvant être postérieure à la date de survenance. Le montant initial du sinistre correspond à un forfait, dit forfait d'ouverture, déterminé par la compagnie d'assurance en attente de plus de finesse dans l'évaluation et la connaissance du sinistre.

— Si 20 jours plus tard l'assureur reçoit une première facture, l'image précédente sera 'fermée' à la date $D + 19$ jours, et une nouvelle image sera créée et 'ouverte' à la date $D + 20$ jours. Cette dernière tiendra compte du nouveau règlement.

— Si au bout de 35 jours l'assureur reçoit une nouvelle facture, la dernière, et permettant de clôturer le dossier, l'image précédente sera fermée et une nouvelle sera ouverte avec le montant additionnel et l'information 'dossier clôturé'.

Base contrat

La base contrat contient toutes les informations liées au contrat d'assurance automobile, comme le numéro de contrat, la date de début, la date d'effet, le numéro de version, le type de paiement (mensuel ou annuel), la suspension ou non du contrat pour quelque motif tels un non-paiement, une mise en demeure ou une interruption dans le cas où le véhicule assuré serait considéré comme une épave suite à un accident. Une image contrat correspond à un identifiant contrat ainsi que sa version de contrat, étendue sur la période entre la date de début et de la fin d'image (ou de version du contrat).

La période d'étude déterminée dans la section 6.1 est de 3 années calendaires permettant ainsi d'analyser la sinistralité avec une saisonnalité comparable. Il convient de scinder les images à cheval sur deux années en deux images, chacune étant rattachée à un exercice. En effet, si un assuré souscrit à un contrat d'assurance le 1er septembre 2012 et n'effectue aucun changement au niveau de son contrat (aucune mise en demeure) ou de critères utilisés en tarification (changement d'adresse ou de situation matrimoniale), il n'aura qu'une version de contrat et l'image du contrat se terminera à échéance, le 31 août 2013. Son exposition sera alors de 1 (cf. définition de l'exposition section 2.1) comme indiqué dans la Figure 3.1. Son image est donc découpée en deux images, l'une en 2012 d'exposition 0.33, l'autre en 2013 d'exposition 0.66.

Lorsqu'une image s'étale entre 2013 et 2014, il conviendra de ne retenir que l'image sur la période 2013, l'année 2014 étant exclue de notre périmètre. Même conclusion pour une image s'étalant entre 2010 et 2011, où seule la seconde période sera retenue.

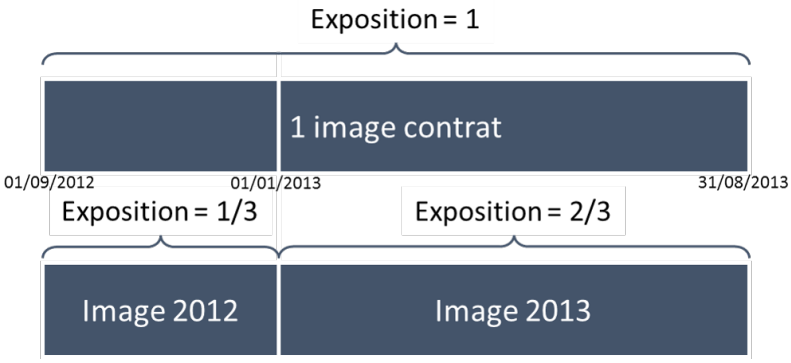


FIGURE 3.1 – Découpage des images de la base contrat

Base personne

La base personne contient toutes les variables liées aux informations des clients, pour chaque version d'un contrat. Il y a plusieurs grandes familles de variables disponibles décrites ci-dessous avec une liste d'exemples non exhaustive.

- Variables personnelles : l'âge du conducteur principal, sa situation matrimoniale, son ancienneté de permis ou le mode d'obtention du permis de conduire (conduite accompagnée, auto-école).
- Variables comportementales : mode d'acquisition du véhicule (crédit, comptant, leasing, héritage), utilisation du véhicule (promenade, étudiant, trajets maison-travail, professionnel)
- Variables géographiques : lieu de stationnement du véhicule le jour ou la nuit, le lieu de résidence de l'assuré.
- Variables contrat : antécédents d'assurance, ancienneté de contrat.

Il est possible de fusionner cette base avec la base des contrats grâce au numéro de contrat et le numéro de version de contrat.

Bases géographiques

D'autres bases sont disponibles telles les bases géographiques, permettant à l'aide du code INSEE, de relier le domicile de l'assuré par exemple à une densité géographique, une moyenne du

nombre de jour de pluie dans la commune, ou la distance par rapport au garage partenaire le plus proche. Ces bases sont le plus souvent issues de partenaires ou fournisseurs officiels, la plus connue étant la base de l'INSEE.

Base sinistre

La base sinistre est également composée d'images, chaque image correspondant à une version du sinistre. Il est possible d'y récupérer les variables comme

- Numéro du sinistre et du contrat
- Date de début et de fin d'image ainsi que le numéro de version
- Date et lieu de survenance du sinistre
- Date de déclaration
- Les garanties mises en jeu (RC Matérielle, Vol, Garantie Personnelle du Conducteur)
- Le type de sinistre (accident 2 véhicules, carambolage, perte de contrôle, vol, etc.)
- Le gestionnaire en charge du dossier
- La responsabilité de l'assuré
- Par garantie
  - L'état du sinistre (clos/en cours/sans suite)
  - Les règlements (REG), encaissements (ENC), reste à régler (RAR) et reste à encaisser (RAE) net de franchise

La charge nette de recours est obtenue par la formule suivante

$$\text{Charge} = \text{REG} + \text{RAR} - \text{RAE} - \text{ENC}$$

Les montants décrits ci-dessus sont nets de franchise. En effet, il convient de créer des modèles sur des montants effectivement payés afin de ne pas pénaliser les clients en leur infligeant une double peine, la franchise permettant justement de baisser le montant des cotisations payées par les assurés. Il n'y a cependant pas de franchises sur la garantie Responsabilité Civile.

Il convient de ne pas étudier les sinistres pour lesquels l'assuré n'est pas responsable, i.e. lorsque la RC de l'assuré est de 0. En effet, on considère qu'un sinistre RC 0 entre dans la garantie Défense Recours pour lesquelles Direct Assurance émettra un recours (forfait ou réel) selon les modalités IRSA, entrainant des coûts moyens potentiellement négatifs. Ne sont également pas pris en compte les sinistres sans suites. En effet on ne modélise la fréquence que sur des sinistres pour lesquels la compagnie a dû indemniser le tiers.

Enfin, il n'est pas nécessaire d'écrêter la charge pour les sinistres au forfait, et donc d'y apporter quelque modification afin de répartir la surcharge des sinistres les plus atypiques sur les autres sinistres. Ceci est néanmoins le cas pour les sinistres au coût réel : les charges dépassant un certain seuil seront coupées et la valeur de point du modèle de coût moyen sera alors ajustée pour tenir compte de cet écrêtement et mutualiser la charge au dessus de la crête (cf. section 6.1).

Chaque dernière version de sinistre est conservée afin d'avoir la vision la plus récente du sinistre (et donc la charge avec le plus de recul). La base sinistre peut alors être fusionnée avec la base des contrats et critères tarifaires évoquée aux paragraphes précédents à l'aide du numéro de contrat, de sorte que la date de survenance du sinistre soit comprise entre la date de début et de fin d'image du contrat.

2 Base de données véhicule

Base brute

La base de données véhicules est fournie par la SRA (cf. section 4.1) et contient des variables descriptives de véhicules selon marque, modèle, type et version. Les variables peuvent être décomposées en plusieurs catégories :

- Performances : vitesse maximale en km/h
- Moteur :
  - Alimentation : GNV, GPL, hybride, injection directe/indirecte/suralimentée
  - Energie : bioéthanol, essence, GNV, gasoil ou GPL
  - Nombre de cylindres
  - Disposition des cylindres : en ligne, à plat, en V ou en W
  - Cylindrée (en m^3)
  - Puissance (en Watt)
  - Régime (en DIN)
- Transmission
  - Transmission : traction, propulsion, 4 roues motrices permanentes et 4 roues motrices débrayables
  - Boîte de vitesse : manuelle, automatique ou semi-Automatique
  - Nombre de rapports
- Châssis :
  - Carrosserie : permet de distinguer les berlines (3 portes ou 5 portes), les fourgonnettes, les coupés, les cabriolets, les automobiles familiales (monospace), etc.
  - Suspension : 4 roues indépendantes, essieu arrière semi rigide, essieu arrière rigide et suspension active
  - Type de freinage : 2 ou 4 disques, 2 ou 4 tambours
- Dimensions
  - Nombre de places : cette variable peut représenter un intérêt potentiel en garantie responsabilité corporelle
  - Longueur, largeur, hauteur
  - Poids à vide et PTAC : Le PTAC (Poids Total Autorisé en Charge) correspond à la charge maximale de marchandises ainsi que le poids du chauffeur et de tous les passagers. Cette variable est intéressante dans la mesure où elle permet de différencier les véhicules destinés au transport de personnes et veux au transport de marchandise (comme les fourgonnettes).
- Équipement
  - Airbags : passager avant, latéraux avants et arrières. Chacune de ces variables a trois modalités (série, option et non). Les véhicules récents ont de plus en plus d'airbag conducteur en série, nous n'avons donc pas retenu cette variable dans l'étude.
  - Antiblocage des roues : série, option ou pas d'anti-blocage des roues. Mécanique ou Automatique.
- Autres
  - la classe de prix (d'origine ou actuelle, mise à jour chaque année), le groupe SRA et la classe de réparation (d'origine ou actuelle). Ces variables seront détaillées dans la section 4.1
  - la date de début et de fin de commercialisation

Il faut noter que les informations sur les options sont assez limitées. En effet, on ne connaît que si l'option est proposée à la carte, en série, ou n'est pas disponible, car elle s'applique à tous les véhicules de même marque, modèle, type et version. En revanche, de nouveaux acteurs apparaissent sur le marché (SIDEXA) et permettent de savoir si tel ou tel véhicule a l'option en question en fonction de la plaque d'immatriculation (et du type mines). Ces bases de données sont néanmoins payantes.

Retraitement

Il conviendra de discuter du traitement de la base de donnée initiale dans cette partie, dans la mesure où la partie application ne concerne que les véhicules faisant partie du portefeuille de Direct Assurance, i.e. après fusion avec la base des contrats.

La base contient environ 100.000 lignes correspondant à tous les véhicules commercialisés. Certaines marques de véhicules ne sont pas beaucoup représentées. Par exemple, seuls deux véhicules de la marque Asia sont présents, il s'agit de la Rocsta 2.2 DX Soft Top et de la Rocsta 2.2 GT Soft Top. De même 4 véhicules de la marque Bertone sont présents. Il s'agit d'une marque italienne spécialisée dans le haut de gamme qui a cessé son activité en 2014. Au vu de cette multitude de marques inconnues et peu représentées, les marques ayant moins de 1000 véhicules sont arbitrairement regroupées dans une catégorie "autres" pour ne pas polluer l'étude. La distribution des marques est visible sur la figure 3.2. La marque ayant le plus de modèles ayant été commercialisés est Renault avec 12.000 modèles, suivie par BMW puis Volkswagen. Ceci ne veut pas dire que BMW est le deuxième acteur sur le marché français, il s'agit ici du nombre modèles différents de véhicules commercialisés et non du nombre de véhicule en circulation, ni même du nombre de véhicules assurés chez Direct Assurance.

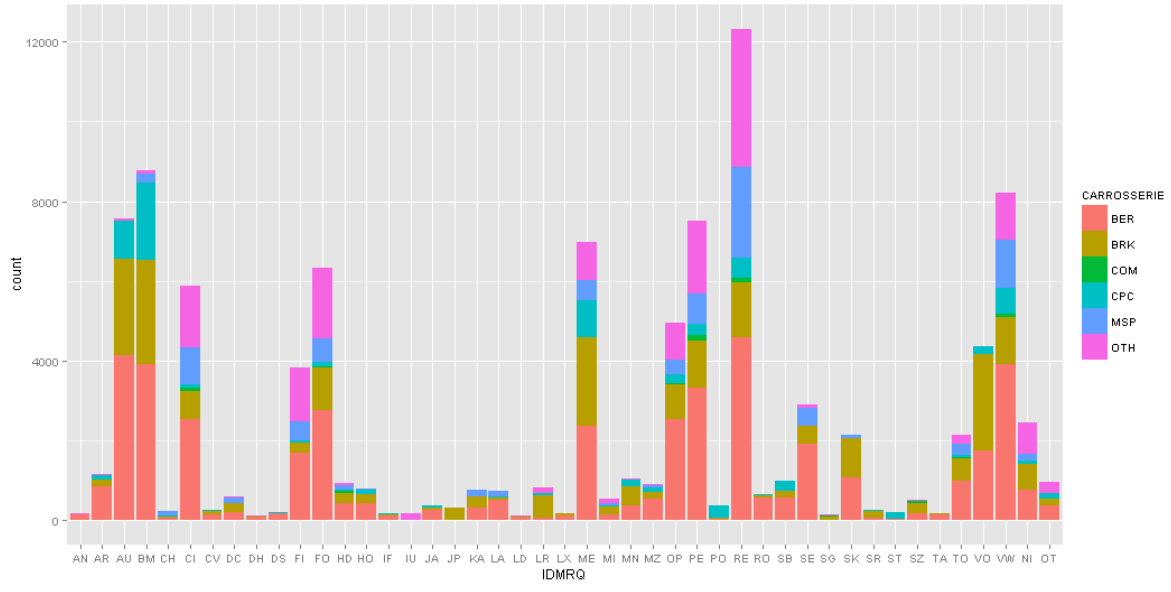


FIGURE 3.2 – Distribution de la marque x carrosserie dans la base SRA

La variable carrosserie contient beaucoup de modalités pour peu d'information apportée, telles BE3 vs BE5 pour différencier les berlines 3 portes des berlines 5 portes. De plus beaucoup de carrosseries ne sont pas assurables selon les règles de souscription, tels les camions. Il conviendra

donc d'avoir les catégories suivantes : Berline, Break, Commerciale, Coupe/Cabriolet, Autres. La figure 3.2 montre que chaque marque peut avoir des cibles différentes. Par exemple, la proportion de Coupés Cabriolets est bien plus importante chez BMW que chez Renault, qui privilégie des modèles Monospaces familiaux et des fourgonnettes (modalité OTH).

Seulement 12% des véhicules de la base totale sont encore commercialisés. Outre ces véhicules, 1% des véhicules de la base n'ont pas de durée de fin de commercialisation renseignée. Il s'agit de véhicules commercialisés avant 1990. Il conviendra d'utiliser la durée moyenne de commercialisation pour calculer la date de fin de commercialisation. La variable Durée de commercialisation est par la même occasion créée. Les fabriquants de véhicules commercalisent en moyenne leurs véhicules pour 1 an et demi en moyenne, mais la majeure partie des véhicules est commercialisée un an ou moins. La distribution est visible sur la figure 3.3. La durée de commercialisation maximale est de 38 ans pour l'Austin Mini. Cependant il n'y a que peu de véhicules commercialisés plus de 10 ans.

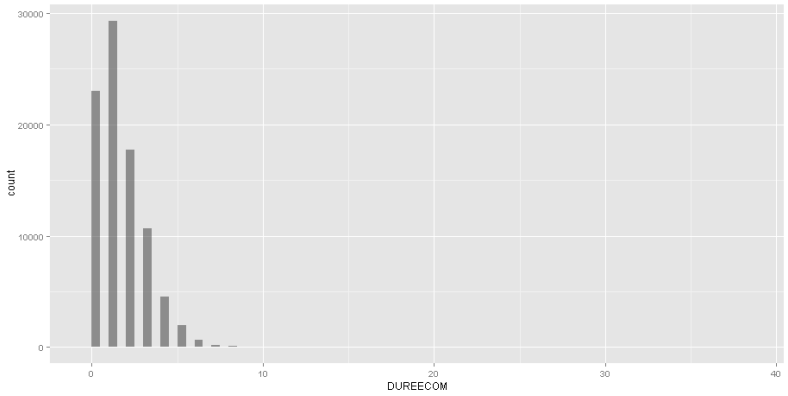


FIGURE 3.3 – Distribution des véhicules selon la durée de commercialisation

La variable nombre de rapport de vitesse est égale à 0 pour 2% des véhicules de la base, la majeure partie pour des années de commercialisation récentes. Les valeurs manquantes sont réparties sur l'ensemble des puissances réelles. Il s'agit d'une grande partie des Lexus, Audi, Toyota, Nissan et Subaru. En poussant les recherches, on remarque que tous ces véhicules sont électriques ou hybrides. La figure 3.4 représente en pourcentage la part du nombre de rapport par énergie. Les rapports à 0 pour les ES (essence), GN (Gaz Naturel) et GO (Gasoil) sont des hybrides. Cette variable reste donc cohérente et il ne s'agit pas d'aberrations. Toutefois, elle est quantitative. Afin de ne pas fausser l'étude future il convient de remplacer les 0 par la modalité la plus représentée, soit 5 vitesses.

La variable freinage d'urgence n'est pas très bien renseignée, notamment commercialisés avant 2000. La première montée en série sur véhicule a eu lieu en 1996 sur les Mercedes-Benz Classe S et Classe SL [1]. En 1998, le système est généralisé sur tous les véhicules de la marque. On peut donc renseigner la modalité "série" pour ces véhicules lorsqu'elle est manquante. Aujourd'hui la totalité des voitures sont équipées de ce système.

Les variables de dimensions ne sont pas très bien renseignées non plus pour les véhicules commercialisés avant 2000. Afin de ne pas avoir de modalités nulles, elles sont remplacées par la moyenne selon la marque et la carrosserie lorsque c'est possible, puis la moyenne selon la carrosserie. Par exemple, la longueur d'un break de la marque Audi, lorsqu'elle est manquante, aura la longueur

1. https ://www.ornikar.com/code/cours/securite/aide-freinage

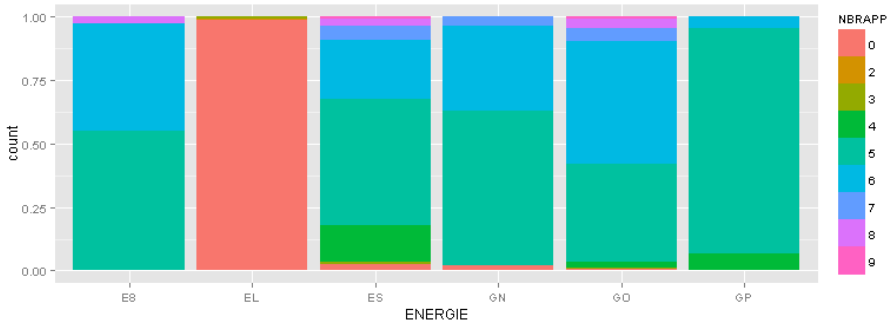


FIGURE 3.4 – Distribution empilée 100% de la marque x nombre de rapports

moyenne des breaks Audi. S'il s'agissait du seul véhicule break commercialisé par la marque, elle aurait été remplacée par la longueur moyenne de tous les breaks, sans différence selon la marque.

La base de véhicules SRA est alors complètement exploitable.

3 Échantillonnage de la base de données

Dans le but d'estimer la prime pure du portefeuille de Direct Assurance, il est nécessaire de modéliser la fréquence et le coût moyen de la garantie étudiée. Ceci est fait à l'aide des modèles linéaires généralisés, qui nécessitent un échantillonnage de la base de donnée initiale. Tout d'abord un échantillon de modélisation, d'apprentissage, ou test set, sur lequel le modèle s'entraînera. Il représente généralement 80% de la base initiale. Ensuite, le modèle sera testé sur un échantillon de validation, ou validation set, constitué de 20% de la base. Cet échantillonnage est fait aléatoirement.

L'un des objectifs de ce mémoire est également de créer une classification de véhicule. Il est donc nécessaire de segmenter l'échantillon de modélisation de façon aléatoire en deux sous échantillons : 60% sera utilisé pour la construction du véhiculier, et 20% serviront à déterminer les indicateurs optimaux de classification.

Afin d'éviter la redondance de langage, l'appellation « échantillon de modélisation » sera utilisée pour nommer les 80%, et l'appellation « échantillon de validation » pour nommer les 20%.

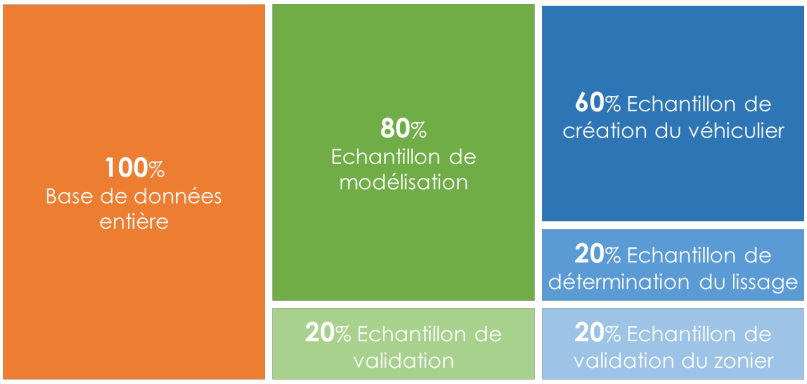


FIGURE 3.5 – Échantillonage de la base de données

Deuxième partie

Approche innovante d'une classification de véhicules

Chapitre 4

Classification actuelle

Avant d'aborder la nouvelle approche de classification de véhicule, nous exposerons dans ce chapitre les classifications utilisées actuellement dans la modélisation des garanties par Direct Assurance.

1 Sécurité et Réparation Automobile

La classification de véhicules la plus répandue en assurance automobile est la classification fournie par l'association Sécurité et Réparation Automobile [1]. Créée en 1977, la SRA (Association Sécurité et Réparation Automobile) est une association composée de toutes les sociétés d'assurance automobiles françaises membres de la FFA (Fédération Française de l'Assurance) et du GEMA (Groupement des Entreprises Mutuelles d'Assurance).

En collaboration avec les constructeurs automobiles, la SRA a pour vocation de diffuser toutes études et de mettre en œuvre tous moyens permettant une limitation de la fréquence et des coûts de sinistres dans l'intérêt des assurés. Elle suit régulièrement les évolutions des prix des pièces de réparation ou des taux horaires de main d'œuvre chez différents constructeurs dans une optique de diffusion de renseignements lors de la publication de lettres trimestrielles. Elle fournit notamment une liste de garages « recommandés » pour les réparations automobiles.

La SRA encourage la sécurité des véhicules et réalise des études de sécurité active et passive. La sécurité active est l'ensemble des éléments permettant la prévention d'un accident, telles les sécurités de perception (visuelle ou acoustique), de conduite (direction assistée, boîte de vitesse) ou de commande (suspension, freinage, régulateur/limiteur de vitesse). La sécurité passive quant à elle permet, au moment de l'accident, de limiter sa gravité à l'aide des airbags, de la ceinture de sécurité ou du pare-brise par exemple. Ces études, faisant partie de la veille technologique de la SRA, seront une aide précieuse lors de la réalisation de la classification.

Outre ses différents suivis et études, la SRA a un devoir d'information envers les assureurs concernant les véhicules. Pour ce faire, elle met à disposition différentes bases de données recueillant les caractéristiques techniques et commerciales de véhicules terrestres à moteurs (automobile, moto et quad, voiturettes). Ces bases de données peuvent alors être utilisées par les adhérents afin de croiser leurs variables de tarification avec leurs propres données, comme l'âge ou l'usage d'un véhicule pour une police de leur portefeuille. En effet, la SRA fournit une classification de véhicules utilisée

1. Sécurité et Réparation Automobile ou SRA, http ://www.sra.asso.fr/

par les assureurs automobiles dans leurs calculs de tarification. Cette classification sera décrite dans les prochaines sections.

La SRA participe également à l'élaboration d'une classification de véhicules centrée sur la protection contre le vol. Cette nouvelle classe de véhicule définie en 'clés' est principalement portée par la présence de système de protection antivol et est notamment utile lors de la création d'un modèle pour la garantie vol présente dans la plupart des formules proposées par les assureurs.

La SRA joue donc un rôle de centre d'étude et de communication, autant à destination des assureurs français que des constructeurs automobiles, dans le but d'optimiser la productivité de ses adhérents et de minimiser les dépenses liées aux accidents de la route.

2 Indicateurs disponibles

La SRA propose trois indicateurs permettant la classification de véhicules et concernant tous les véhicules particuliers et utilitaires d'un poids inférieur à 3,5 tonnes, commercialisés à destination du marché français par un constructeur ou un importateur officiel.

Le Groupe

Cet indicateur reflète la dangerosité intrinsèque des véhicules, indépendamment du conducteur, de l'usage ou de la zone géographique de circulation. Il est lié à la probabilité de survenance d'un accident à travers les caractéristiques techniques d'un véhicule. Le Groupe est la base de l'assurance obligatoire : la Responsabilité Civile.

La valeur du groupe s'obtient par la formule suivante

$$20 + \left(a * \frac{P_{reelle}}{M_{vide} + \varepsilon} + b * (V_{max} - V_{ref}) + c * PTAC \right) + (1 + d * N_{conception})$$

avec

- P_{reelle} la puissance du véhicule en chevaux
- M_{vide} le poids à vide du véhicule en kg
- $PTAC$ le Poids Total Autorisé en Charge en kg
- V_{max} la vitesse maximale du véhicule en km/h
- V_{ref} une vitesse de référence établie à 130km/h
- $N_{conception}$ une note de conception
- a, b, c, d et ε des constantes

Le facteur ε, fixé à 200 kg, correspond environ à la masse de deux passagers et leurs bagages, tandis que la vitesse de référence V_{ref}, établie à 130 km/h correspond à la limitation de vitesse actuelle.

Le rapport Puissance/Masse du véhicule permet de distinguer les véhicules performants et dynamiques des véhicules nécessitant une forte puissance au vu de leur volume imposant. Il permet donc de refléter la dangerosité intrinsèque du véhicule.

Le paramètre mesurant la sécurité globale du véhicule est $N_{conception}$. Celle-ci dépend de la tenue de route du véhicule (type de transmission, présence d'un système de contrôle dynamique de stabilité), du freinage (présence d'un système d'antiblocage des roues ou d'aide au freinage d'urgence) ainsi que du comportement sécuritaire du véhicule (sécurité passive, airbags, résultats de crash tests). Le groupe est donc un indicateur prenant les valeurs de 20 à 50, avec une distribution pseudo normale centrée autour de 32 (cf. Fig. 4.1).

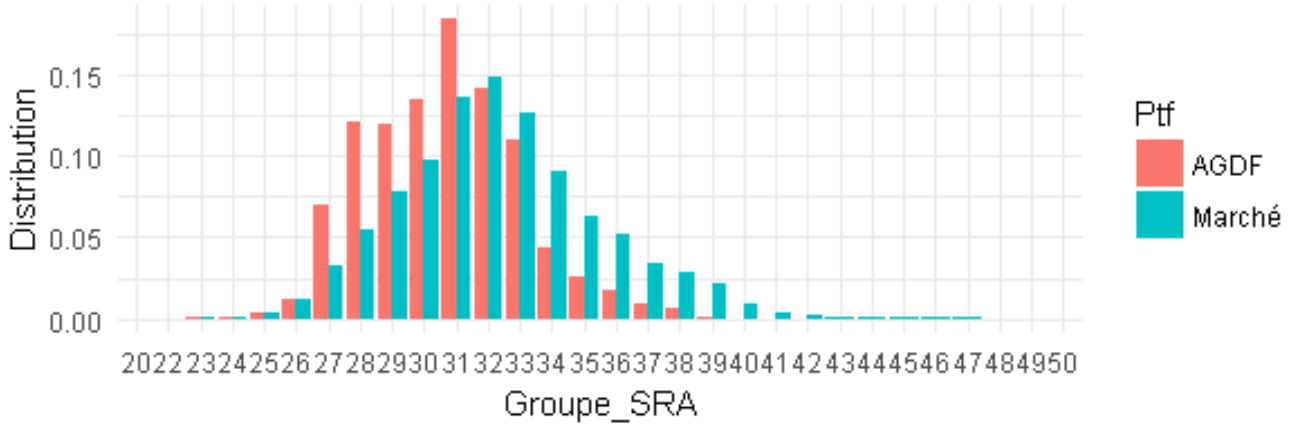


FIGURE 4.1 – Distributions du groupe du marché et du portefeuille Direct Assurance

La Classe de Prix

La classe de prix est un indicateur lié à la valeur à neuf TTC du véhicule lors de sa mise sur le marché (hors options et remises). Elle est définie par tranches (cf. Table 4.1). Cet indicateur est utilisable par les assureurs en cas de vol ou de perte totale (véhicule économiquement irréparable) ainsi qu'en tarification.

FIGURE 4.2 – Distributions de la classe de prix du marché et du portefeuille Direct Assurance

La classe de prix actuelle est un deuxième indicateur lié à la valeur de marché du véhicule, après dépréciation. Elle est remise à jour chaque année. On peut remarquer que la distribution au sein du portefeuille de Direct Assurance est différente de celle des véhicules du marché (cf. Fig. 4.2). En effet Direct Assurance n'accepte pas de souscrire des véhicules de classe SRA supérieure à T.

La Classe de Réparation

La classe de réparation est également un indicateur monétaire (hors taxe). C'est une estimation du coût total des réparations d'un véhicule, calculé en fonction du prix actuel des pièces de remplacement, de l'inflation, du taux horaire de la main d'œuvre ainsi que des coûts de réparations

Min <	Tranche	≤ Max
	A	7683
7683	B	8561
8561	C	9549
9549	D	10646
10646	E	11853
11853	F	13170
13170	G	14706
14706	H	16463
16463	I	18439
18439	J	20744
20744	K	23267
23267	L	26122
26122	M	29305
29305	N	32925
32925	O	37097
37097	P	41816
41816	Q	47194
47194	R	53340
53340	S	60365
60365	T	68377
68377	U	77486
77486	V	87913
87913	Z	

TABLE 4.1 – Tranches de la classe de prix SRA

Min <	Tranche	≤ Max
	A	564
564	B	677
677	C	795
795	D	917
917	E	1046
1046	F	1181
1181	G	1321
1321	H	1468
1468	I	1622
1622	J	1782
1782	K	1950
1950	L	2124
2124	M	2307
2307	N	2498
2498	O	2698
2698	P	2907
2907	Q	3124
3124	R	3352
3352	S	3593
3593	T	3840
3840	U	4100
4100	V	4373
4373	Z	

TABLE 4.2 – Tranches de la classe de réparation SRA

facturés suite à des résultats de crash test fournis par les constructeurs. Elle est définie par tranches actualisées annuellement en fonction des évolutions de marché (cf. Table 4.2).

Limites

Les trois indicateurs fournis par la SRA sont utilisés par la majeure partie des assureurs, ne permettant pas une segmentation du tarif plus fine et adaptée au portefeuille de chacun. En effet, un assureur dont les règles de souscription telle que les véhicules de classes SRA supérieure à S afin d'éviter d'avoir des véhicules trop coûteux dans son portefeuille pourrait affiner les indicateurs. On remarque cette distorsion entre le marché et le portefeuille de Direct Assurance dans les Classes de prix faibles et élevées (Fig. 4.2) ainsi que sur les groupes pour lesquels les distributions ne sont pas tout à fait similaires, notamment entre les groupes 31, 32 et 33 (Fig. 4.1). Le développement d'une classification de véhicule propre à un portefeuille peut être un atout pour un assureur et lui permettrait d'avoir un tarif plus compétitif que ses concurrents.

D'autre part, les informations comme l'énergie ou la carrosserie du véhicule ne sont pas prises en compte dans les classes de la SRA. La création d'une classification unique, regroupant les mêmes informations que les indicateurs SRA ainsi que d'autres variables disponibles, est un moyen de simplifier les modèles de tarification.

Chapitre 5

Création d'une Nouvelle Classification

Dans ce chapitre, nous allons décrire la méthodologie de classification de véhicule. Il conviendra d'exposer dans un premier temps la théorie des zoniers. En effet, la création du véhiculier est inspirée de ce processus, à la grande différence qu'il n'existe pas de représentation spatiale des véhicules. Il est alors nécessaire de créer une "carte" de véhicules, ou une relation de voisinnage. Ceci est possible à l'aide de méthodes d'analyse de données ainsi qu'à la Triangulation de Delaunay, dont les principes seront décrits dans la section suivante. Enfin, le processus de lissage spatial et différentes méthodes de classification, permettant d'obtenir le véhiculier final, seront évoquées. La dernière section de ce chapitre proposera un récapitulatif du processus global.

1 Inspiration de la théorie des zoniers

L'utilisation de classification géographique est de plus en plus commune en tarification de produits d'assurance non vie. Communément appelées zoniers, ces variables sont développées sur un portefeuille spécifique afin de lier un risque à sa localisation sous forme de classes, habituellement numérotées ou nommées alphabétiquement, de la zone la moins risquée à la zone la plus risquée. Le nombre de classes peut varier en fonction de l'exposition totale du portefeuille, afin que les résultats restent robustes. Ces zoniers permettent de cartographier le risque du portefeuille d'un assureur sous différentes granularités, telles les départements, les codes postaux, les codes commune (INSEE), les codes IRIS, et récemment un carroyage du territoire (carrés de 100m de côté).

Le développement initial de la classification géographique avait pour but d'estimer le risque des zones dans lesquelles l'assureur ne disposait pas de suffisamment d'exposition pour en déduire des résultats fiables et robustes. L'idée principale pouvait se résumer de la façon suivante : « tout est relatif, mais le risque d'une zone est plus lié au risque des zones voisines que des zones lointaines ».

Au-delà d'une estimation du risque dans les zones à faible exposition, l'hypothèse principale des zoniers est que les résidus des modèles linéaires généralisés (autrement dit la part non expliquée du modèle) contiennent, outre du bruit, une information liée à des données géographiques, appelée variation résiduelle géographique. Ce pourrait être le cas de la densité, rarement intégrée dans les modèles, qui pourrait être intégrée dans un modèle de fréquence de sinistres : plus la densité de population est importante dans la commune de résidence de l'assuré, plus la probabilité d'avoir un accident est importante. De même, le caractère urbain ou rural pourrait influencer le coût moyen des sinistres : les coûts de réparation peuvent être moins chers dans les zones urbaines où la compétitivité est plus importante, et inversement pour les zones rurales. La classification géographique repose donc sur l'isolation, puis la classification de cette variation résiduelle, créant ainsi les zoniers.

Nous pouvons transposer l'hypothèse évoquée ci-dessus dans le cadre d'une classification de véhicules (Fig. 5.1). Le résultat d'un GLM se décompose entre la sinistralité estimée (Fitted) et les résidus (Residuals), tentant d'estimer au mieux la sinistralité observée (Observed). La sinistralité estimée dépend de critères liés ou non aux véhicules, tandis que les résidus sont formés de bruit (Noise), ainsi que d'un signal non capté par la sinistralité estimée, appelé variation résiduelle véhicule. Une classification sera appliquée sur ce dernier afin d'obtenir les classes de véhicules.

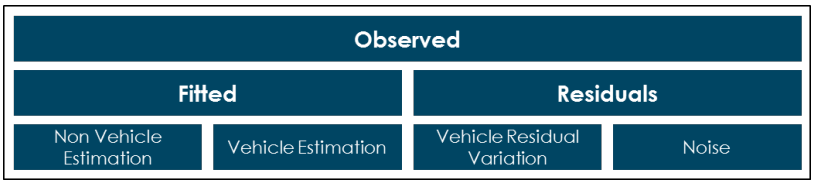


FIGURE 5.1 – Hypothèse de décomposition du signal de sinistralité

Il est important de connaître une alternative à la méthode de classification développée ci-dessus. En effet, dans le cas de la classification géographique, il est probable que l'assureur veuille remplacer la totalité des variables géographiques par son zonier. C'est notamment le cas lorsqu'une variable sociodémographique par exemple, intégrée au GLM, ne peut pas être récupérée périodiquement. Dans ce cas, les classes du zonier seront déterminées à l'aide de la variation résiduelle ainsi que de la sinistralité estimée par les variables géographiques. Il en résultera une perte d'information par rapport au modèle initial accompagnée d'une simplification du modèle, dans la mesure où ce dernier pourra s'affranchir de toutes les variables géographiques initialement inclues dans le modèle.

Un problème se pose dans l'une ou l'autre des deux méthodes évoquées ci-dessus. Une donnée essentielle que l'on peut facilement récupérer en géographie est absente du processus de classification de véhicule : une carte. En effet, la classification géographique dispose de cartes de différentes granularité (départementales, INSEE ou autres) permettant de déterminer une distance ou une relation de voisinage entre divisions. Il est donc nécessaire de passer par une étape de création de «carte» de véhicules avant d'utiliser le processus géographique.

2 Création d'une table de contiguïté

2.1 Représentation spatiale de véhicules

Le processus de classification nécessite donc une carte. Cette question ne se pose guère en classification géographique dans la mesure où plusieurs cartes sous différentes granularités sont disponibles. Quant à la classification des véhicules, comment obtenir une carte d'automobile ?

Afin de résoudre ce problème, il conviendra de s'inspirer des concepts de l'analyse de données, plus particulièrement de l'analyse factorielle. A la différence de l'analyse descriptive qui permet d'étudier séparément chaque variable sans toutefois donner une vue d'ensemble des liaisons et interactions entre variables, l'analyse factorielle permet d'avoir une vision macro de ces relations. Habituellement utilisée dans le but de synthétiser l'information, l'analyse factorielle permet l'étude d'un nuage de points de dimension > 3 résumé dans un nuage de points dans l'espace, voire le plan, tout en minimisant la distorsion par rapport aux données initiales. C'est ce nuage de points qui constituera notre carte.

Parmi les différents types d'analyse factorielle disponibles, deux grandes méthodes sont prédominantes :

— Analyse en Composantes Principales (ACP)
— Analyse par Correspondances Multiples (ACM)

Nous allons en décrire ci-dessous les grands principes.

Analyse en Composantes Principales

L'Analyse en Composantes Principales est une méthode factorielle de réduction de dimension pour l'exploration statistique de variables quantitatives. En d'autres termes, une ACP cherche à construire un espace de dimension réduite qui préserve les distances entre les individus. On peut la voir comme une compression avec perte (contrôlée) de l'information [18].

Soit X une matrice d'ordre n x p, où X_{ij} représente la valeur de la variable quantitative j prise par l'individu i. Afin de déterminer la ressemblance entre deux individus il est nécessaire d'introduire la distance euclidienne

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

Il est également nécessaire de définir l'inertie du nuage de points dans l'espace d'origine :

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

avec G le barycentre. L'inertie indique la dispersion autour du barycentre.

Le premier objectif de l'ACP est de construire la meilleure représentation graphique des individus i dans un espace de dimension q inférieure à p, telle que la déformation du nuage de points soit minimale. Pour y parvenir, il faut déterminer les composantes principales, vecteurs définissant les projections optimales telles que la différence entre la distance euclidienne entre deux points A et B et la distance projetée soit la plus petite. Ces projections devront être orthogonales. Il faut également que ce plan maximise l'inertie des points. Cela revient à maximiser la moyenne des coefficients de corrélation entre chaque variable et les facteurs dans la nouvelle projection.

$$\sum_{q} r^2(z,q)$$

avec z un facteur dans la nouvelle projection, et r le coefficient de corrélation linéaire.

Le deuxième objectif d'une ACP consiste à représenter graphiquement les variables dans un espace de dimension q inférieure à n, de façon à minimiser la perte d'information des liaisons entre les variables. De façon similaire aux projections des points, la projection optimale est celle qui minimise la déformation des angles entre les vecteurs des variables.

Les graphiques utilisés pour l'interprétation d'une ACP sont l'histogramme des variances des composantes principales (vecteurs des projections des individus déterminées ci-dessus), le nuage d'individus dans le nouvel espace ainsi que le cercle des corrélations, obtenu par les projections des variables réduites.

Afin de déterminer la dimension à retenir, c'est-à-dire le nombre de composantes principales, plusieurs méthodes sont possibles :

— la part d'inertie expliquée, fixée à priori par l'utilisateur
— le critère de Kaiser, tel que les composantes conservées sont celles ayant une variance supérieure à 1
— le critère du "coude", méthode visuelle recherchant une cassure dans la décroissance des variances des composantes principales
— le critère de Karlis, Saporta & Spinakis, donné par l'expression

$$floor = 1 + 1.65\sqrt{\frac{p-1}{n-1}}$$

avec p l'inertie totale et n le nombre d'observations.

Pour plus de détails mathématiques, le lecteur est invité à se référer aux cours de CLOT [6], aux TD de DUFOUR [9] ou au mémoire GONNET [2010] [12].

Analyse par Correspondance Multiple

L'Analyse par Correspondance Multiple est une méthode factorielle de réduction de dimension pour l'exploration statistique de données qualitatives. C'est une généralisation de l'Analyse Factorielle des Correspondances et permet une description des relations entre q ($q > 2$) variables qualitatives simultanément observées sur n individus. Le cours de RAKOTOMALALA [17] parmi d'autres références, permet de comprendre les enjeux et la théorie de l'ACM. Il conviendra de n'exposer que l'objectif et le principe de cette méthodologie.

Soit X une matrice d'ordre n x p, où X_{ij} représente la valeur de la variable qualitative j prise par l'individu i. L'objectif de l'ACM est de trouver un repère factoriel préservant au mieux les distances entre les individus, i.e. permettant de discerner au mieux les individus entre eux. Le principe est d'obtenir une projection telle que la moyenne des rapports de corrélation entre chacune des variables qualitatives et les nouveaux facteurs dans la projection soit la plus grande possible (cf. TENENHAUS [2006] [21]) :

$$\sum_q \eta^2(z, q)$$

avec z un facteur dans la nouvelle projection, et η le rapport de corrélation.

Concrètement, le facteur est construit de manière à ce que, globalement, on distingue au mieux entre elles les modalités de chaque variable, i.e. pour chaque variable, "ses modalités soient le plus étalées possibles sur l'axe factoriel" [17].

Pour plus de détails mathématiques, le lecteur est invité à se référer aux cours de CLOT [6], aux TD de DUFOUR [9] ou au mémoire GONNET [2010] [12].

Mélange entre une ACP et une ACM

Lorsque les données à explorer sont de même nature, l'utilisation de l'Analyse en Composantes Principales ou l'Analyse des Correspondances Multiples est privilégiée. Cependant, l'introduction de variables qualitatives et quantitatives au sein d'une même analyse factorielle est une problématique récurrente. Une possibilité est de coder les variables quantitatives en variables qualitatives. Ceci peut s'avérer problématique si on ne souhaite pas changer la nature ni définir de groupes pour des

variables quantitatives (e.g. un premier groupe pour les âges entre 18 et 25 ans, un second pour ceux entre 26 et 35 ans, etc.).

ESCOFIER [1979] [11] proposa l'introduction des variables quantitatives à l'aide d'un code approprié dans une ACM. SAPORTA [1990] [19] démontre qu'il est également possible de juxtaposer des variables qualitatives codées dans un tableau disjonctif complet dans une ACP avec des variables quantitatives réduites. Ces deux méthodes convergeant, PAGES [2004] [14] publie les propriétés d'une méthodologie commune appelée Analyse Factorielle de Données Mixtes. Mixtes se réfère au fait que les variables qualitatives sont étudiées conjointement avec les variables quantitatives. Une autre publication de PAGES [2004] [15] permet également de rappeler la méthodologie de l'AFDM et propose une application. Ce dernier document est repris ci-dessous pour présenter le principe de l'AFDM.

Soit X une matrice d'ordre n x $(p+q)$, où l'individu i est décrit par :
— K_1' variables quantitatives, $k \in [[1; K_1']]$. Ces variables seront toujours supposées centrées réduites. Ceci n'est pas une commodité mais une nécessité due à la présence des deux types de variables ;
— Q variables qualitatives, $q \in [[1; Q]]$. La q^e variable présente K_q modalités, $k \in [[1; K_q]]$. L'ensemble des modalités a pour cardinal $\sum_q K_q = K_2'$.

Soit $K = K_1' + K_2'$ le nombre total de variables quantitatives et de variables indicatrices. La figure 5.2 (proposé par [Pagès]) rassemble les notations. Les variables qualitatives apparaissent à la fois sous leur forme condensée et sous leur forme disjonctive complète.

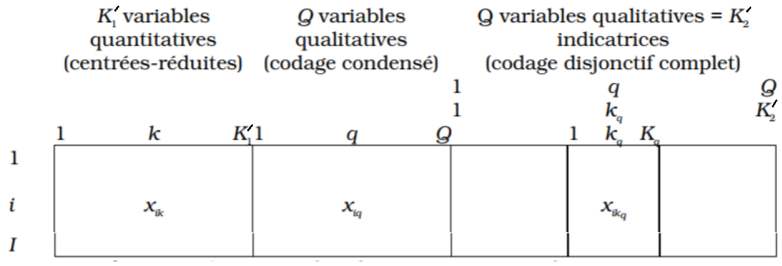


FIGURE 5.2 – Structure des données et principales notations

avec :
— x_{ik} la valeur de i pour la variable (centrée-réduite) k ;
— x_{iq} la modalité de i pour la variable q ;
— $x_{ikq} = 1$ si i possède la modalité k de la variable q et 0 sinon.

Le principe de l'AFDM est de trouver les facteurs z de façon à maximiser l'inertie projetée du nuage N_K (comportant à la fois les variables quantitatives et qualitatives). Cela revient à maximiser le coefficient de corrélation pour les variables quantitatives et le rapport des corrélations pour les variables qualitatives, soit la quantité :

$$\sum_{k \in K} r^2(z,k) + \sum_{q \in Q} \eta^2(z,q)$$

Les graphiques utilisés pour l'interprétation d'une AFDM sont les mêmes que pour les autres analyses factorielles, à savoir le nuage des individus par sa projection sur ses axes d'inertie, le cercle

des corrélations des variables quantitatives, et les modalités de variables qualitatives par les centres de gravité des individus correspondant.

Pour plus de détails mathématiques, le lecteur est invité à se référer à PAGES [2004] [14] ou à PAGES (Agrocampus) [2004] [15].

2.2 Triangulation de Delaunay

Une fois le nuage de véhicules constitué à l'aide de l'analyse factorielle de données mixtes, il est nécessaire de définir la relation de voisinage entre les points. Une première approche a été de considérer que tous les points situés dans un cube de longueur définie étaient considérés comme voisins. Elle n'a cependant pas été retenue dans la mesure où la longueur du cube, déterminante pour créer les liaisons, n'a pas été implémentée en vue d'un calibrage automatique.

L'approche retenue ne doit donc pas comprendre de paramètres humains (du moins dans cette étape) dans un but de simplification, d'automatisation et de déploiement de la classification de véhicules sur d'autres garanties et d'autres portefeuilles.

La triangulation de Delaunay permet de satisfaire cette condition. C'est une méthode commune de discrétisation spatiale d'un milieu continu, ou maillage, dont de nombreux algorithmes ont déjà été codés et sont disponibles sur internet pour différents environnements informatiques [1].

Il conviendra d'exposer ci-dessous les définitions proposées par PILLAUD [2006] [16] et les quelques propriétés démontrées par BREVILLIERS [2008] [3] de la triangulation de Delaunay, ainsi qu'un exemple visuel permettant de bien comprendre ce procédé.

Soit m un entier et E l'espace euclidien de dimension m. La distance euclienne sur E est notée d. On considère un ensemble S de points de E appelés sites, de cardinal n, dont on note $\wp(S)$ l'ensemble des parties. On définit la fonction $\varphi_S : E \mapsto \wp(S)$ qui à un point $x \in E$ associe $\{y \in S | d(x,y) = min_{z \in S} d(x,z)\}$, c'est-à-dire l'ensemble de ses plus proches voisins dans S.

Définition (Position générale)**.** *Les sites sont en position générale en dimension m s'il n'y a pas $m+1$ points dans le même hyperplan et $m+2$ points sur la même hypersphère.*

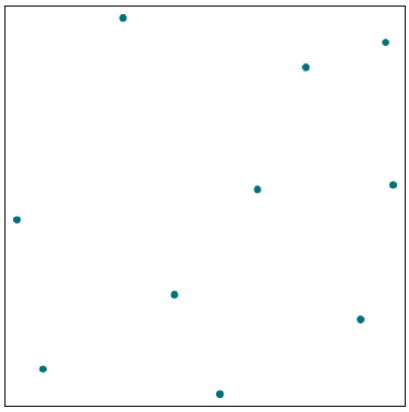
La position générale en dimension 2 est telle qu'il n'y a pas trois points alignés ou quatre points cocycliques, tels les sommets d'un rectangle.

Définition (Triangulation de Delaunay)**.** *Pour toute partie R de S telle qu'il existe une boule circonscrite à R qui ne contient aucun site de S autre que les sites de R, on définit la face de Delaunay $Del_S(R)$ comme l'enveloppe convexe de R. On appelle triangulation de Delaunay de S l'ensemble $Del(S)$ des faces de Delaunay de S.*

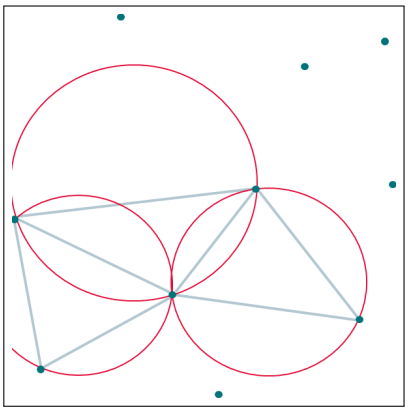
Propriété. *Tout ensemble S d'au moins trois sites en position générale admet une et une seule triangulation de Delaunay.*

En prenant l'hypothèse que le nuage de points résultant de l'analyse factorielle de données mixtes satisfait la condition de position générale, l'unicité de la Triangulation de Delaunay permet d'obtenir une unique relation de voisinage.

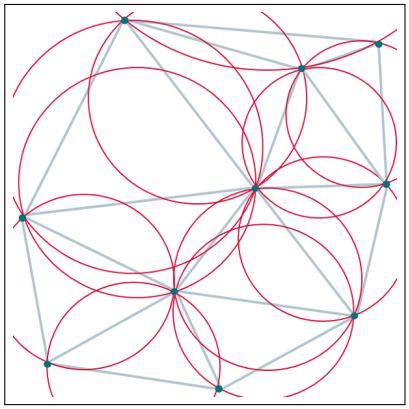
1. Cette partie étant développée sous R, le package geometry propose une utilisation directe de la triangulation de Delaunay.



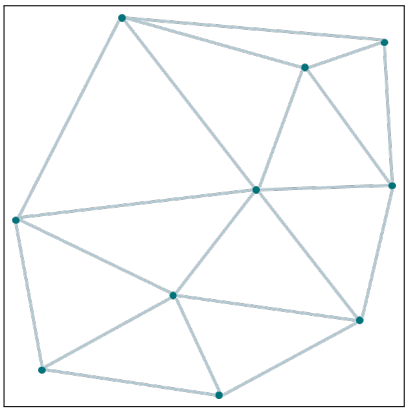
(a) Sites de S



(b) Cercles circonscrit satisfaisant les conditions de Delaunay



(c) Totalité des cercles circonscrits satisfaisant les conditions de Delaunay



(d) Résultat de la triangulation de Delaunay

FIGURE 5.3 – Exemple de création de la Triangulation de Delaunay

SIBSON (cf. BREVILLIERS [2008] [3]) démontre également que les triangulations de Delaunay maximisent le minimum des angles des triangles. En d'autres termes, elles évitent les simplexes (triangles en dimension > 2) allongés, ce qui correspond à notre but de définir les plus proches voisins du nuage de points de l'analyse factorielle de données mixtes.

Pour plus de clarté, un exemple visuel en dimension 2 est proposé en figure 5.3. Soit S un ensemble de 10 points du plan (Fig. 5.3a). Afin de satisfaire la définition de la triangulation de Delaunay, il faut que 3 sites soient sur le même cercle circonscrit et que ce cercle ne contienne aucun autre site de S. On peut donc commencer à construite les cercles (Fig. 5.3b). Chaque triangle permet de définir la relation de voisinage entre les points, et ce comme plus proche voisin. Une fois la totalité des cercles tracés (Fig. 5.3c), la triangulation de Delaunay apparaît. Il apparaît que le nombre minimal de voisins est 3 (pour la plupart des sites situés sur l'enveloppe convexe). Le nombre maximal de voisins est de 6 pour le site le plus central.

La triangulation de Delaunay apparaît donc comme la solution à notre problème de définition des plus proches voisins. Cependant, comme le précise sa définition, la triangulation de Delaunay créé une enveloppe convexe autour des sites de S. Ainsi, les sites fortement éloignés et n'étant apparemment pas voisins sont rapprochés par cette méthode (Fig. 5.4). Il conviendra alors de retraiter ces "faux voisins" par une méthode décrite dans le chapitre Application et Résultats, nous

permettant de créer des "îles" de véhicules de façon similaire aux îles géographiques.

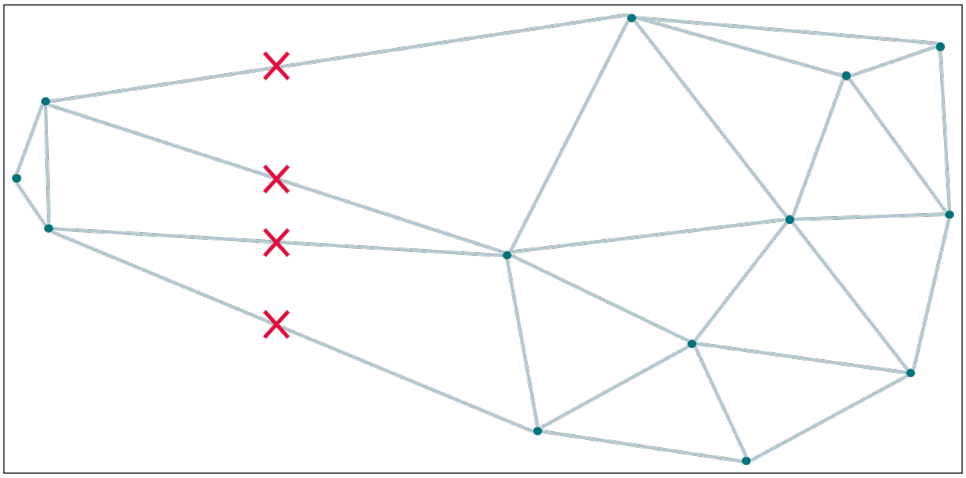


FIGURE 5.4 – Retraitement du caractère convexe de la Triangulation de Delaunay

3 Techniques de lissage

Les deux sections précédentes donnent quelques éléments théoriques afin d'obtenir une représentation spatiale des véhicules ainsi qu'une relation de voisinage entre eux. Transposée dans le cadre de la création d'une classification géographique, cette étude préalable n'est pas nécessaire dans la mesure où des cartes géographiques sont déjà disponibles. Ces dernières peuvent être plus ou moins granulaires : région, département, code postal, code INSEE, code IRIS, rue. L'objectif du zonier est de classer ces unités géographiques en fonction de leur niveau de risque, déterminé en fonction de la sinistralité observée et de la sinistralité estimée, soit par ses résidus. Pour rappel, les résidus sont calculés à l'aide de la formule suivante :

$$Residu = \frac{Estime}{Observe}$$

Il est possible que certaines communes n'aient pas suffisamment d'exposition pour que la quantité à classer soit robuste. Par exemple, il n'y a que très peu de personnes vivant dans cette commune et peu voire aucune assurée chez AXA. Il est donc nécessaire de procéder à un lissage en fonction du territoire, i.e. le risque d'une commune dépend du risque des communes qui l'entourent ou qui lui sont adjacentes. Il s'agit en d'autres termes de crédibiliser le risque d'une commune en fonction des autres. Le lissage peut se faire par voisinage ou par distance et permet ainsi d'éviter les grands sauts de risque entre communes adjacentes. Il est en effet difficile d'expliquer que le risque se divise par deux en fonction du côté de rue sur lequel on se trouve.

Cette méthodologie est exactement la même dans le cas d'une classification de véhicule, sauf que la carte géographique en question correspond au nuage de points. Les deux types de lissage, par voisinage et par distance, peuvent être appliqués.

Il conviendra de détailler la théorie de ces deux techniques ci-dessous. L'outil utilisé pour les implémenter est le logiciel Classifier fourni par Towers Watson.

Lissage par distance

Le lissage basé sur la distance permet de lisser les paramètres d'une unité géographique en attribuant plus de poids à l'information contenue dans les unités géographiques proches, proportionnellement à leur distance de celle étudiée. Les unités les plus proches ont donc un poids plus important.

Il est nécessaire de déterminer une distance pour effectuer ce lissage. Il est d'usage d'utiliser la distance euclidienne donnée par la formule suivante : $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$, où (x_i, y_i, z_i) et (x_j, y_j, z_j) sont les coordonnées de deux unités géographiques (e.g. centroïdes de communes INSEE en deux dimensions ou représentations spatiales de véhicule).

Soient :
- r_i le résidu non lissé de l'unité géographique i,
- w_i l'exposition totale de l'unité géographique i,
- w_0 l'ajustement de crédibilité,
- p la puissance de crédibilité. Cette valeur permet d'ajuster l'effet de la crédibilité entre les zones à forte et à petite exposition,
- $\overline{r_i}$ la distance moyenne pondérée des résidus non lissés, donnée par la formule suivante :

$$\overline{r_i} = \frac{\sum_{j \neq i} r_j w_j e^{-d_{ij} p}}{\sum_{j \neq i} w_j e^{-d_{ij} p}}$$

Le résidu lissé est donné par la formule :

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

avec Z_i le facteur de crédibilité dont la formule est :

$$Z_i = \left(\frac{w_i}{w_i + w_0}\right)^p$$

Ce lissage est basé sur la moyenne pondérée des expériences de l'unité géographique et des zones environnantes, avec un poids variant selon la distance par rapport à l'unité étudiée. Le lissage basé sur la distance intègre donc l'information des unités géographiques en fonction de la distance entre elles : plus la distance sera élevée, moins son expérience aura de poids.

Cette technique est habituellement utilisée en lissage géographique afin d'étudier des risques liés à la météorologie, où l'information de frontières géographiques telles une aire urbaine et une aire rurale, ou le passage d'un département à un autre, n'a pas d'importance. Pour plus d'information, le lecteur est invité à se référer à BRUBAKER [1996] [4], CHRISTOPHERSON et WERLAND [1996] [5] et à ANDERSON et *al.* [2007] [2].

Lissage par voisinage

Le lissage basé sur le voisinage intègre l'information du voisinage aux unités géographiques étudiées. Chaque unité dépend de ses voisins, lesquels sont également influencés par leurs voisins directs. Cette méthode se base sur la théorie bayésienne, supposant que le risque de chaque unité se comporte de façon similaire à celui de ses voisins. Le nombre de sinistres d'une unité géographique i suit une loi de Poisson N_i telle que

$$N_i \sim \mathcal{P}(E.e^{m+b_i})$$

avec
— E l'exposition,
— m le risque de base issu du modèle linéaire généralisé,
— b_i l'effet géographique de l'unité i lui même défini par

$$b_i \sim \mathcal{N}(\overline{b_j}, \sigma^2)$$

où
— $\overline{b_j}$ est la moyenne des b_j avec j les unités voisines de i,
— σ^2 le niveau de lissage

Pour plus de détails sur cette méthode, le lecteur peut se référer à BOSKOV et VERRALL [1993] [23].

4 Méthodes de classification

La classification est une méthode d'analyse de données visant à regrouper un ensemble d'individus en groupes homogènes, de façon à ce que chaque individu d'un groupe partage des caractéristiques communes. Il existe une multitude de méthodes de classifications. Il conviendra de n'en exposer que trois, celles utilisées par la suite dans ce document.

Soit I un ensemble d'individus et $i \in I$ un individu. Soit w_i l'exposition ou le poids de cet individu.

Méthode des quantiles
Cette méthode consiste à créer des groupes G_k, k étant le nombre de groupes souhaités, de telle sorte à avoir le même nombre d'individus i par groupe. Dans le cas d'une classification du risque, elle permet de différencier fortement les premières classes des dernières. Cependant elle peut créer des groupes dont les intervalles de risque sont très larges.

Méthode des poids égaux
Cette méthode consiste à créer des groupes G_k, k étant le nombre de groupes souhaités, de telle sorte à avoir le même poids $w = \sum_i w_i$ par groupe. Cette méthode permet d'avoir des classes plus robustes lorsque la classification est intégrée dans un GLM. Cependant la différenciation entre les premières et les dernières classes seront limitées.

Méthode de Ward
La méthode de Ward est une technique de classification ascendante hiérarchique. Le principe est de fournir un ensemble de groupes de plus en plus gros obtenus par regroupements successifs. Les groupes sont déterminés de manière à maximiser l'homogénéité intra-groupe et à maximiser l'hétérogénéité intergroupe. Une hypothèse sur une mesure de dissimilarité entre les individus est alors nécessaire. Dans le cas de points situés dans un espace euclidien, il est par exemple possible d'utiliser la distance comme mesure de dissimilarité. Dans une même classe, les individus seront les plus semblables possibles tandis qu'ils seront le plus dissemblables possibles dans des classes différentes.

5 Processus global

La figure 5.5 résume graphiquement le processus global de création du véhiculier. Tout d'abord, la création de la base de données ainsi que tous les traitements nécessaires sont faits sur le logiciel de SAS. Au terme de cette étape, deux types de sorties sont générées : les fichiers nécessaires à la création des modèles GLM sur le logiciel Emblem développé par Towers Watson, et la base de données véhicules présents dans le portefeuille de Direct Assurance. C'est sur cette dernière qu'une Analyse Factorielle de Données Mixtes est effectuée avec le logiciel R. A la suite de l'AFDM, la triangulation de Delaunay est appliquée sur le nuage de points obtenu, ainsi que les retraitements nécessaires, toujours dans le logiciel R. Le nuage de points et les relations de voisinage servent à construire une table de contiguïté sur R, dont le format respecte celui nécessaire pour le logiciel Classifier, également développé par Towers Watson.

Les entrées du logiciel Classifier sont les fichiers issus d'Emblem et une table de contiguïté. L'étude sur les résidus, le lissage spatial ainsi que la classification des résidus lissés s'effectuent à l'aide du logiciel Classifier. Le fichier de sortie est une base comportant chaque véhicule et la classification correspondante. Cette nouvelle variable est alors fusionnée avec la base de donnée initiale, et elle pourra être introduite et testée dans le GLM à l'aide du logiciel Emblem.

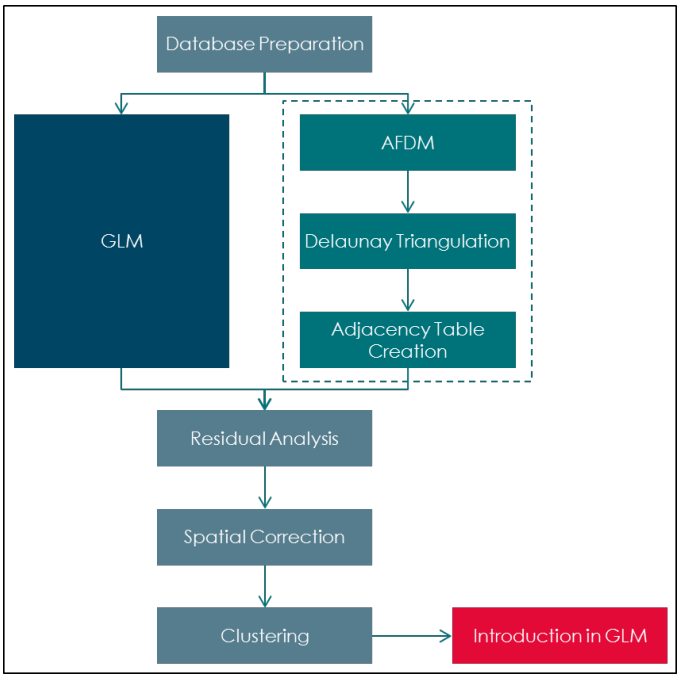


FIGURE 5.5 – Méthodologie de la classification de véhicules

Troisième partie

Application et résultats

Chapitre 6

Application de la Modélisation

Dans ce chapitre, nous allons nous concentrer sur les deux premiers objectifs de ce mémoire : la revue du modèle de la garantie responsabilité civile matérielle ainsi que la segmentation des périls entre sinistres conventionnés et sinistres au coût réel. Il conviendra de déterminer le périmètre de l'étude et d'exposer les indicateurs décrits dans la section 2.1. Nous nous focaliserons sur la description du modèle de fréquence des sinistres IRSA et sa validation (ceux des modèles de fréquence et de coût moyen ayant subi les mêmes tests, nous évitons ainsi la redondance des résultats). Enfin, nous vérifierons que la mise à jour et la segmentation des périls permet d'améliorer la prédiction de la prime pure de cette garantie.

1 Détermination du périmètre de l'étude

La sinistralité est généralement étudiée par année de survenance de sinistre, ce qui permet de ne pas biaiser les analyses avec certains problèmes de saisonnalité. En effet, on remarque par exemple que pour la garantie Dommages Tous Accidents, visant à indemniser le véhicule de l'assuré lorsqu'il a un sinistre responsable ou un accident n'impliquant que son véhicule (perte de contrôle dûe à la neige ou à un défaut de maîtrise, accident de parking, etc.), un pic de sinistralité revient chaque année à l'époque des vacances d'été.

Les indicateurs assurantiels se stabilisent en fonction du recul par rapport au 1er janvier de l'année de survenance. Cependant la fréquence au 31 décembre de l'année est, dans la majeure partie des garanties, loin de représenter la fréquence ultime de la survenance. Certaines garanties se stabilisent au deuxième mois de l'année N+1 (deux mois après la fin de l'année de survenance) comme la garantie bris de glace pour laquelle le nombre de sinistres tardifs, i.e. survenus l'année N mais déclarés les années ultérieures, est minime. D'autres comme les garanties corporelles prennent des années avant de se stabiliser (environ 10 ans). En effet, la volatilité de ces garanties s'explique par leur faible volume de sinistres ainsi que par l'aléa de l'état de santé de certaines victimes.

Comme discuté dans la section 2.3 Modélisation de la Responsabilité Civile Matérielle, nous avons décidé de segmenter les périls afin d'avoir une meilleure prédiction des deux types de sinistres : ceux suivant le processus IRSA et ceux remboursés au coût réel. Il conviendra dans cette étude de ne se concentrer que sur les sinistres IRSA afin d'éviter la redondance, l'étude sur l'autre type de sinistre étant fortement similaire.

Le déroulé de la fréquence Responsabilité Civile matérielle est tracé en fig. 6.1. Le point de la courbe 2013 situé à l'abscisse 11 (i.e. 30 novembre 2016) correspond à la fréquence Year To Date (YTD) de novembre de cette survenance. C'est le nombre de sinistres de survenance 2013, déclarés

entre le 1 janvier et le 30 novembre 2013, divisé par le nombre d'exposition de la même période. A partir de décembre (abscisse 12), il semble y avoir une hausse sensible de la fréquence correspondant aux sinistres tardifs : tandis que le nombre de sinistres augmente, l'exposition ne varie plus, l'année étant écoulée.

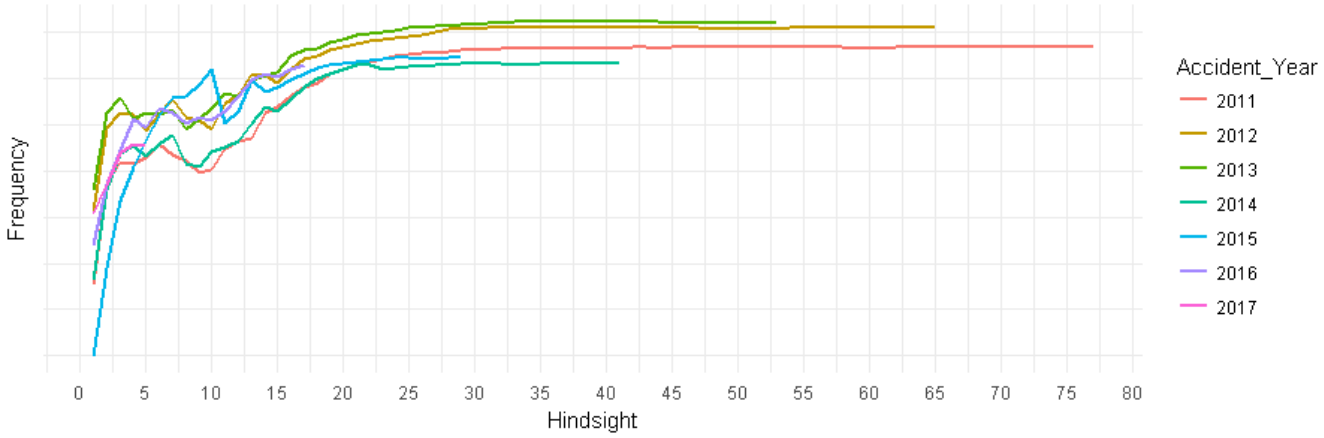


FIGURE 6.1 – Déroulé de la fréquence YTD RCM IRSA par année de survenance

On retrouve la même structure de la fréquence sur les années précédentes. Ceci dénote qu'il n'y a pas eu de changement majeur dans la gestion des sinistres ou dans le fonctionnement des équipes opérationnelles. On peut donc se baser sur les fréquences historiques homogènes afin de projeter la fréquence future.

La fréquence est quasi stabilisée avec un recul de 18 mois, notamment pour les années 2014, 2015 et 2016. De plus, le taux de clôture à fin juin 2016, i.e. le nombre de sinistres clôturés divisé par le nombre de sinistres total, est de 93,4% en moyenne sur ces trois années, soit la quasi-totalité des sinistres. Ceci est visible sur la fig. 6.2. Il conviendra alors d'intégrer ces trois années de survenance dans notre modèle de fréquence afin d'avoir suffisamment de volume ainsi que d'avoir trois années nous permettant de vérifier le caractère de robustesse du modèle évoqué dans la section 2.2 (Modèles Linéaires Généralisés).

Même si il n'y aura pas de modélisation du coût moyen, ce dernier étant déterministe, nous pouvons analyser le déroulé des coûts moyens hors sinistres sans suites sur la fig. 6.3. On peut voir qu'il varie en fonction du recul, alors qu'il est censé être constant et égal aux forfaits IRSA. De plus il semble être moins élevé que le forfait. Prenons par exemple l'année 2015. Le coût moyen varie de 1200€ à 1250€, alors que le forfait est égal à 1308€. Cet effet n'est qu'un reflet de la proportion du taux de responsabilité dans le portefeuille (responsabilité partielle vs responsabilité totale). Le coût moyen croît et se stabilise autour de 15 mois. Le taux de mi forfait est stable en fonction de l'année de survenance et oscille entre 19% et 20% du total des sinistres IRSA. La proportion des sinistres IRSA elle-même est très stable sur l'ensemble des sinistres RCM et oscille entre 62% et 63% des sinistres. L'évolution à la hausse du coût moyen annuelle est entièrement portée par l'évolution des forfaits IRSA sensés représenter l'inflation, l'augmentation du prix des pièces de réparations ou du coût de la main d'œuvre.

A noter que pour la modélisation du coût moyen des sinistres de droit commun ou ne rentrant pas dans le giron des conventions IRSA (dépassement de montant, assurance ne faisant pas partie de la convention), il convient d'écrêter les sinistres atypiques, i.e. de répartir la charge au-dessus d'un seuil

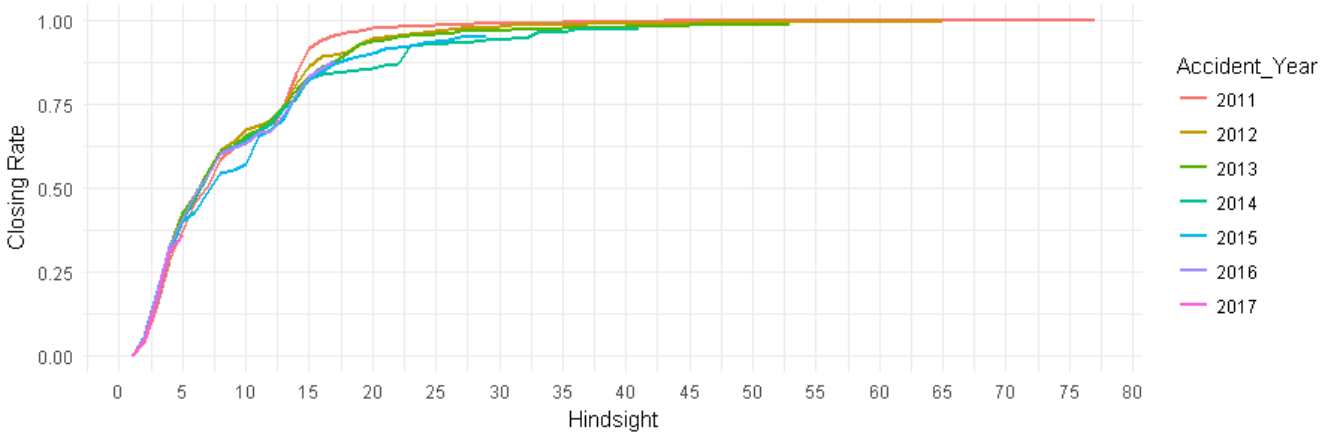


FIGURE 6.2 – Déroulé du taux de clôture YTD RCM IRSA par année de survenance

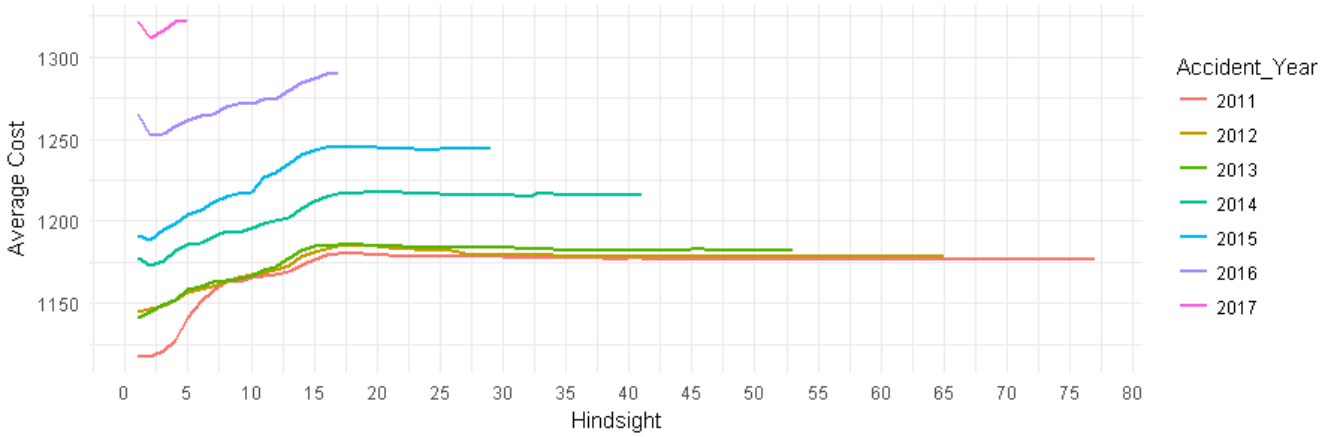


FIGURE 6.3 – Déroulé du coût moyen YTD RCM IRSA par année de survenance

déterminé par les actuaires sur tout ou une partie du portefeuille afin de ne pas pénaliser tout un segment lors de la modélisation. Un exemple serait le cas d'un assuré de 34 ans dont le sinistre aurait entraîné l'incendie dans le parking d'un centre commercial. Le sinistre pourrait atteindre plusieurs millions dû aux réparations du parking et aux pertes économiques subies par le centre commercial. Si cette charge était laissée comme telle lors de la modélisation du coût moyen, le coefficient de la tranche d'âge 30-35 ans serait bien supérieur aux coefficients des autres tranches, d'où l'intérêt d'un écrêtement. Il est toutefois intéressant de noter que si 96% des sinistres hors forfait IRSA ont une charge inférieure à 10000€, les 4% restant représentent 28% de la charge totale. Le seuil d'écrêtement de 10000€ est alors arbitrairement choisi. Cependant, plusieurs méthodes de choix de seuil d'écrêtement peuvent être utilisées, telles que la méthode du Hill Plot ou la Mean Excess Function.

Il est également possible de voir sur le graphique fig. A.2 en Annexes représentant la distribution des coûts moyens non IRSA par année de survenance de sinistres, vision au 30 juin 2017, la sur-représentation des sinistres aux montants 1500€ et 2120€ pour les années 2016 et 2017. Il s'agit des forfaits d'ouvertures des sinistres de droit commun et des carambolages (sinistres impliquant plus que deux véhicules). Ces sinistres sont habituellement plus longs à se clôturer que les sinistres

IRSA de par les lenteurs administratives lorsqu'il s'agit de sinistres avec les propriétés dommage au bien (poteau, grille de maison) qui nécessitent qu'un PV soit dressé, ou bien les carambolages pour lesquels les responsabilités ne sont déterminées qu'à l'issue d'un procès-verbal des forces de l'ordre.

Suite à l'étude ci-dessus, il est possible de travailler sur les années de survenances jusqu'à 2016 inclus, ces années étant quasi-stabilisées. Il conviendra d'utiliser une période de 3 ans pour la modélisation de la fréquence des sinistres IRSA afin d'avoir la possibilité d'effectuer des tests de robustesse et de conserver suffisamment de volume. La période choisie est du 1er janvier 2014 au 31 décembre 2016, c'est-à-dire tous les contrats présents et couverts sur cette période ainsi que leur sinistralité vue au 30 juin 2017. Ceci permettra de conserver un mix de portefeuille suffisamment récent pour qu'il soit similaire à celui des années futures. Le déroulé des sinistres ne relevant pas des conventions IRSA étant plus long, il conviendra de travailler sur une période allant du 1er janvier 2013 au 31 décembre 2015 pour la modélisation de la fréquence et du coût moyen (cf. fig. A.3 et fig. A.4 en Annexes).

2 Statistiques descriptives

Un point sur les indicateurs statistiques définis dans la section 2.1 peut se révéler utile. Il est intéressant de constater la forte croissance du portefeuille de Direct Assurance France (+36% entre 2013 et 2016), ce qui montre que la part de l'assurance directe gagne peu à peu des parts de marché au sein de l'assurance automobile.

Les indicateurs de sinistralité restent assez stables dans le temps, comme la fréquence entre 4% et 5%, ou le taux de sans suites autour de 17%. Quelques écarts sont néanmoins à noter pour les années récentes (2015 et 2016), la sinistralité n'étant toujours pas complètement stabilisée. Ces écarts sont tout de même minimes (moins de deux points d'écart).

La fréquence est également stable une fois le type de sinistre déterminé : la fréquence des sinistres ne relevant pas de la convention IRSA est constante par année, autour de 1%, tandis que les sinistres IRSA représentent une fréquence d'environ 3%. Cette clé de répartition est également stable dans le temps, avec environ 78% des sinistres hors sans suites relevant de la convention.

Le seul indicateur ayant des variations est le coût moyen. Pour les sinistres IRSA, il s'agit de l'évolution des forfaits comme décrit dans la section précédente. L'évolution du coût moyen des sinistres non IRSA dépend quant à elle de l'inflation et des sinistres graves. Elle est plus sujette à de la volatilité que les autres sinistres.

3 Corrélation des variables

Avant de commencer la modélisation, il est utile de s'attarder sur la corrélation entre les variables. Ceci est possible à l'aide du V de Cramer permettant de mesurer l'intensité du lien entre deux variables nominales et basé sur la statistique du test du χ^2. Soit un tableau de contingence des variables A et B, de taille n, avec n_{ij} le nombre de fois où les valeurs (A_i, B_j) ont été observées. $i \in [1, r]$, $j \in [1, s]$, avec r le nombre de lignes et s le nombre de colonnes. Le V de Cramer, défini ci-dessous, varie entre 0 (liaison nulle, i.e. les deux variables étudiées sont indépendantes) et 1 (liaison maximale, i.e. dépendance complète).

$$V = \sqrt{\frac{\phi^2}{min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}}$$

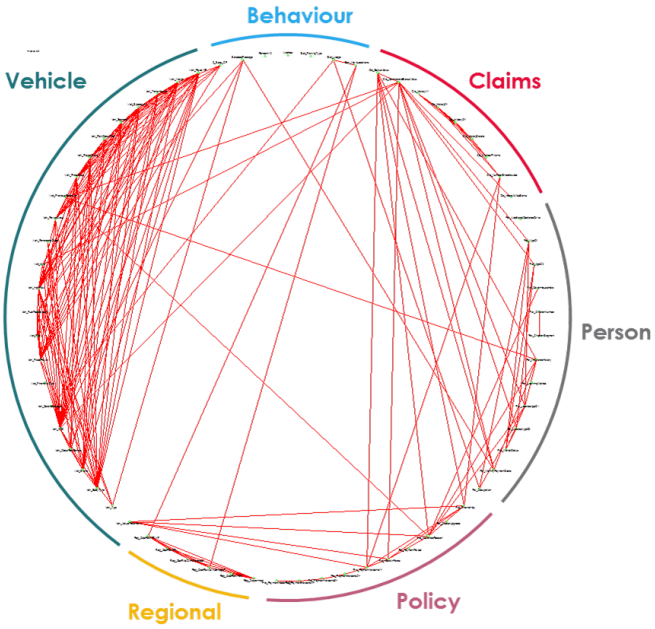


FIGURE 6.4 – Graphe des corrélations

La Figure 6.4 permet de visualiser directement l'intensité des corrélations entre variables de façon très pratique. Plus l'intensité est grande, plus la ligne est épaisse. Il est clair que les corrélations entre les variables régionales sont très fortes, et qu'elles sont totalement dé-corrélées des autres types de variable : ceci est un résultat attendu. En effet, prenons le cas de la variable moyenne du nombre de jours de pluie par an et l'âge du conducteur secondaire. Cela ne fait pas de sens de penser que seuls les jeunes habitent dans des régions pluvieuses et les personnes âgées au soleil. On peut remarquer également que les variables véhicules sont suffisamment dé-corrélées des autres variables, sauf les variables âge du véhicule, durée de détention du véhicule actuel et précédent. Elles ne sont pas tout à fait des variables véhicules mais plutôt des variables comportementales, et ne sont pas liées aux caractéristiques intrinsèques du véhicule, comme la puissance réelle ou le prix de la voiture. Le risque lié aux variables véhicule est donc indépendant des variables comportementales ou autre et il y a un potentiel signal sous-jacent qui ne pourra pas être capté sans quelques caractéristiques véhicules.

4 Comparaison des modèles

La revue de la garantie Responsabilité Civile Matérielle, et le choix de segmenter la garantie en deux périls (forfaits conventionnels et sinistres au coût réel) conduisent à la création de trois modèles, deux de fréquence et un de coût moyen. Les sinistres conventionnels ayant des coûts fixes, il n'est pas nécessaire de faire une modélisation du coût moyen. La modélisation se fait à l'aide du logiciel Emblem fourni par Towers Watson.

Principales variables sélectionnées et supprimées

Pour des raisons de confidentialité, les variables sélectionnées ne seront décrites que par catégorie de variable. L'ancien modèle comportait 26 variables. Le nouveau modèle IRSA (fréquence) ne contient plus que 24 variables, tandis que le modèle non IRSA (fréquence x coût moyen) en contient 32.

Catégorie	Ancien modèle	Modèle IRSA	Modèle Non IRSA
Comportementales	1	0	2
Sinistralité	4	4	6
Personne	7	9	11
Police	3	3	5
Géographique	1	1	1
Vehicule	10	7	7
Total	26	24	32

TABLE 6.1 – Catégories de variables sélectionnées dans l'ancien et les nouveaux modèles

Les variables conservées dans les deux modèles sont des variable usuelles, telles l'âge du conducteur principal, la carrosserie du véhicule ou le segment du véhicule. Certaines variables ont été rajoutées, notamment des variables de sinistralité telles le nombre de sinistres durant la vie du contrat. Enfin, des variables ont été supprimées selon deux critères. Pour certaines l'effet était déjà capté par d'autres variables, et la valeur ajoutée de la variable en question était infime. C'est le cas pour plusieurs variables véhicule. Pour d'autres, il s'agissait de variables déclaratives, comme la durée de détention du véhicule précédent : cette variable est déclarative dans le sens où il n'y a pas de vérification de la part de l'assureur de son exactitude. L'assuré peut donc mentir sur sa réponse et bénéficier d'un tarif plus avantageux. De plus, la plupart du temps l'assuré renseigne la variable de façon aléatoire ou arrondie (0 ans, 5 ans, 10 ans) en fonction de la valeur par défaut du questionnaire.

Les graphiques ci-dessous correspondent au modèle de fréquence des sinistres conventionnés (sur lequel la classification des véhicules sera faite dans la section suivante). Les éléments de validation de la fréquence et du coût moyen au coût réel sont toutefois présentés en Annexes (fig. A.5 à fig. A.8).

L'histogramme jaune correspond à l'exposition par modalité de la variable sélectionnée (par exemple, l'âge du conducteur principal dans la Figure 6.5). La mesure est en pourcentage du total et se lit sur les ordonnées à droite. Dans l'exemple de l'âge du conducteur principal, la part des polices dont le conducteur a 38 ans est d'environ 4% et représente la catégorie la plus importante du portefeuille.

Les autres courbes se lisent avec l'ordonnée de gauche. La courbe rose avec les marques carrées représente les valeurs observées pour chacune des catégories. La courbe verte fluo avec les marques triangulaires représente les relativités de chaque modalité, et la courbe verte foncée avec les marques triangulaires correspond aux valeurs prédites. Ces trois indicateurs sont remise à l'échelle par rapport à la modalité la plus importante ("rescaled").

La modalité la plus représentée a une relativité à 1. Les coefficients sont alors positifs ou négatifs par rapport à cette valeur en fonction de leur risque : les 18 ans représentent un risque plus important

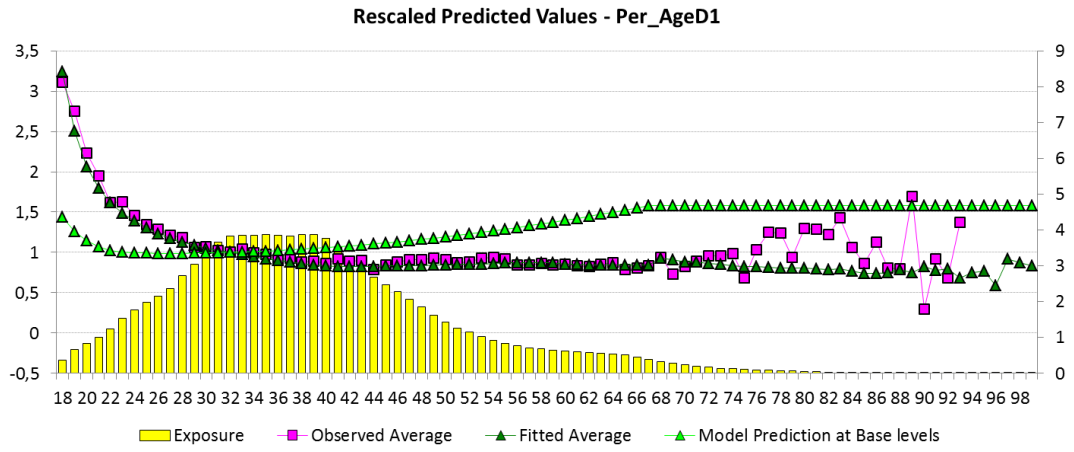


FIGURE 6.5 – Résultats GLM sur l'Age du Conducteur Principal

"toutes choses égales par ailleurs" que les 38 ans et leur coefficient est de 1,5. Il se peut que la relativité d'une modalité soit inférieure à une autre alors que l'observé est inverse. C'est ici le cas pour les âges supérieurs à 68 ans par rapport aux 18-20 ans. Ceci est possible dans la mesure où le risque des 18-20 ans est déjà capté par une autre variable : l'ancienneté de contrat ou le type de voiture par exemple. D'où l'important de la notion "toutes choses égales par ailleurs".

"Toutes choses égales par ailleurs", les personnes de plus de 68 ans sont plus risquées que les 18-20 ans. Ce n'est bien évidemment pas le cas en observé au vu de la courbe des observés. En effet, la sinistralité des plus jeunes est bien plus importante que celle des autres. Ce résultat est attendu dans la mesure où leur expérience de conduite est moindre.

La sinistralité baisse en fonction de l'ancienneté de contrat, visible sur la figure 6.6. Les relativités conservent cette tendance. Ceci veut dire qu'un assuré ayant 5 ans d'ancienneté de contrat aura un risque moins élevé qu'un autre assuré ayant exactement les mêmes caractéristiques, mais n'ayant que 2 ans d'ancienneté.

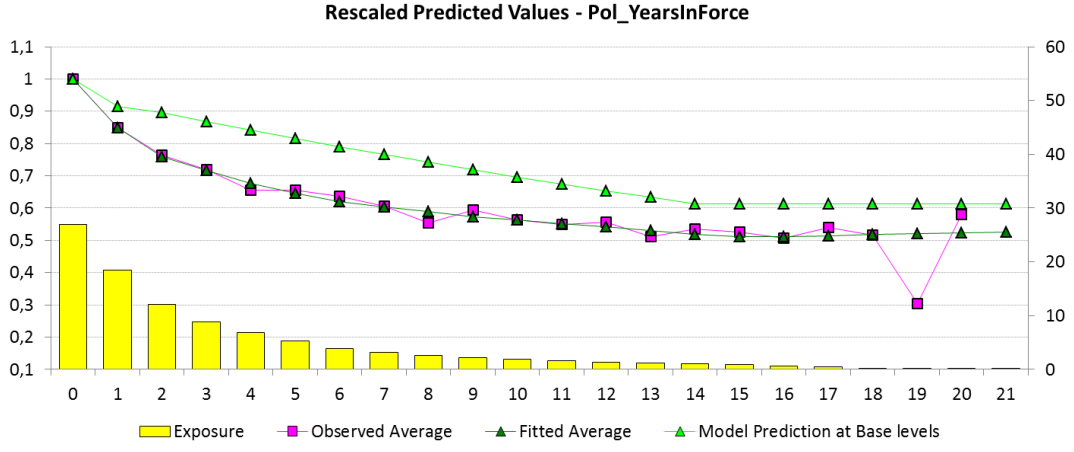


FIGURE 6.6 – Résultats GLM sur l'ancienneté de contrat

Le modèle final étant créé, il convient de valider la robustesse du modèle.

Validation du modèle

Distribution des résidus

Une fois le modèle créé, la validation du modèle peut être effectuée à l'aide de tests dans le but d'identifier les potentiels problèmes, dont la détection et la correction renforceront la robustesse du modèle.

Le bon choix de la fonction de lien peut être vérifié en traçant la dispersion des résidus de l'échantillon de modélisation en fonction des valeurs prédites. Si le modèle choisi est approprié, la déviance standardisée devrait suivre une distribution normale centrée réduite. Les résidus ne doivent donc pas présenter de structure particulière, mais distribuées de façon aléatoire autour de l'axe des abscisses.

Pour les modèles de fréquences, beaucoup de résidus sont égaux à zéro au vu de la variable réponse. Un graphique de résidus typiques montrerait une tendance ni centrée autour de l'axe des abscisses, ni symétrique. Une conclusion rapide serait de considérer la fonction de lien ou la structure des résidus incorrectes. Cependant les résidus sont calculés par individu et le graphique aussi granulaire n'apporte pas d'information utile.

Dans ces circonstances, il est plus intéressant de tracer ce qu'on appelle les "crunched residuals" dans le logiciel Emblem. Il s'agit des résidus calculés par classes et non pas par individus. Ce graphique est visible sur la figure 6.7. Nous pouvons voir que les résidus sont bien disposés de façon aléatoire autour de 0, peu importe la valeur des prédictions.

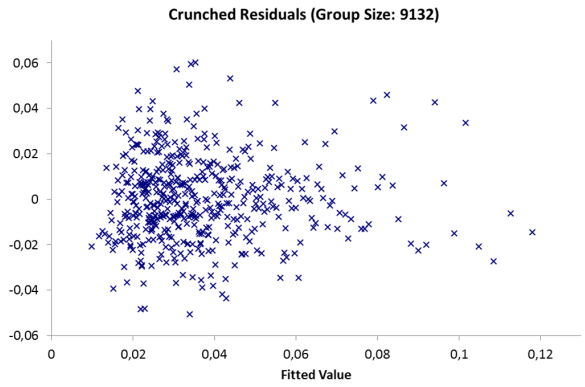


FIGURE 6.7 – Crunched Residuals

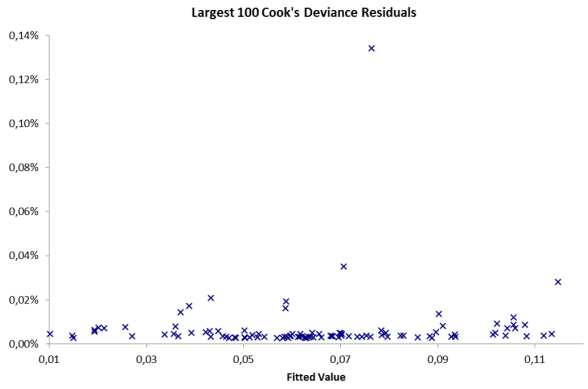


FIGURE 6.8 – 100 plus grandes valeurs de la Déviance de Cook

Déviance de Cook

Il est également intéressant d'étudier les plus grandes valeurs de la déviance de Cook dans la mesure où elles permettent d'identifier des observations particulières ayant une influence excessive sur le modèle. Elles sont la résultante d'une valeur extrême de la réponse ou de la prédiction, soit des contrats avec une fréquence très élevée. Ainsi le modèle est trop dépendant de ces observations au détriment d'autres plus nombreuses. Une déviance de Cook supérieure à 2% est généralement considérée comme trop élevée au vu de la faible probabilité d'observer ces individus. La solution est de retirer ce genre d'observations afin de conserver un bon caractère prédictif pour la majorité du modèle.

Nous pouvons voir sur la figure 6.8 qu'il y a un point potentiellement aberrant. Cependant la valeur est moins élevée que les 2% préconisés par AXA. Nous pouvons alors considérer que le modèle est valide.

Courbe d'ajustement

Une courbe d'ajustement est une mesure permettant de comparer les valeurs prédites aux valeurs observées par groupe de valeurs prédites croissantes, sur l'échantillon de modélisation et sur celui de validation. Nous pouvons ainsi évaluer le caractère prédictif du modèle. Ces graphiques sont disponibles sur les figures 6.9 et 6.10. La courbe bleue représente la moyenne des valeurs prédites par classe et la moyenne des valeurs observées est en rouge. Les courbes sont quasiment superposées et leur comportement est le même sur les deux graphiques. Le modèle ne sur-apprend donc pas et son ajustement est correct.

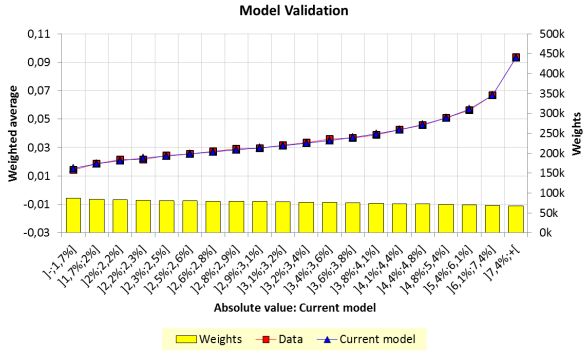


FIGURE 6.9 – Courbe d'ajustement sur l'échantillon de modélisation

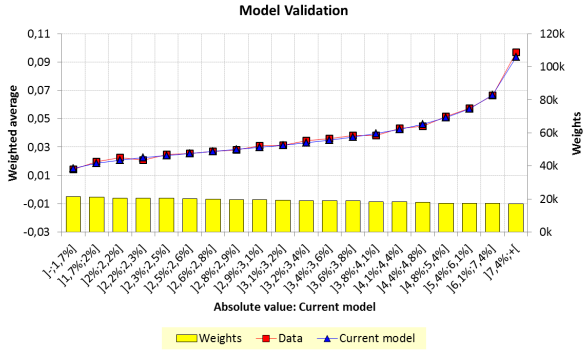


FIGURE 6.10 – Courbe d'ajustement sur l'échantillon de validation

Mesure de l'amélioration de la nouvelle modélisation

Afin de vérifier que le rafraîchissement des données a été bénéfique et que la segmentation des périls est justifiée, nous allons calculer et comparer les GINI entre les différentes modélisations de la prime pure (et non seulement la modélisation de la fréquence des sinistres IRSA). Pour rappel, l'ancienne prime pure était calculée de la façon suivante, avec une modélisation de la fréquence mais un coût moyen fixe, au vu de la distribution du coût des sinistres incluant des forfaits conventionnels.

$$\text{Ancienne Prime Pure} = \text{Fréquence} * \text{Coût Moyen}$$

La formule de la nouvelle prime pure est la suivante, avec deux modélisations pour les fréquences, une pour les coûts moyens non IRSA et un coût moyen déterministe pour le coût moyen IRSA.

$$\text{Nouvelle Prime Pure} = \text{Fréquence IRSA} * \text{CM IRSA} + \text{Fréqence Non IRSA} * \text{CM Non IRSA}$$

Modèle	Échantillon de Modélisation	Échantillon de Validation	Écart
Nouveau Modèle	26.98%	27.33%	1.28%
Ancien Modèle	18,16%	20,09%	10,62%
Amélioration	49%	36%	

TABLE 6.2 – Comparaison des coefficients de GINI entre les modèles

Nous pouvons voir dans le Tableau 6.2 que le rafraîchissement des données ainsi que la segmentation des périls permet d'améliorer considérablement le GINI, de l'ordre de 36% sur l'échantillon de validation. De la même manière, la revue du modèle réduit le sur-apprentissage présent sur l'ancien modèle. En effet, le coefficient de GINI calculé sur l'échantillon de validation est très proche de celui calculé sur l'échantillon de modélisation. Ceci n'est pas le cas pour l'ancien modèle, où la différence de GINI est de 10.62% entre les deux échantillons.

Chapitre 7

Application de la Nouvelle Méthodologie de Classification

Dans ce chapitre, la nouvelle méthodologie de classification des véhicules sera appliquée sur le portefeuille de Direct Assurance. Il conviendra de se concentrer sur les résultats de l'Analyse Factorielle de Données Mixtes, étude nécessitant beaucoup de graphiques afin de comprendre la structure du nuage de points qui servira par la suite. Puis la création de la table de contiguïté sera détaillée ainsi que son retraitement permettant de parer à l'effet convexe de la triangulation de Delaunay. Enfin, les différentes classifications testées seront exposées.

1 Résultats de l'AFDM

Traitements préalables

Comme précisé dans la section 1.4, l'un des objectifs de ce mémoire est d'appliquer la théorie des zoniers à la création d'une classification de véhicules. Pour ce faire il est nécessaire de représenter les véhicules dans l'espace. L'Analyse Factorielle de Données Mixtes sera utile pour cet exercice dans la mesure où seront utilisées à la fois des données quantitatives et des données qualitatives.

Avant d'appliquer cette méthodologie sur la base de données véhicules, laquelle a subi quelques modifications sur ses colonnes décrites dans la section 3.2, il est nécessaire de trier ses lignes. En effet, il convient de se concentrer sur des véhicules présents dans le portefeuille de Direct Assurance. Ainsi la classification obtenue sera propre à son portefeuille et ne sera polluée ni par des véhicules ne satisfaisant pas les règles de souscription de Direct Assurance décrites dans la section 1.3, ni par des véhicules qui ne sont plus présents sur le marché. Une fois ce filtre effectué, nous devons sélectionner les variables sur lesquelles nous allons appliquer l'AFDM. En effet, les variables telles la voie avant, la voie arrière ou l'empattement sont très corrélées à la largeur du véhicule et l'information contenue n'apporterait que plus de complexité à l'exercice. De même la puissance en Chevaux DIN n'a un intérêt que limité dans la mesure où la puissance réelle en KWH est retenue.

Une fois cette liste définie, nous allons créer des Codes Clé, chacun correspondant à une combinaison unique de caractéristiques de véhicules. Par exemple, il n'y aura qu'un code pour deux versions de véhicules pour lesquelles nous ne sommes pas intéressés de les distinguer. Choisissons deux véhicules de la base, codés AR22017 et AR22018. Ces codes correspondent à une Alpha Romeo – Alpha 159 – 1.9 JTD 110 et à une Alpha Romeo – Alpha 159 – 1.9 JTD 110 PACK dont la seule différence est le prix d'origine (mais elles appartiennent à la même classe de prix). Or nous n'avons

conservé que la classe de prix et non le prix d'origine dans la base de données. Nous avons donc créé un Code Clé propre à la combinaison de variables que nous avons conservées, regroupant alors ces deux véhicules. De cette façon, un modèle de véhicule ayant une myriade de versions n'aura pas un poids démesuré par rapport à un véhicule n'ayant que quelques modèles et plus représenté sur le marché.

Finalement, la base véhicule utilisée pour l'AFDM contient 45371 Codes Clé et 30 variables caractéristiques véhicule.

Tableau des valeurs propres

Le tableau des valeurs propres indique la variance expliquée par les axes. Nous avons 12 variables quantitatives et 17 variables qualitatives avec pour chacune un nombre de modalité indiqué dans la figure 7.2.

L'inertie totale est expliquée par 103 axes et est égale à $p = 108$. En effet, l'inertie totale est calculée de la façon suivante, pour n variables qualitatives :

$$\text{nombre de variables quantitatives} + \sum_{i=1}^{n} (\text{nombre de modalité de la variable qualitative } i - 1)$$

$$12 + [(44 - 1) + (7 - 1) + (16 - 1) + (6 - 1) + \ldots + (4 - 1) + (6 - 1)] = 108$$

Dans le cadre d'une analyse de données pure, nous aurions conservé 46 axes selon le critère de Kaiser, lequel consiste à ne conserver que les axes dont l'inertie est supérieure à 1. Karlis, Saporta et Spinakis ont proposé en 2003 un nouveau seuil défini selon la formule suivante :

$$seuil = 1 + 1.65\sqrt{\frac{p-1}{n-1}} = 1.080129$$

Où :

— n est le nombre d'observations = 45371

— p est la somme des valeurs propres = 108

Suivant ce dernier critère, il conviendrait de conserver 29 axes dans une analyse de données. La totalité de l'inertie expliquée en fonction du nombre de facteur est disponible en Annexe (fig A.11).

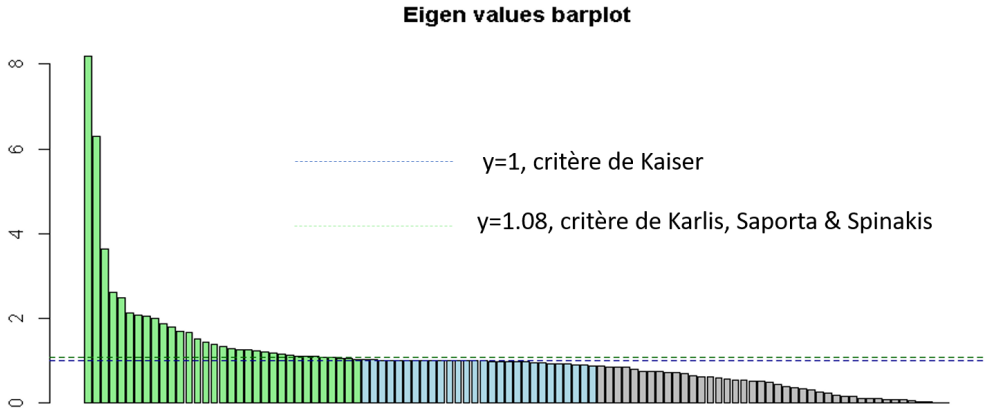


FIGURE 7.1 – Histogramme des valeurs propres de l'AFDM

Le but de notre étude étant de créer une représentation spatiale des véhicules, nous n'allons conserver que 3 axes afin d'obtenir les coordonnées en dimension 3 de ces véhicules, correspondant à leur projection sur ces 3 axes. La part d'inertie conservée est de 18.12 sur 108, soit 16.78%. Ceci peut paraître faible mais comme précisé, le but n'est pas de faire une analyse de données pure. Il serait envisageable de conserver plus d'axes afin d'obtenir les projections en dimension supérieure, i.e. une carte de « n » dimension. Cependant nous mènerons notre étude en dimension 3 à cause de limitations informatiques (puissance de calcul des ordinateurs) : l'ajout d'une dimension multiplie par 7 les temps de calcul lors du lissage des résidus (section 7.3). L'étude sera également plus visualisable en dimension 3.

Tableau des coordonnées

Le tableau des coordonnées (fig. 7.2) décrit l'impact des variables quantitatives et qualitatives sur la définition des axes. Pour les variables qualitatives, il s'agit du carré de rapport de corrélation et la somme en ligne est égale au nombre de modalités de la variable moins 1. Pour les variables quantitatives, il s'agit du carré du coefficient linéaire et la somme en ligne vaut 1.

Le pourcentage en ligne permet de constater la part d'information de la variable retranscrite par l'axe. Par exemple, 89% de la variable classe de prix SRA est retranscrite par l'axe 1. Il y a donc peu de chances que cette variable pèse de façon significative sur les autres axes.

La somme des coordonnées en colonne est égale à la valeur propre associée à l'axe correspondant (Axe 1 : 8.18).

			Axis 1				Axis 2				Axis 3			
name	type	levels	RS1	% col	% row	% row cum	RS2	% col	% row	% row cum	RS3	% col	% row	% row cum
sra_IDMRQ	f	44	0,4083	5%	1%	1%	0,1610	3%	0%	1%	0,2982	8%	1%	2%
sra_SEGMENT	f	7	0,5218	6%	9%	9%	0,7228	11%	12%	21%	0,1107	3%	2%	23%
sra_ALIMENTATION	f	13	0,2627	3%	2%	2%	0,0487	1%	0%	3%	0,4622	13%	4%	6%
sra_ENERGIE	f	6	0,0593	1%	1%	1%	0,1275	2%	3%	4%	0,2531	7%	5%	9%
sra_PFISCALE	q	0	0,5585	7%	56%	56%	0,0061	0%	1%	56%	0,1702	5%	17%	73%
sra_GENRE	f	2	0,0084	0%	1%	1%	0,6962	11%	70%	70%	0,0078	0%	1%	71%
sra_NBPLACES	q	0	0,0300	0%	3%	3%	0,1266	2%	13%	16%	0,0112	0%	1%	17%
sra_POSMOT	f	3	0,0061	0%	0%	0%	0,0043	0%	0%	1%	0,0459	1%	2%	3%
sra_TRANSMIS	f	4	0,2579	3%	9%	9%	0,0691	1%	2%	11%	0,3264	9%	11%	22%
sra_BTVIT	f	5	0,1529	2%	4%	4%	0,0319	1%	1%	5%	0,1061	3%	3%	7%
sra_NBRAPP	q	0	0,1195	1%	12%	12%	0,0002	0%	0%	12%	0,1129	3%	11%	23%
sra_SUSPENS	f	4	0,1485	2%	5%	5%	0,5150	8%	17%	22%	0,0529	1%	2%	24%
sra_TYPFREIN	f	3	0,5302	6%	27%	27%	0,0206	0%	1%	28%	0,0655	2%	3%	31%
sra_ASSFREIN	f	2	0,0527	1%	5%	5%	0,0060	0%	1%	6%	0,0646	2%	6%	12%
sra_DIRECASS	f	3	0,2214	3%	11%	11%	0,0370	1%	2%	13%	0,2419	7%	12%	25%
sra_FREINURG	f	3	0,1761	2%	9%	9%	0,0626	1%	3%	12%	0,2329	6%	12%	24%
sra_GRSRA	q	0	0,8371	10%	84%	84%	0,0107	0%	1%	85%	0,0155	0%	2%	86%
sra_CLSRA	q	0	0,8902	11%	89%	89%	0,0069	0%	1%	90%	0,0000	0%	0%	90%
sra_VEHHRJC	f	2	0,1543	2%	15%	15%	0,3978	6%	40%	55%	0,0086	0%	1%	56%
sra_VNA	f	2	0,0104	0%	1%	1%	0,0017	0%	0%	1%	0,0033	0%	0%	2%
sra_PROPULS	f	4	0,1119	1%	4%	4%	0,0270	0%	1%	5%	0,3317	9%	11%	16%
sra_CARROSSERIE	f	6	0,1425	2%	3%	3%	0,8157	13%	16%	19%	0,1377	4%	3%	22%
sra_DTDEBCOM	q	0	0,1432	2%	14%	14%	0,0351	1%	4%	18%	0,4857	13%	49%	66%
sra_VITESSE	q	0	0,5349	7%	53%	53%	0,3159	5%	32%	85%	0,0097	0%	1%	86%
sra_PREELLE	q	0	0,8185	10%	82%	82%	0,0162	0%	2%	83%	0,0220	1%	2%	86%
sra_LONGUEUR	q	0	0,2552	3%	26%	26%	0,4908	8%	49%	75%	0,0028	0%	0%	75%
sra_LARGEUR	q	0	0,1878	2%	19%	19%	0,5165	8%	52%	70%	0,0008	0%	0%	71%
sra_HAUTEUR	q	0	0,0003	0%	0%	0%	0,7933	13%	79%	79%	0,0418	1%	4%	84%
sra_PDSVIDE	q	0	0,5848	7%	58%	58%	0,2426	4%	24%	83%	0,0104	0%	1%	84%
			8,1854				6,3058				3,6325			

FIGURE 7.2 – Tableau des coordonnées

Nous utilisons, arbitrairement, les mêmes règles de sur-brillance que Tanagra, logiciel d'exploration de données destiné à l'enseignement et la recherche. Ces règles sont les suivantes :

— l'axe doit être significatif, i.e. sa valeur propre doit être supérieure au seuil développé par Karlis, Saporta et Spinakis.
— p est le nombre total de valeurs propres théoriquement non nulles,
— Le pourcentage en ligne doit être supérieur à $\frac{1}{p} = \frac{1}{108} = 0.00925$,
— Le pourcentage en colonne doit être supérieur à $\frac{1}{q} = \frac{1}{29} = 0,03448$ avec q le nombre de variables utilisées dans l'AFDM,

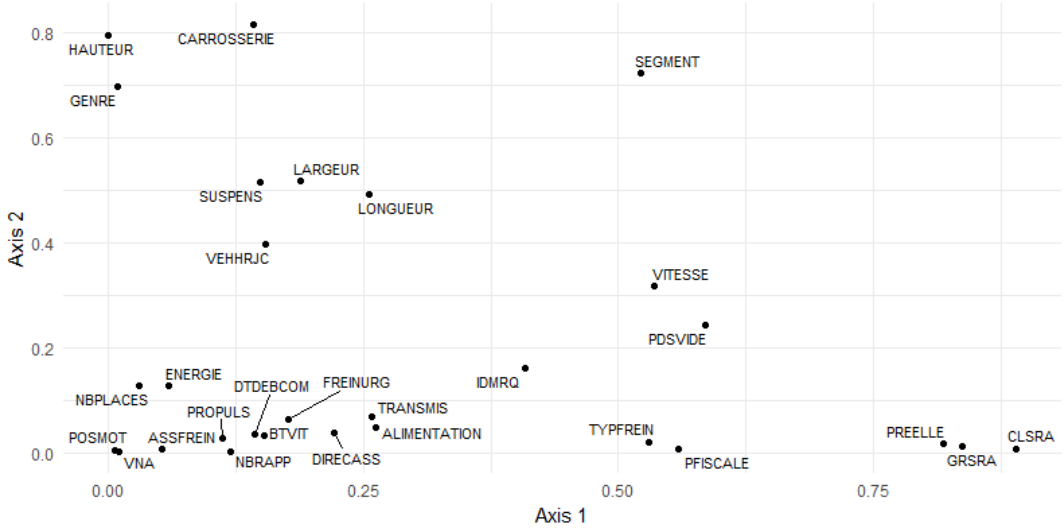


FIGURE 7.3 – Nuage des variables sur l'axe 1 et l'axe 2

Les variables expliquant le mieux le premier axe sont en majorité des variables quantitatives ou ordinales, telles le groupe et la classe de prix SRA, la puissance réelle ou fiscale et la vitesse. Il s'agit de caractéristiques en lien avec la dangerosité ou la gamme du véhicule. Il y a également la marque et le segment, qui vont dans le sens du commentaire précédent.Le second axe semble plus porté par les dimensions (hauteur, largeur, longueur), la carrosserie, le genre (particulier ou utilitaire), et une nouvelle fois le segment du véhicule. Il apparaît donc que cet axe est plutôt défini par l'aspect ou l'image du véhicule. Enfin, le dernier axe est porté par des caractéristiques techniques du véhicule, comme l'alimentation, l'énergie, la transmission, la présence de système de freinage d'urgence ou de direction assistée. Notons également l'importance de la variable date de commercialisation, ce qui est logique avec la chronologie des avancées technologiques ci-dessus.

Nous pouvons retrouver ces résultats graphiquement à l'aide des nuage des variables (fig. 7.3), largement utilisé lors des ACM. Les trois variables puissance réelle, classe de prix et groupe SRA se trouvent en bas à droite du graphique. Ces variables sont donc quasiment expliquées par l'axe 1, et nullement par l'axe 2. Contrairement à la vitesse maximale ou le poids à vide, au milieu du graphique, qui sont à la fois expliqués par l'axe 1 et 2. L'ensemble de ces graphiques, sur les axes 1 et 3 et sur les axes 2 et 3, est disponible en Annexes (fig. A.9 et A.10).

A noter que la variable VNA, acronyme de Véhicule Non Assurable, liste créée par les équipes opérationnelles jugeant de la dangerosité d'un véhicule et décidant si ce véhicule peut être assuré ou non, n'apporte de l'information sur aucun des 3 axes. Ceci est potentiellement un indicateur de révision de ce critère.

Tableau des corrélations

Le cercle de corrélation sert à obtenir une vue synthétique du positionnement des variables quantitatives par rapport aux axes, le sens des relations entre eux. Ils sont visibles sur la Figure 7.4. Il existe un fort effet taille dans notre étude. En effet, les véhicules puissants sont chers et lourds. Sur le second axe les variables de puissance et vitesse s'opposent aux variables de dimension, à savoir hauteur largeur et longueur. En d'autres termes, les voitures lourdes et hautes s'opposent aux voitures rapides. Enfin, la puissance fiscale s'oppose légèrement à la date de commercialisation sur l'axe 3, les véhicules récents auraient donc des puissances fiscales plus élevées que par le passé (la table des corrélations avec les valeurs est disponible en Annexe fig. A.12).

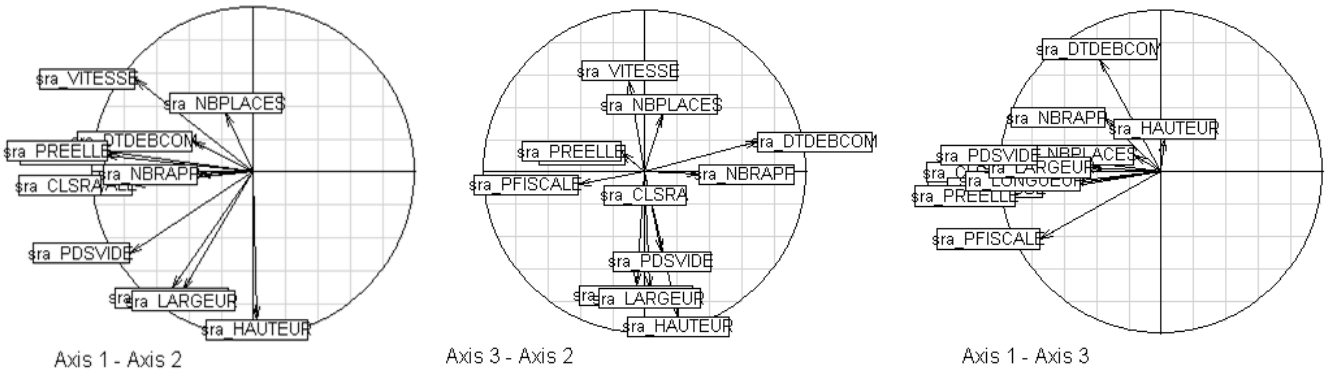


FIGURE 7.4 – Cercles des corrélations

Projection des individus

Afin de pouvoir interpréter correctement le résultat de l'Analyse Factorielle de Données Mixtes, nous pouvons représenter le nuage des individus en fonction des modalités de chaque variable. Ce genre de graphique est similaire à ceux disponibles lors d'une Analyse par Correspondance Multiple. Pour ce faire nous avons créé une fonction R largement inspirée de la fonction s.class du package ade4. Les variables continues ne conservent qu'un gradient de couleur, tandis que le barycentre de chaque modalité est représenté pour les variables qualitatives, en plus des individus. Nous allons détailler ci-dessous plusieurs variables.

La figure 7.5 représente donc le nuage des individus sur les trois axes retenus de l'AFDM en fonction de la vitesse maximale, le vert représentant les véhicules de faible vitesse et le rouge, véhicules les plus rapides. Un premier constat est l'échelle de couleurs. En effet, la vitesse maximale la plus élevée de la base SRA est 350km/h alors que celle du graphique ne dépasse pas 250km/h. En effet, le nuage ne comporte que des véhicules présents dans le portefeuille de Direct Assurance, et la vitesse fait partie des variables utilisées lors des règles d'acceptation des contrats. Ensuite, il y a une tendance claire de la vitesse sur l'axe 1, sauf pour les vitesses les moins élevées pour lesquelles l'axe 2 joue un rôle important. Ces résultats (importance de la tendance en fonction des axes) étaient attendus au vu du tableau des coordonnées (Figure 7.2).

Nous réitérons l'analyse avec le poids à vide du véhicule représenté sur la figure 7.6, le vert représentant les véhicules légers et le rouge, véhicules les plus lourds. Comme pour la vitesse, le poids est fortement porté par l'axe 1, excepté pour les valeurs élevées réparties sur l'axe 2. Ces

FIGURE 7.5 – Nuage des individus selon la vitesse

résultats étaient attendus au vu du tableau des coordonnées (Figure 7.2). De même, les cercles des corrélations (Figure 7.4) indiquaient une corrélation positive entre le poids et la vitesse sur l'axe 1, et négative sur l'axe 2. Les véhicules lourds sont généralement plus rapides, notamment à cause du poids du moteur, excepté pour les véhicules utilitaires qui sont lourds à cause de leur carrosserie. Ces véhicules sont représentés en bas du nuage selon l'axe 1 et l'axe 2 (selon les valeurs les plus faibles de l'axe 2), ce qui est visible sur la figure 7.7 représentant le nuage des individus selon la carrosserie.

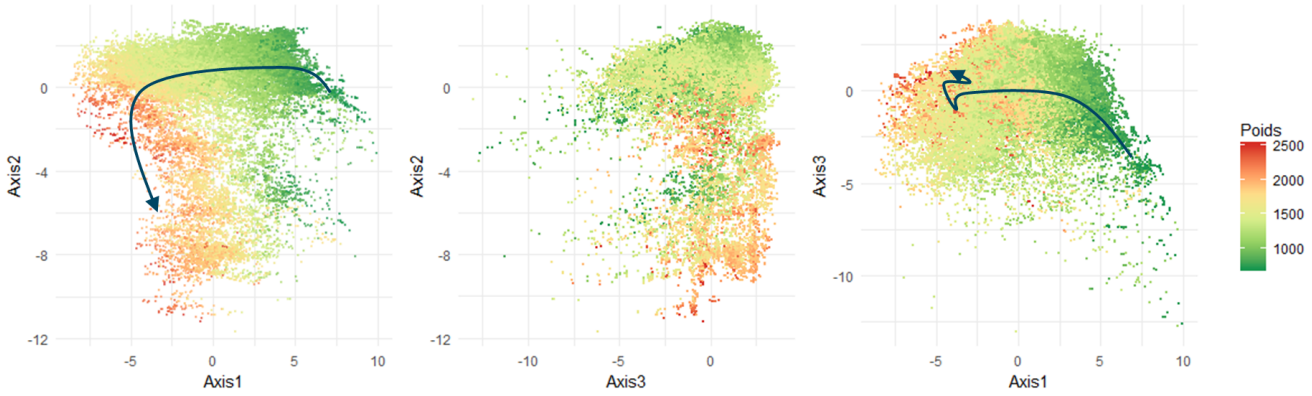


FIGURE 7.6 – Nuage des individus selon le poids à vide

Il est clair que l'information sur la carrosserie est portée par l'axe 2. Cependant la distinction n'est faite qu'entre les véhicules utilitaires (OTH) et les autres véhicules de type berline (BER), commerciales (COM), monospaces (MSP), coupés cabriolets (CPC) et breaks (BRK) sont regroupés. On peut confirmer ce point avec les barycentres des modalités représentés avec des étiquettes sur les graphiques.

Enfin, nous pouvons regarder la structure du nuage selon la classe de prix (figure 7.8). Comme attendu, l'information est portée par l'axe 1 et est positivement corrélée avec la vitesse et le poids, ce qui est logique, plus les véhicules sont rapides, plus ils coûtent chers. Il n'y a cependant aucune information récupérée par l'axe 2.

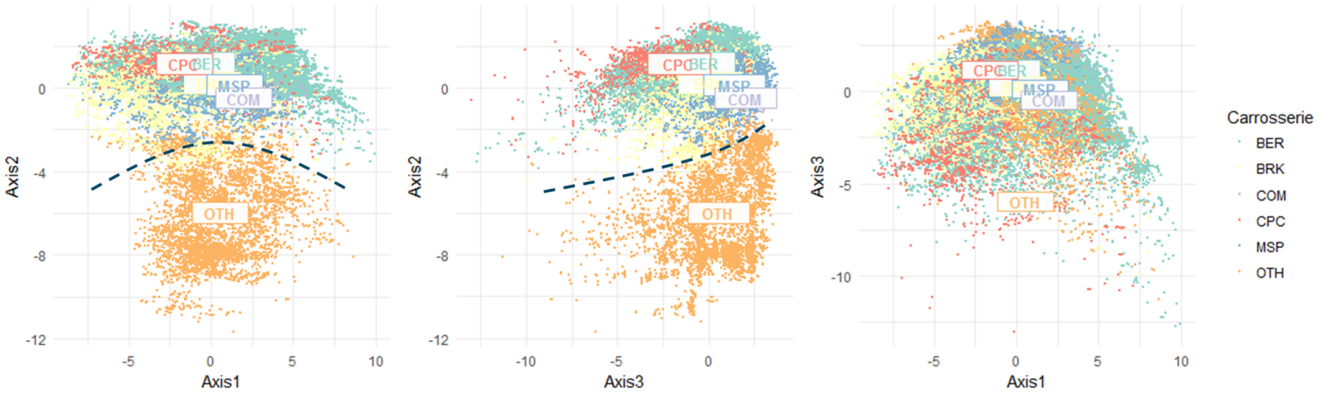


FIGURE 7.7 – Nuage des individus selon la carrosserie

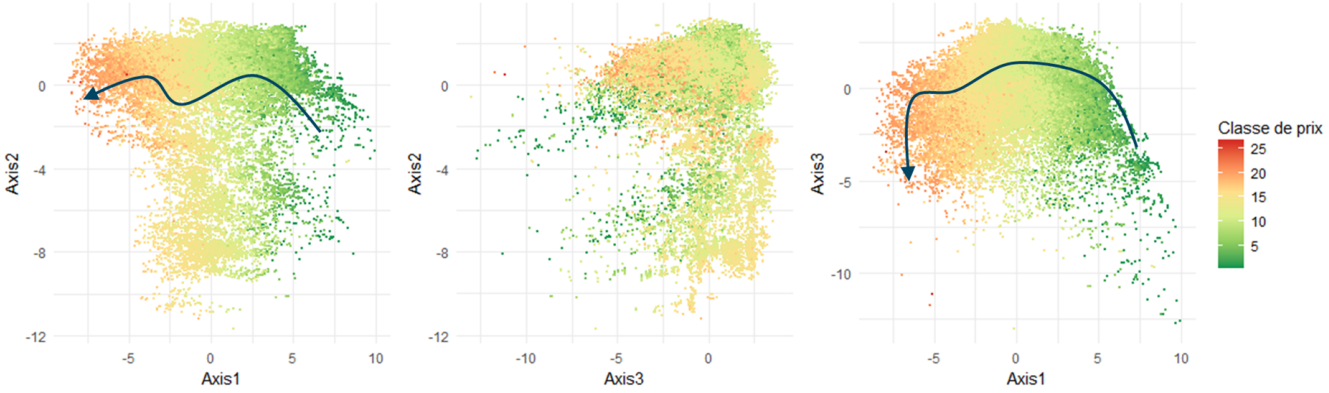


FIGURE 7.8 – Nuage des individus selon la classe de prix

D'autres graphiques intéressants sont disponibles en Annexe telles le groupe SRA fig. A.13 (résultats très similaires à la classe de prix) et la propulsion (fig. A.14) dont l'information est majoritairement portée par l'axe 3.

A la suite de cette étude du résultat de l'Analyse Factorielle de Données Mixtes, le nuage de points semble cohérent avec les caractéristiques véhicule.

2 Étude de la table de contiguïté

Le résultat de la section précédente, obtenu à l'aide de l'Analyse Factorielle de Données Mixtes, est une liste de points géolocalisés dans l'espace avec des coordonnées x, y et z. Chacun d'eux représente un véhicule, ou plus précisément une combinaison unique de caractéristiques véhicules. Tous sont présents dans le portefeuille de Direct Assurance. Ce nuage de point constitue l'entrée des données de la création de la table de contiguïté.

Les points sont liés via la création d'une relation de voisinage. Cette relation découle de la méthode de Triangulation de Delaunay, évoquée dans la sous-section 5.2.2. La fonction *delaunayn* du package *geometry* est appliquée au nuage de point. A partir des 45371 Codes Clé (ou points), la fonction rend une matrice de dimension 1.109.192 x 4, chaque ligne correspondant à un tétraèdre composé de 4 points. Cette dernière est convertie en table de contiguïté à proprement parler à l'aide

d'un code R. La figure 7.9 permet de représenter graphiquement cette table : chaque Codes Clé est relié à ses voisins par un segment. La forme du nuage de point peut être devinée, mais l'image reste cependant peu lisible au vu du nombre de liaisons. Il y en a 321066 au total. La figure est néanmoins importante car elle fait apparaître le caractère convexe de la triangulation de Delaunay. Certains points, notamment en périphérie du nuage, ont beau être fort éloignés les uns des autres, ils sont tout de même considérés comme voisins (par exemple les liaisons du point $y = -8$ et $z = -10$). Ce problème a déjà été soulevé lors de la définition de la Triangulation de Delaunay dans la section 5.2.2.

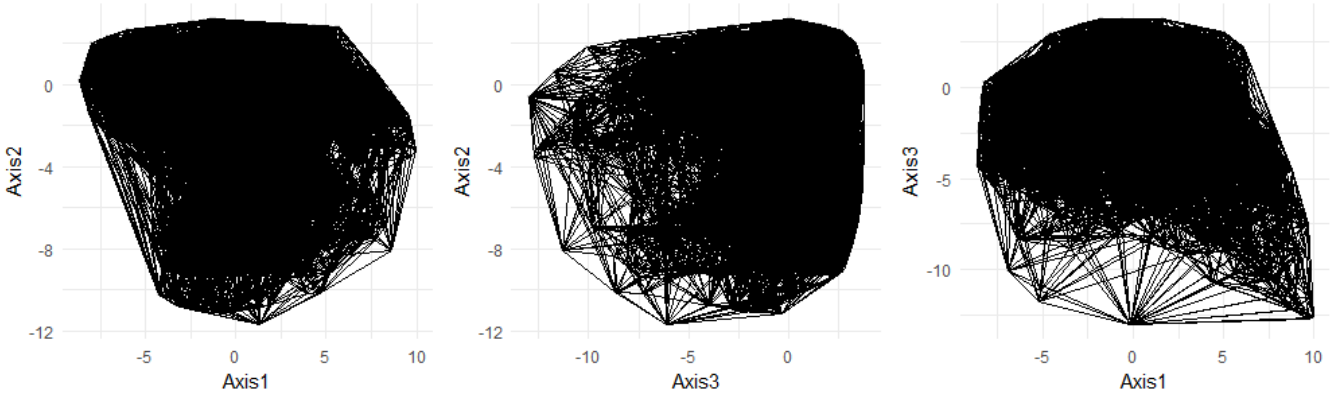


FIGURE 7.9 – Représentation de la table de contiguïté initiale

La convexité de la Triangulation de Delaunay a donc pour effet de rendre voisins deux véhicules alors que leurs projections dans l'espace ne le sont pas. Ainsi, un véhicule de classe de prix faible peut se retrouver voisin d'un véhicule au prix élevé. Ceci est visible sur la table des modalités 7.10 obtenue à l'aide d'une code R. Cette dernière contient le nombre de liaisons par classe de véhicules : les liaisons entre véhicules de même classe de prix se trouvent dans les cellules vertes foncées. Le code couleur est arbitraire, plus l'écart entre les classes est grand, plus la couleur devient chaude. Les liaisons qualifiées d'aberrantes (arbitrairement) sont dans les cases colorées en rouge. Ces dernières sont au total 891. Par exemple, la liaison entre le véhicule de classe de prix V avec un autre de classe F peut naturellement être considérée non pertinente.

Une solution pour résoudre ce problème consiste à supprimer les liens dont la longueur est supérieure à un seuil défini. Cependant, la méthodologie de création de classification de véhicule se doit d'être le plus automatique possible pour des contraintes d'efficacité et d'uniformité entre les différentes garanties ou entités d'AXA, et doit contenir le moins de paramètres "humains" que possible. Il conviendra alors de déterminer ce seuil en fonction des quantiles de la longueur des liaisons (évaluée selon la distance euclidienne), et ne supprimer que les liens se trouvant dans le dernier centile.

La table 7.1 contient les intervalles des derniers centiles de la distance entre chaque voisins ainsi que la distance moyenne de ce centile. Le dernier quantile a une distance moyenne plus de deux fois supérieure à celle du quantile précédent. L'écart entre la moyenne du quantile 98 et 99 est en comparaison acceptable. De plus, la longueur de l'intervalle du dernier centile est nettement supérieure à celle des autres, dû à la valeur maximale de 15.9. Il n'y a donc pas d'incohérence à choisir un seuil à 1.50632554 afin de ne supprimer que les liaisons du dernier centile, limitant ainsi les aberrations de la table de contiguïté.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	HC
A	1 645	594	495	406	396	286	216	180	80	40	25	35	20	5	16	0	0	3	0	0	0	0	0
B	594	547	723	475	380	274	223	140	55	21	5	15	6	0	6	0	0	0	0	0	0	0	0
C	495	723	1 581	1 568	951	528	362	285	111	33	24	16	9	1	3	2	0	0	0	0	0	0	0
D	406	475	1 568	2 909	2 903	1 585	819	489	198	64	34	16	8	1	4	1	0	0	0	0	0	0	0
E	396	380	951	2 903	4 785	4 468	1 961	979	470	174	44	20	14	5	5	1	0	0	0	0	0	0	0
F	286	274	528	1 585	4 468	7 361	5 976	2 635	1 195	474	127	39	20	5	7	3	1	0	0	0	0	1	0
G	216	223	362	819	1 961	5 976	9 603	7 957	3 437	1 306	360	112	41	12	15	3	3	0	0	0	0	0	0
H	180	140	285	489	979	2 635	7 957	12 364	8 743	3 295	1 116	449	131	45	22	13	2	0	0	0	0	0	0
I	80	55	111	198	470	1 195	3 437	8 743	13 494	9 595	3 985	1 473	396	138	26	12	3	1	1	0	0	0	0
J	40	21	33	64	174	474	1 306	3 295	9 595	14 469	11 420	4 332	1 526	490	90	38	13	2	0	0	0	0	0
K	25	5	24	34	44	127	360	1 116	3 985	11 420	18 101	12 503	4 880	1 649	492	89	35	8	2	0	0	1	0
L	35	15	16	16	20	39	112	449	1 473	4 332	12 503	20 361	15 290	5 341	1 888	483	153	20	8	0	0	1	0
M	20	6	9	8	14	20	41	131	396	1 526	4 880	15 290	23 408	15 140	5 605	1 703	607	146	54	11	0	4	1
N	5	0	1	1	5	5	12	45	138	490	1 649	5 341	15 140	20 421	13 148	4 777	1 770	553	133	17	0	4	1
O	16	6	3	4	5	7	15	22	26	90	492	1 888	5 605	13 148	15 638	10 000	3 890	1 330	360	57	0	0	0
P	0	0	2	1	1	3	3	13	12	38	89	483	1 703	4 777	10 000	11 270	7 315	2 871	789	200	11	3	2
Q	0	0	0	0	0	1	3	2	3	13	35	153	607	1 770	3 890	7 315	8 771	5 720	1 811	484	5	9	2
R	3	0	0	0	0	0	0	0	1	2	8	20	146	553	1 330	2 871	5 720	5 955	3 288	1 000	23	19	1
S	0	0	0	0	0	0	0	0	1	0	2	8	54	133	360	789	1 811	3 288	3 229	1 802	62	46	7
T	0	0	0	0	0	0	0	0	0	0	0	0	11	17	57	200	484	1 000	1 802	1 868	86	72	20
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	5	23	62	86	6	3	1
V	0	0	0	0	0	1	0	0	0	0	1	1	4	4	0	3	9	19	46	72	3	4	3
HC	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2	2	1	7	20	1	3	0

FIGURE 7.10 – Table des modalités - Classe de Prix

Centile	Intervalle	Distance moyenne
97	]0.86007875 ; 0.96400626]	0.90914166
98	]0.96400626 ; 1.13146671]	1.03766643
99	]1.13146671 ; 1.50632554]	1.28727080
100	]1.50632554 ; 15.95733289]	2.55139843

TABLE 7.1 – Derniers centiles des distances entre voisins

Une fois ces aberrations retirées, le nuage de points est plus cohérent sur le critère de la distance entre les Codes Clé (cf. Figure 7.11). Cependant, des aberrations selon la classe de prix sont toujours présentes (619 au total). Cette table des modalités sur la classe de prix, une fois les liaisons du dernier centile retirées, est disponible en Annexe fig. A.15. Il conviendra de supprimer ces liaisons selon un "rayon" de 6, i.e. tous les voisins d'un véhicule de classe L, de classe non comprise entre les classes E et R, ne seront plus liés avec ce dernier. Le choix de cette variable est justifié par l'importance de son pouvoir explicatif du nuage, tandis que le rayon de 6 est basé sur le jugement d'expert.

3 Résultats des classifications

Nous disposons maintenant d'une représentation spatiale des véhicules ainsi que des relations de voisinages entre eux. Selon l'hypothèse de décomposition du signal vue sur la figure 5.1, nous estimons qu'il y a au sein même des résidus une variation résiduelle expliquée par des caractéristiques de véhicule non captée par les variables véhicule déjà intégrées dans les modèles. Nous allons alors pouvoir récupérer la variation résiduelle définie dans la section 5.1 et l'associer à chacun des véhicules de notre carte. C'est sur ces données que nous allons pouvoir effectuer une classification afin d'avoir des classes robustes et homogènes. Ces classes constitueront ce que nous appelons un véhiculier.

La classification est effectuée à l'aide du logiciel Classifier fourni par l'éditeur Willis Towers Watson. Le processus défini pour la classification géographique est disponible en Annexe (fig A.16). Nous l'avons adapté pour qu'il puisse correspondre à une classification de véhicules. Le logiciel a

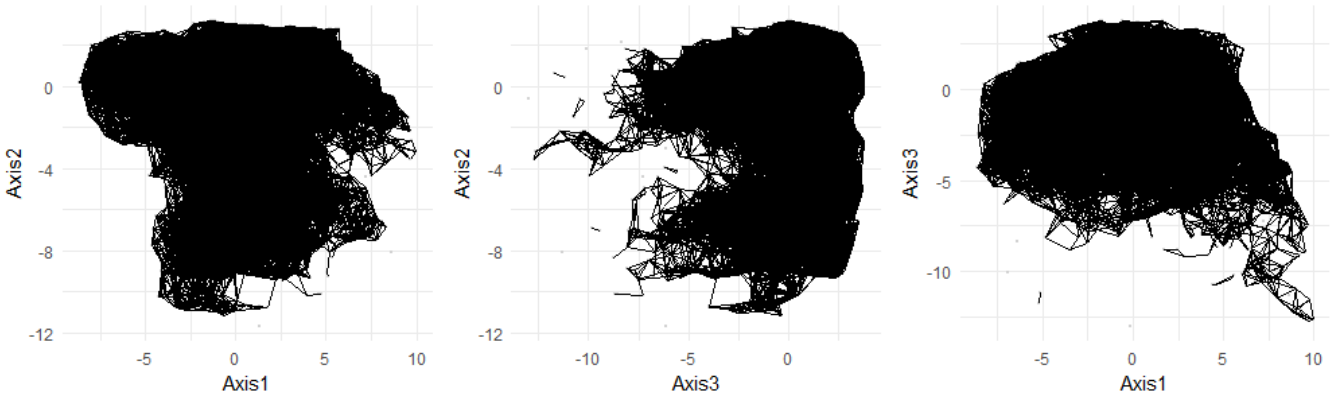


FIGURE 7.11 – Représentation de la table de contiguïté n'ayant conservé que 99% des liaisons

besoin d'un modèle linéaire généralisé en entrée ainsi qu'une table de contiguïté (habituellement une carte de France et les communes voisines). Une fois les variables véhicules spécifiées (incluses ou non dans le modèle), Classifier va tenter de déceler un signal dans les résidus à l'aide de ces variables, qui correspondra à un signal de véhicule sous-jacent. Ces résidus sont liés entre eux à l'aide de la table de contiguïté. Les résidus et les valeurs prédites du GLM sont ensuite standardisées afin de ne pas biaiser les effets dûs aux véhicules. Les résidus sont alors lissés à l'aide des méthodes décrites dans la section 5.3 avec un niveau de lissage (précisé dans la suite de cette section).

La classification intervient à ce moment et deux choix sont possibles :

— Les résidus standardisés lissés sont combinés avec les prédictions standardisées afin de créer des classes. Ainsi, une fois la classification réintégrée dans le GLM, les variables externes n'auront plus besoin d'être intégrées. Cette solution est la plus souvent utilisée lors de la création de zoniers dans la mesure où les variables externes telles la densité de population ou le taux de cadres dans une commune donnée est bien souvent indisponible dans les systèmes informatiques et bien laborieux à mettre en place. Il en résulte cependant une perte légère de l'information ;

— Seuls les résidus standardisés lissés sont utilisés pour créer les classes et les variables externes peuvent être conservées dans le GLM lors de l'ajout de la classification. Ainsi le signal reste capté et le gain d'information est totalement porté par la variation résiduelle des véhicules.

Nous allons tester ces deux méthodes afin de mesurer la perte d'information due à l'intégration des prédictions standardisées dans le véhiculier, compensée par une simplification du modèle. Cependant, Direct Assurance dispose déjà des variables véhicules dans son système informatique. La première solution n'est donc pas nécessaire, à moins que les résultats soient meilleurs que le deuxième type de classification. Par ailleurs, nous allons tester trois méthodes de classification (Equal Weigth, Quantiles et Ward) décrites dans la section 5.4, afin de vérifier laquelle donne de meilleurs résultats. Cela fait un total de 6 classifications à tester et à réinjecter dans les modèles GLM dont les résultats seront présentés dans le chapitre suivant.

Lissage des résidus

Comme indiqué dans l'hypothèse de décomposition du signal, les variables véhicule n'expliquent qu'une partie du signal. Le reste du signal se trouve dans les résidus du modèle. Le lissage et la classification des résidus ne va s'effectuer que sur l'échantillon de modélisation, soit sur 80%

des données. Cet échantillon est décomposé en deux sous-échantillons de 60% et de 20%, appelés respectivement échantillon d'entraînement du zonier et échantillon de détermination du niveau de lissage (cf. section 3.3).

En effet, nous avons besoin de lisser les résidus car ils ne peuvent être intégrés en tant que tels dans le modèle dans la mesure où ils sont trop hétérogènes entre voisins, ce que nous pouvons voir sur la figure 7.12 de façon analogue à une étude de lissage pour des résidus géographiques. Notre représentation spatiale est néanmoins en trois dimensions tandis que les cartes géographiques n'ont que deux dimensions. Beaucoup de Codes Clé ont des résidus nuls (représentés par la couleur bleue) : ce sont tous les Codes Clé n'ayant pas eu de sinistres, soit 71% des Codes Clé qui contiennent environ 18% de l'exposition de l'échantillon d'entraînement du zonier. En revanche, les Codes Clé rouges sont ceux avec très peu d'exposition mais où il y a eu un sinistre. La prédiction est ainsi souvent biaisée par manque de robustesse et loin de l'observé. Les Codes Clé de couleur intermédiaire sont des véhicules pour lesquels l'exposition est suffisante et ayant eu beaucoup de sinistres.

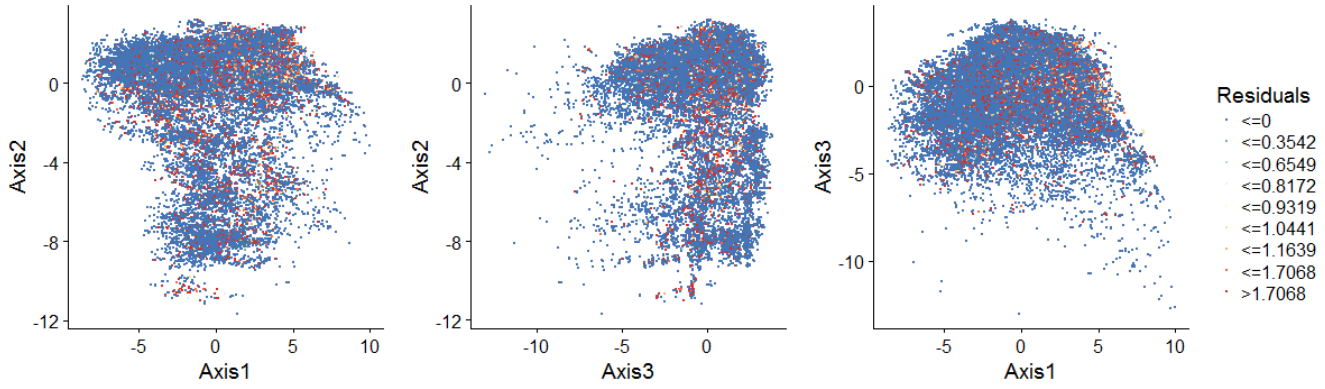


FIGURE 7.12 – Nuage des individus, résidus non lissés

Nous pouvons trier ces résidus dans l'ordre croissant (en vert foncé) et les comparer aux résidus de l'échantillon de détermination du niveau de lissage (en vert clair), ce que nous pouvons voir sur les figures 7.13 et 7.14. Ces graphiques sont issus du logiciel Classifier. Le but est de trouver une tendance dans les résidus de l'échantillon de détermination, ce qui à première vue n'est pas le cas. Cependant lorsque nous effectuons un agrandissement de l'échelle, nous constatons qu'un signal croissant est bel et bien présent.

L'objectif du lissage est d'améliorer les estimations du risque pour chaque Code Clé en fonction de l'expérience des véhicules voisins. Le lissage est basé sur la table de contiguïté précédemment créée. Il s'agit maintenant de définir le bon niveau de lissage. Si le niveau de lissage est trop peu élevé, il restera trop de bruit dans les résidus lissés. S'il est au contraire trop grand, il y aura une perte de signal. Le lecteur pourra trouver plus d'information théorique concernant le lissage de résidus dans le cadre de lissage spatial en tarification géographique dans l'étude MICU [2012] [13].

Concrètement, il faut trouver le niveau de lissage optimal pour pouvoir aligner les résidus de l'échantillon d'entraînement et ceux de l'échantillon de détermination du lissage. En fonction du niveau de lissage choisi, la courbe rose (initialement égale à la courbe vert foncé) deviendra de moins en moins discriminante et s'aplatira pour se rapprocher de la courbe vert clair. Le niveau de lissage optimal se déterminera à l'aide de l'erreur quadratique moyenne. Pour chaque Code Clé, le carré de

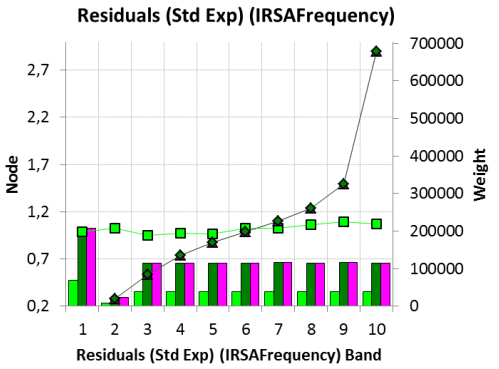


FIGURE 7.13 – Résidus Tries

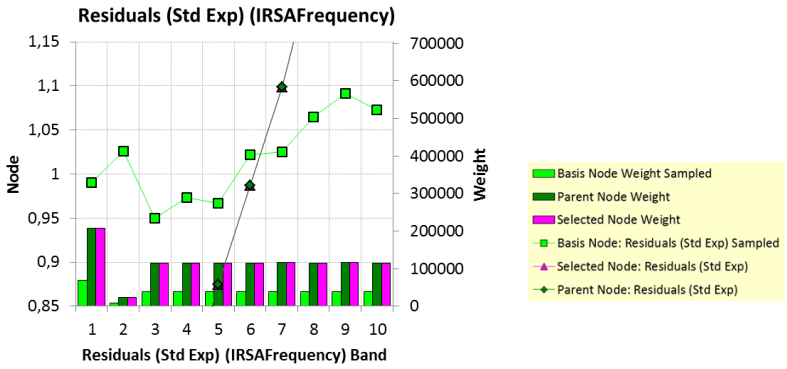


FIGURE 7.14 – Zoom sur les résidus

la distance entre les résidus lissés et les résidus de l'échantillon a été calculé, puis la moyenne des distances au carré. Nous obtenons alors la figure 7.15 suivante, la moyenne des distances au carré en fonction du niveau de lissage. Le niveau de lissage optimal apparaît donc comme le niveau à 90% [1].

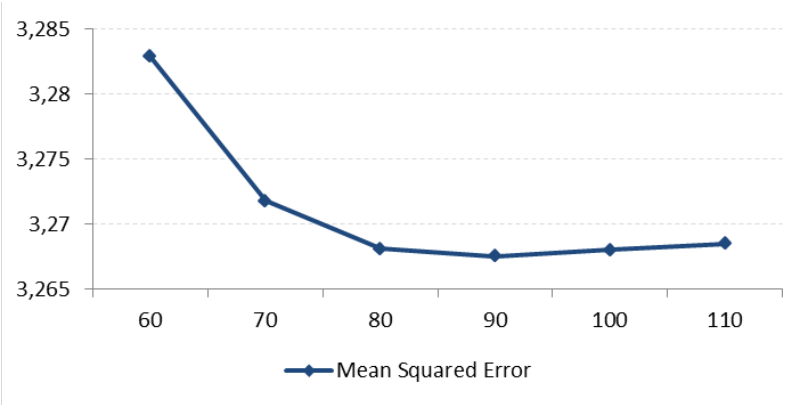


FIGURE 7.15 – Erreur quadratique moyenne du niveau de lissage

Une fois les résidus lissés, nous pouvons retracer notre nuage d'individus (figure 7.16) et constater que ces résidus auront plus de potentiel à être classés et intégrés dans les modèles. Le nombre de Codes Clé ayant un résidu à 0 a fortement diminué. Il s'agit des véhicules isolés, à l'extérieur du nuage d'individus, des "îles" créées lors du retraitement de la table de contiguïté. Afin de palier à ce problème nous aurions pu décider de conserver un lien avec le plus proche voisin. Cependant cette création d'îlot est considérée comme satisfaisante. Par ailleurs, le nombre de classes a fortement diminué, réduisant ainsi les valeurs extrêmes : en conservant les mêmes 9 classes de résidus que dans la figure 7.12, nous n'en avons que 6 représentées.

Classification des résidus

Comme précisé auparavant, nous allons utiliser 3 types de classification : les classes égales (equal weigth), les quantiles, et la classification de Ward. Nous allons également classer deux types

1. La valeur du niveau de lissage est spécifique à Classifier.

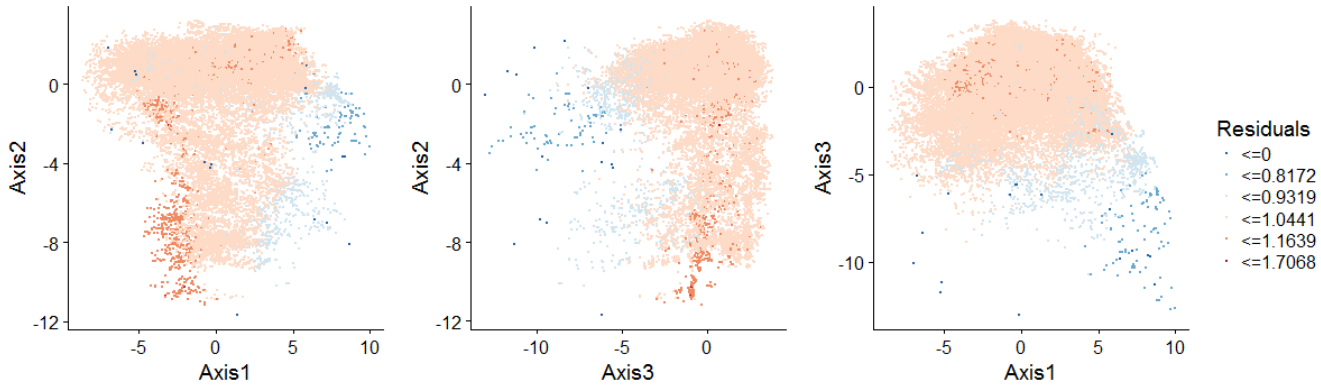


FIGURE 7.16 – Nuage des individus, résidus lissés à 90

de quantités : les résidus lissés et les prédictions standardisées, qui sont obtenus en multipliant les coefficients des variables véhicules issus du modèle avec les résidus lissés. Cette dernière méthode nous permettra de nous affranchir des variables véhicules dans le modèle et de le simplifier considérablement, au détriment d'une perte d'information. Il y aura donc 6 types de classifications testées dans les modèles. Pour plus de clarté, l'ensemble des classifications ainsi que leur nom est résumé dans le tableau 7.2.

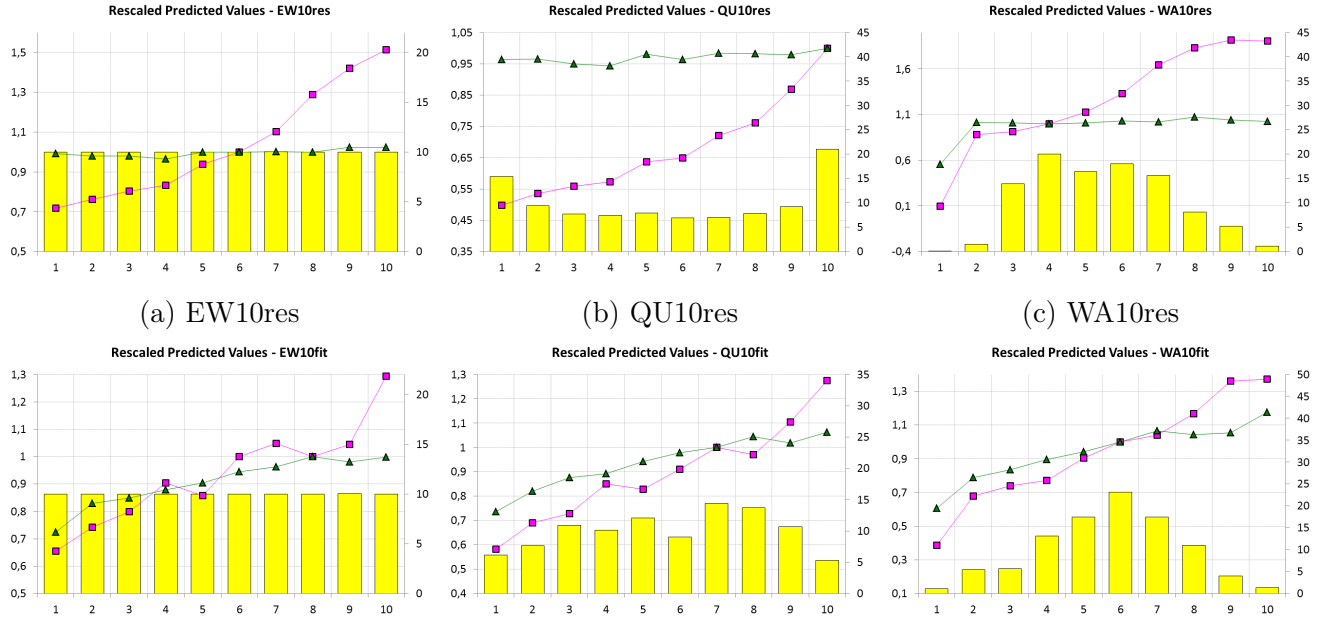
	Equal Weigth	Quantiles	Ward
Résidus Lissés	EWres	QUres	WAres
Prédictions Standardisées	EWfit	QUfit	WAfit

TABLE 7.2 – Synthèse des classifications testées

Le nombre de classes optimal a également été testé au préalable, incluant la génération des 6 types de classifications et leur test dans les modèles. Les nombres de classes testées ont été 5, 10, 15 et 20. Une classification à 5 classes ne permettait pas de segmenter suffisamment le risque, tandis que le nombre de classes de 15 et 20 entraînaient des problèmes de convergence du modèle sur la méthode des quantiles. Ainsi nous avons retenus et ne présenterons que des classifications avec un nombre de classes égal à 10.

Nous avons représenté dans la figure 7.17 l'ensemble des classifications sur l'échantillon de modélisation (80% de la base initiale). Nous pouvons remarquer qu'il y a bien une tendance croissante de l'observé (en rose) : plus les classes sont élevées, plus le risque est élevé, et ce peu importe la méthode de classification utilisée ou la quantité classifiée. La méthode permettant de segmenter le plus le risque est celle de Ward : l'écart entre la valeur de l'observé réduit de la classe 2 et celle de la classe 10 est de 1.3 pour la classification des résidus et de 0.78 pour la classification des prédictions corrigées standardisées. Attention cependant, le faible volume des classes extrêmes ne permettra sûrement pas d'obtenir des résultats robustes et elles devront sans doute être regroupées avec les classes adjacentes.

Les graphiques de la première ligne représentent les classifications sur les résidus seulement. Aucun signal de cette variable n'est capté (les prédictions sont représentées en vert foncé) : l'intégration de ces variables dans le modèle peut potentiellement être bénéfique et améliorer ce dernier.



(a) EW10res (b) QU10res (c) WA10res

(d) EW10fit (e) QU10fit (f) WA10fit

FIGURE 7.17 – Observé et prédit de chacune des classifications dans le modèle

Les graphiques de la seconde ligne représentent les classifications des valeurs prédites par les variables véhicule combinées aux résidus. Il est donc logique de voir qu'une partie du signal est déjà captée par le modèle actuel qui contient les variables véhicules. L'ajout de ces classifications n'aurait que peu d'intérêt dans le modèle, sauf si les variables véhicules sont retirées, permettant ainsi une simplification du modèle.

Nous disposons désormais de 6 véhiculiers distincts que nous allons pouvoir tester dans le modèle linéaire généralisé. Cette étude sera présentée dans le chapitre 8.

Chapitre 8

Résultats et limites

Dans ce chapitre, nous allons intégrer les différentes classifications créées dans le chapitre précédent dans le modèle de fréquence des sinistres conventionnés afin de mesurer l'apport prédictif de cette variable. Les résultats des véhiculiers créés sur les résidus seulement seront d'abord exposés avant de se concentrer sur les véhiculiers issus de la combinaison entre prédictions et résidus. La classification optimale sera enfin choisie avant de mesurer les limites de cette méthodologie de classification de véhicules et de réfléchir aux améliorations et aux alternatives.

1 Références du modèle fréquence IRSA

Avant de tester les différents véhiculiers dans le modèle, nous devons définir un point de référence en terme de Déviance, de coefficient de GINI et d'Expexted Déviance Ratio (EDR) afin de mesurer l'amélioration ou la dégradation du modèle. Ces chiffres sont disponibles dans le tableau 8.1. Le coefficient de GINI du modèle initial est de 25.254% sur l'échantillon de modélisation et de 25.375% sur l'échantillon de validation, soit un écart relatif de 0.48%, ce qui est très satisfaisant. Pour rappel, il ne faut pas que l'écart entre les deux GINI soit trop important car cela se traduit par un mauvais ajustement du modèle sur l'un ou l'autre des échantillons. L'EDR, calculé à l'aide de la déviance nulle, est de -2.5%, ce qui est également satisfaisant.

La description du modèle ainsi que sa validation ont été évoqué dans le chapitre 6.

2 Résultat des véhiculiers sur résidus

Le but de cette section est de tester chacun de ces véhiculiers dans le modèle de fréquence IRSA. Etant donné qu'ils n'ont été créés qu'à l'aide des résidus (et non la combinaison des prédictions des variables véhicule et des résidus), les variables liées aux caractéristiques automobiles présentes dans le modèle sont conservées. Comme évoqué dans le chapitre précédent et visible sur les figures 7.17a, 7.17b et 7.17c, le signal de cette variable n'est pas capté par le modèle actuel (la courbe verte des valeurs prédites est horizontale).

La structure du modèle final peut donc être résumée de la façon suivante :

$$\text{Variables non véhicule} + \text{Variables véhicules} + \text{Véhiculier(résidus)}$$

La méthode habituelle pour tester une variable est de l'inclure dans le modèle sur l'échantillon de modélisation, vérifier qu'elle remplit les critères de sélection de variable décrits dans la section 2.2, puis d'étudier sa robustesse sur l'échantillon de validation. Nous allons la nommer Méthode 1. Seuls les résultats sur la classification des véhicules utilisant la méthode de Ward seront exposés. Les résultats des autres sont néanmoins disponibles en Annexe (fig. A.17 à fig. A.22).

Une fois introduite dans l'échantillon de modélisation, nous constatons que les résultats de la variable WA10res sont très bons. Les deux premières classes ont été regroupées pour cause de robustesse et de fiabilité. Le spread est de 136% (cf. figure 8.1), permettant de segmenter énormément les assurés. Le coefficient de GINI augmente de 20% (5 pts) et l'EDR baisse de 41%, ce qui est assez spectaculaire. Cependant, après vérification sur l'échantillon de validation (cf. figure 8.2), les prédictions sont très éloignées des observés : la courbe vert foncé n'est pas du tout en ligne avec la courbe rose. Le coefficient de GINI diminue même de 12.5%, creusant ainsi l'écart de cette mesure entre les deux échantillons. Enfin, l'EDR se dégrade sur l'échantillon de validation. Ceci dénote la présence d'un sur-apprentissage au détriment de la validation. Ce résultat pouvait être attendu, dans la mesure où 75% de l'échantillon de modélisation a été utilisé pour créer cette classification.

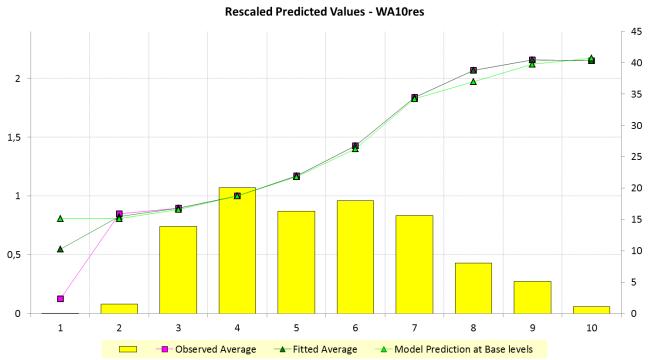


FIGURE 8.1 – WAres sur l'échantillon de modélisation, Méthode 1

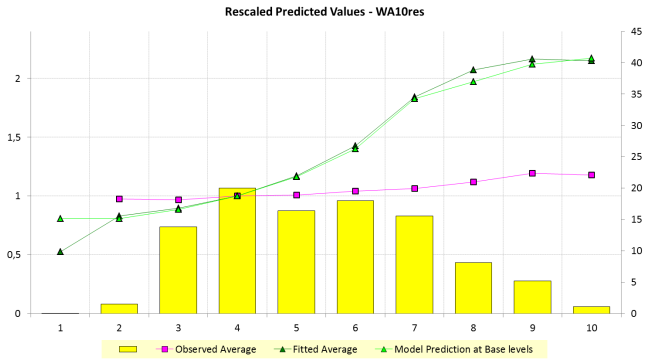


FIGURE 8.2 – WAres sur l'échantillon de validation, Méthode 1

Nous obtenons le même résultat non concluant avec la classification en classes égales EW10res. Il n'est donc pas nécessaire de tester la dernière classification avec la Méthode 1. Cette méthodologie pousse le modèle à sur-apprendre, expliquant la différence de prédiction entre l'échantillon de modélisation et celui de validation. Afin d'éviter ce problème, deux solutions sont possibles : lisser davantage les résidus quitte à perdre énormément de pouvoir prédictif sur l'échantillon d'apprentissage, ou changer la méthodologie d'introduction dans le modèle. Nous avons alors développé une Méthode 2, consistant à déterminer les coefficients du GLM sur l'échantillon de détermination du lissage, lequel n'a pas été utilisé lors de la création du zonier (seulement pour obtenir le niveau de lissage optimal). Concrètement, cette méthode consiste à effectuer les étapes suivantes. On fixe les coefficients du modèle initial (sur l'échantillon de modélisation, les 80%), on l'applique sur l'échantillon de détermination du lissage. La variable véhiculier est alors ajoutée, permettant de calculer des coefficients pour chaque classe. On ajoute alors ces coefficients dans le modèle initial libéré (et l'échantillon de modélisation), entraînant un ajustement potentiel des autres variables. Enfin, c'est ce modèle final qui sera appliqué sur l'échantillon de validation pour obtenir les résultats.

C'est cette Méthode 2 qui a été utilisée pour générer les graphiques sur les figures 8.3 et 8.4. Les prédictions sont proches des observés sur l'échantillon de validation, prouvant que cette méthode

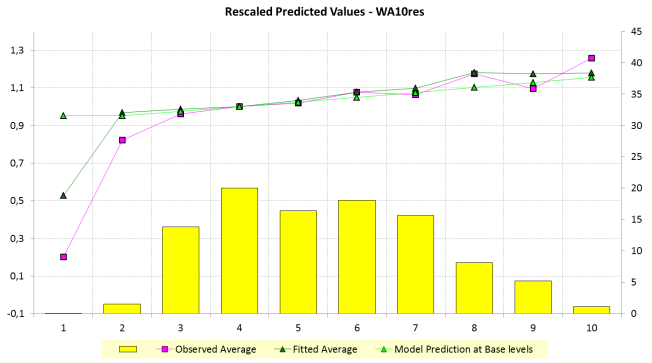


FIGURE 8.3 – WAres sur l'échantillon de détermination du lissage, Méthode 2

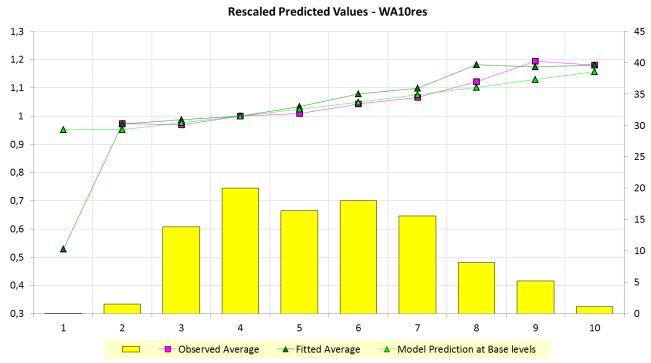


FIGURE 8.4 – WAres sur l'échantillon de validation, Méthode 2

est plus adaptée et accorde plus de poids à ce dernier plutôt qu'à l'échantillon de modélisation. La première classe a été regroupée avec la seconde comme précédemment pour des questions de robustesse au vu du manque d'exposition. Il reste alors à vérifier que les critères de sélection de variables sont satisfaits. Au vu de la tendance et des relativités, il ne semble pas y avoir de problème : il y a bien une vraie distinction entre les classes avec une tendance croissante, le spread étant de 20%. Toutes choses égales par ailleurs, les classes 1 et 2 auront un coefficient de réduction de -5% tandis que les véhicules situés dans la classe la plus élevée auront une majoration de 16%. Le test du Chi2 est également largement satisfait, et le critère de jugement est naturellement positif.

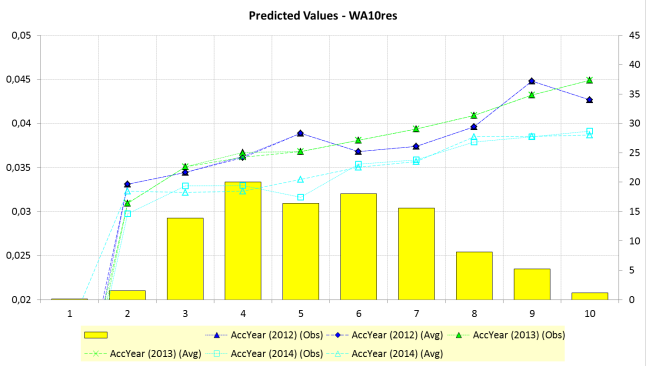


FIGURE 8.5 – Test de robustesse du véhiculier WAres

Enfin, le test de robustesse temporelle est également réussi. En effet nous pouvons voir les observés et les prédictions de l'échantillon de validation pour chacune des années sur la figure 8.5. La tendance est conservée aussi bien pour l'une ou l'autre des valeurs, peu importe les années, et ce sur des données complètement indépendantes du processus de création du véhiculier. La variable passe donc sans problèmes les tests d'inclusion des variables.

Cette variable nous permet donc de bien segmenter le risque au vu du spread élevé entre la classe minimale et la classe maximale. De plus, le coefficient de GINI s'améliore de 5.7% et l'EDR de 11%. Les résultats sur l'échantillon de validation sont également satisfaisants, avec un écart de GINI inférieur à 5%, ce que nous considérons comme acceptable. Les autres mesures statistiques comme l'AIC et le BIC s'améliorent également. Tous ces résultats sont disponibles dans le tableau 8.1.

Les mêmes tests ont été effectués sur les variables QUres et EWres et ont également été satisfaisants. Le coefficient de GINI de la classification WAres sur l'échantillon de modélisation est sensiblement supérieur à ceux des deux autres, pour des écarts par rapport au GINI de l'échantillon de validation quasi équivalent. Parmi les classifications sur des résidus seulement, c'est donc le véhiculier de 9 classes, au lissage 90, classé selon la méthode de Ward et intégré au modèle à l'aide de la Méthode 2 qui est la meilleure (cf. tableau 8.1).

3 Véhiculiers sur prédictions et résidus combinés

Dans cette section, nous allons nous concentrer sur les véhiculiers construits en combinant la prédiction des variables véhicules et les résidus lissés. Ces véhiculiers seront intégrés dans le modèle de fréquence IRSA en suivant la Méthode 1 et la Méthode 2 définies dans la section précédente. Comme évoqué dans le chapitre précédent et sur les figures 7.17d, 7.17e et 7.17f, le signal du véhiculier est en partie déjà capté dans le modèle actuel dans la mesure où les variables des caractéristiques automobiles sont dans le modèle. Une fois enlevées, il n'y a plus de signal (cf. figure 8.6 avec la classification de Ward) et chaque classe du véhiculier a environ les mêmes valeurs prédites (représentées en bleu foncé). Le véhiculier peut alors être testé, permettant également de simplifier fortement le modèle.

La structure du modèle final peut donc être résumée de la façon suivante :

$$\text{Variables non véhicule} + \text{Véhiculier}(\text{Variables véhicules} + \text{résidus})$$

Seuls les résultats de la classification de Ward seront exposés, les résultats des autres classifications seront néamoins disponibles dans la section suivante (cf. tableau 8.1) et les graphiques en Annexes (fig. A.23 à fig. A.28). Tout d'abord, nous enlevons les variables véhicules et intégrons le véhiculier directement dans le modèle, suivant la Méthode 1, traditionnellement utilisée pour le test de variables dans les modèles. Les coefficients sont visibles sur la figure 8.6. Il y a bien la tendance croissante attendue en fonction de la classe de risque, et le spread est de 113%, faisant de cette variable un atout fort discriminant dans le modèle. Nous pouvons constater que l'effet n'est pas le même que lorsque les véhiculiers créés sur résidus seulement étaient testés : l'overfitting est bien moins important sur l'échantillon de validation, comme on peut le voir sur la figure 8.7. Cependant, la prédiction s'écarte de l'observé dans les extrêmes.

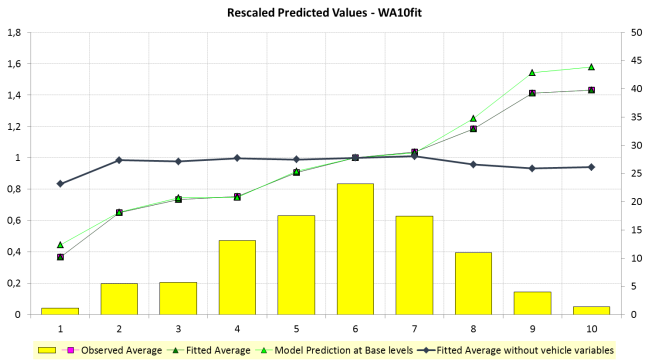


FIGURE 8.6 – WAfit sur l'échantillon de modélisation, Méthode 1

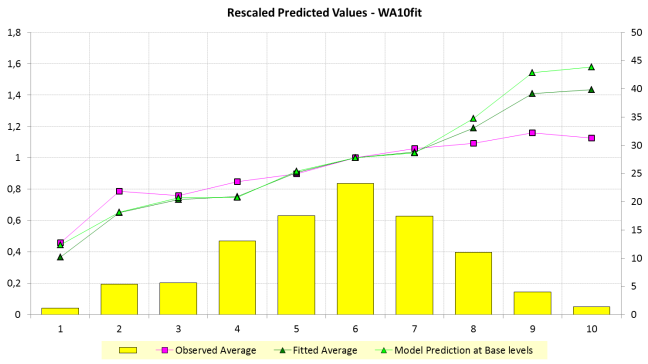


FIGURE 8.7 – WAfit sur l'échantillon de validation, Méthode 1

Le coefficient de GINI augmente de 7.3% par rapport au modèle initial, passant de 25.254% à 27.117% sur l'échantillon de modélisation. Cependant lorsqu'on se concentre sur l'échantillon de validation, il diminue (légèrement, moins de 1%). Ceci n'est donc pas satifaisant. De même l'écart entre les coefficients de GINI du nouveau modèle est de 8% entre l'échantillon de modélisation et l'échantillon de validation, ce qui est insatisfaisant. L'EDR diminue également. La Méthode 1, quoique moins impactante que pour les véhiculiers sur les résidus seuls, dégrade donc les indicateurs et n'est à priori pas la méthode à utiliser pour l'intégration du véhiculier.

Nous pouvons alors tester la Méthode 2 pour intégrer la classification de véhicules dans le modèle. Comme dans le chapitre précédent, cette méthode permet d'avoir des prédictions plus semblables à l'échantillon de validation et évite d'avoir un sur-apprentissage sur l'échantillon de modélisation. Nous pouvons voir sur la figure 8.8 que la variable a une tendance croissante, et robuste sur l'échantillon de validation visible sur la figure 8.9. Le spread est légèrement moins important que celui obtenu avec la Méthode 1 (82% vs 113%, ce qui reste tout de même extrêmement discriminant). Toutes choses égales par ailleurs, la classe 1 aura un coefficient de réduction de -50% tandis que les véhicules appartenant à la classe la plus élevée auront une majoration de 32%. Il n'est d'ailleurs pas nécessaire de regrouper les premières classes car elles contiennent suffisamment de volume. Ces coefficients sont bien plus importants que ceux obtenus avec les véhiculiers créés à partir des résidus seulement. En effet, dans le premier cas, les variables liées au véhicules du modèle, à savoir la carrosserie (spread de 14%), l'énergie (spread de -11%), le segment (spread de 30%), la vitesse (spread de -20%) et le poids (spread de -4%). Il faut alors compenser cette perte d'information via les autres variables avec une discrimination plus importante de la classification.

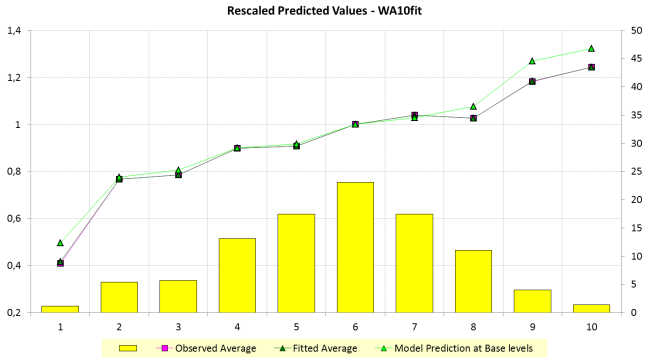


FIGURE 8.8 – WAfit sur l'échantillon de détermination du lissage, Méthode 2

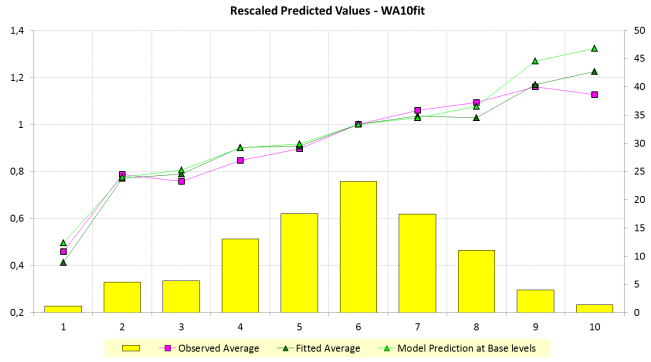


FIGURE 8.9 – WAfit sur l'échantillon de validation, Méthode 2

Cette variable nous permet donc de bien segmenter le risque au vu du spread important et de sa tendance. Le test du Chi2 ainsi que le test de robustesse sont également satisfaits (fig. A.29 en Annexe). Quant aux indicateurs statistiques, les résultats sont très corrects : le coefficient de GINI s'améliore de 5% par rapport à celui du modèle initial, et l'EDR de 9.7%. De plus, l'écart entre le GINI de l'échantillon de modélisation et de l'échantillon de validation n'est que de 0.5%, ce qui est comparable avec le modèle initial.

Comme pour la section précédente, l'ensemble de ces tests a été effectué sur les classifications EWfit et QUfit. Cependant c'est la classification de Ward qui a les meilleurs résultats. Parmi la classification sur la combinaison entre prédiction des variables véhicules et résidus, c'est le véhiculier de 10 classes au lissage 90 classé selon la méthode de Ward et intégré à l'aide de la Méthode 2 qui donne les meilleurs résultats.

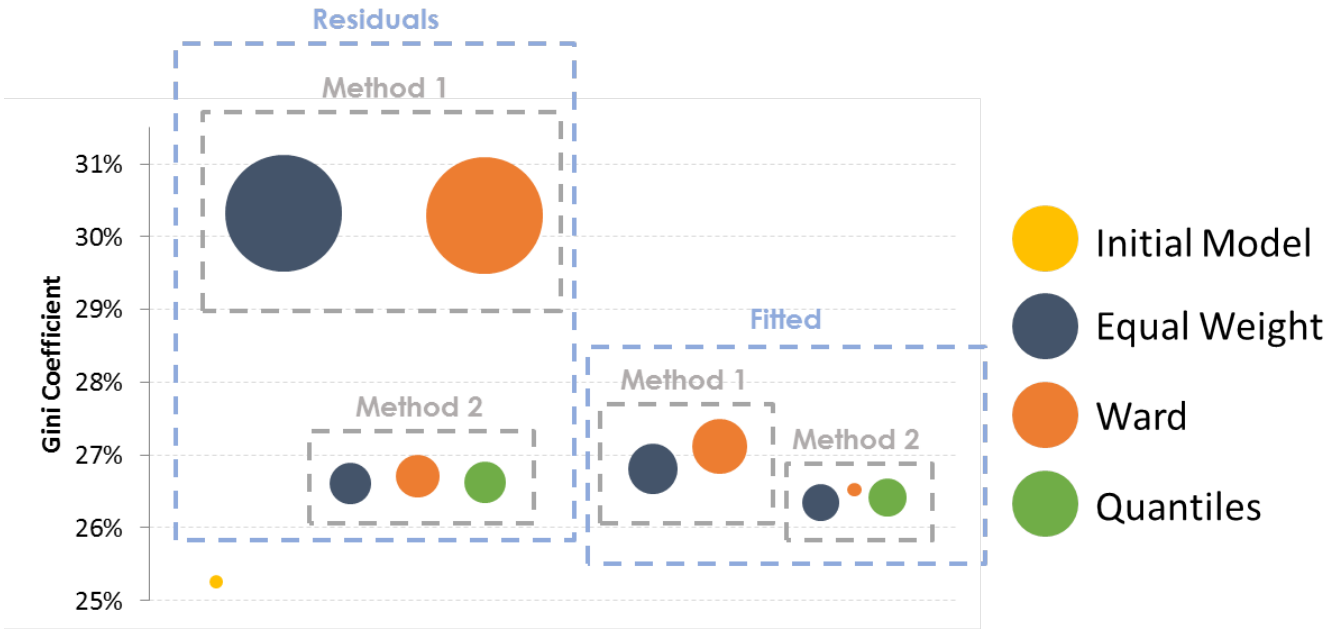


FIGURE 8.10 – Représentation graphique des coefficients de GINI

4 Choix de la classification optimale

Dans les sections précédentes, nous avons testé 6 classifications différentes :
- 3 types de classification : Equal Weight, Ward et Quantiles ;
- 2 types de quantité classée : résidus seulement et combinaison entre les prédictions de variables véhicules et les résidus.

De plus, ces 6 véhiculiers ont été introduits de deux façons différentes dans le modèle de fréquence IRSA :
- Méthode 1 qui consiste à rajouter le véhiculier directement dans l'échantillon de modélisation (80% des données) ;
- Méthode 2 avec laquelle les coefficients de la variable sont déterminés à l'aide de l'échantillon de détermination du lissage (20% parmi l'échantillon de modélisation, et n'ayant pas servi à la construction de la classification).

Nous avons donc 12 classifications possibles (10 en réalité car la Méthode 1 n'a pas été testée sur les classifications par quantiles au vu des résultats sur les classification de Ward et Equal Weight) parmi lesquelles nous ne souhaitons retenir que celles dont les résultats sont optimaux. L'ensemble de ces résultats est disponible dans le tableau 8.1. Cependant pour plus de facilité d'interprétation, nous avons représenté graphiquement les coefficients de GINI sur la figure 8.10.

L'axe des coordonnées représente le coefficient de GINI de l'échantillon de modélisation tandis que la taille de la bulle correspond à l'écart entre ce dernier et celui de l'échantillon de validation. La bulle en bas à gauche, avec un coefficient de 25.254%, est celle du modèle initial. Nous pouvons clairement voir les résultats à priori concluants de la Méthod 1 appliquée sur les véhiculiers sur les résidus seulement au vu de leur coefficient particulièrement élevé. Cependant la taille de la bulle est très grande, ce qui peut se traduire par un sur-apprentissage sur l'échantillon de modélisation. De plus, pour terminer de décrédibiliser cette approche, le coefficient sur l'échantillon de validation décroît par rapport au modèle initial. Les résultats de la Méthode 1 sont similaires pour la classi-

fication des "fitted", combinaison entre prédictions et résidus, même si les effets sont bien moins importants.

La méthode ayant les meilleurs résultats est donc la Méthode 2, avec des améliorations de coefficient de GINI ainsi que des écarts entre modélisation et validation acceptables. Comme indiqué dans les sections précédentes, c'est la méthode de Ward qui donne les meilleurs résultats, que la classification se fasse sur les résidus seulement ou sur la combinaison. La première donne un meilleur coefficient de GINI sur l'échantillon de modélisation, et l'écart entre les coefficients des échantillon de modélisation et échantillon de validation est acceptable. La seconde permet néanmoins de réduire fortement cet écart, au détriment d'une légère perte du coefficient de GINI sur l'échantillon de modélisation. De plus, son coefficient de GINI sur l'échantillon de validation est meilleur que l'autre méthode.

Cette analyse sur le coefficient de GINI permet donc de ne retenir que deux véhiculiers, mais ne réussit pas à les démarquer pour ne retenir que la solution optimale. Nous pouvons alors regarder l'EDR. Le véhiculier sur les résidus a un EDR de respectivement -2.79% et -2.58% sur l'échantillon de modélisation et de validation. Celui sur la combinaison des prédictions et résidus a quant à lui un EDR de respectivement -2.75% et -2.57% sur l'échantillon de modélisation et l'échantillon de validation. Même si les différences sont très faibles, la solution optimale est la classification sur les résidus seulements.

Il faut noter que dans le cas de notre étude, nous pouvons facilement intégrer l'une ou l'autre des classifications dans la mesure où nous disposons déjà de la majorité des caractéristiques de véhicule dans nos systèmes informatiques. Dans le cas d'une classification géographique, il est plus difficile de récupérer les données du code postal, code INSEE ou code iris. De plus, les variables comme la densité des ville dépend du temps, tandis que les variables véhicules sont figées pour un véhicule donné. Dans ce cas, la classification de la combinaison des prédictions et des résidus aurait été retenue, permettant ainsi de retirer les variables potentiellement incluses dans les modèles, telles la densité de population ou la proportion de retraités d'une commune.

Échantillon	Modèle	Modèle	Méthode 1		Méthode 2			Méthode 1		Méthode 2		
de modélisation	sans variables	initial	EWres	WAres	EWres	WAres	QUres	EWfit	WAfit	EWfit	WAfit	QUfit
Deviance	473 225	461 345	456 459	456 496	460 119	460 033	460 110	459 894	459 600	460 338	460 188	460 293
AIC	479 833	468 189	463 320	463 357	466 964	466 877	466 955	466 712	466 435	467 155	466 797	466 901
BIC	479 847	469 776	465 014	465 050	468 551	468 464	468 542	468 126	467 955	468 555	466 810	466 915
AICc	479 833	468 189	463 320	463 357	466 964	466 877	466 955	466 712	466 435	467 155	466 797	466 901
GINI		25,254%	30,324%	30,297%	26,613%	26,711%	26,624%	26,81%	27,117%	26,35%	26,521%	26,418%
GINI Saturated		25,571%	30,705%	30,677%	26,947%	27,046%	26,957%	27,146%	27,457%	26,68%	26,853%	26,749%
EDR	0%	-2,51%	-3,54%	-3,54%	-2,8%	-2,8%	-2,8%	-2,8%	-2,9%	-2,7%	-2,8%	-2,7%
Coef min			-34%	-19%	-7%	-5%	-11%	-39%	-56%	-24%	-50%	-23%
Coef max			60%	117%	6%	16%	0%	15%	58%	8%	32%	14%
Spread			93%	136%	13%	20%	11%	54%	113%	32%	82%	37%

Échantillon	Modèle	Modèle	Méthode 1		Méthode 2			Méthode 1		Méthode 2		
de validation	sans variables	initial	EWres	WAres	EWres	WAres	QUres	EWfit	WAfit	EWfit	WAfit	QUfit
Deviance	118 910	115 865	116 812	116 810	115 847	115 848	115 849	115 931	115 950	115 862	115 857	115 858
AIC	120 575	117 530	118 477	118 475	117 512	117 513	117 514	117 596	117 615	117 527	117 522	117 523
BIC	120 589	117 543	118 490	118 489	117 526	117 526	117 528	117 609	117 629	117 540	117 535	117 536
AICc	120 575	117 530	118 477	118 475	117 512	117 513	117 514	117 596	117 615	117 527	117 522	117 523
GINI		25,375%	22,22%	22,21%	25,444%	25,442%	25,436%	25,165%	25,066%	25,412%	26,386%	25,412%
GINI Saturated		25,698%	22,503%	22,491%	25,768%	25,766%	25,76%	25,486%	25,384%	25,736%	25,708%	25,735%
EDR	0%	-2,56%	-1,765%	-1,766%	-2,6%	-2,6%	-2,6%	-2,5%	-2,5%	-2,6%	-2,6%	-2,6%

TABLE 8.1 – Indicateurs sur les échantillons de modélisation et de validation

Le véhiculier finalement retenu est donc celui sur les résidus seulement avec un lissage à 90,

utilisant la classification de Ward en 10 classes, dont les coefficients du modèle ont été déterminés sur l'échantillon de détermination du lissage.

5 Validation du modèle

Nous avons décidé de retenir la classification WAres dans le modèle. Il faut néanmoins vérifier que le modèle est juste via les outils présentés dans la section 6.4.

Distribution des résidus
Nous avons tracé les "crunched residuals" (résidus moyennés par groupes) à l'aide du logiciel Emblem. Le graphique est disponible sur la figure 8.11. Nous pouvons voir que les résidus sont distribués aléatoirement autour de 0 indépendamment de la valeur des prédictions. Ce test valide donc le modèle avec la variable WAres.

Déviance de Cook
Quant aux valeurs extrêmes de la déviance de Cook, nous constatons sur la figure 8.12 la même chose que dans la section 6.4, à savoir la présence d'un point potentiellement aberrant. Cependant la valeur est moins élevée que les 2% préconisés par AXA. Nous pouvons alors considérer que le modèle est valide. Pour plus d'explications au sujet de l'utilité de ce graphique, le lecteur est invité à se référer à la section 6.4.

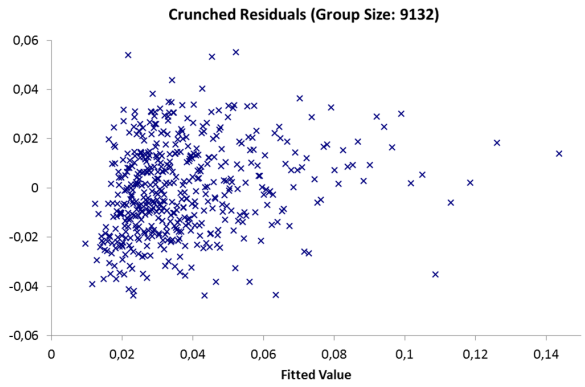


FIGURE 8.11 – Crunched Residuals

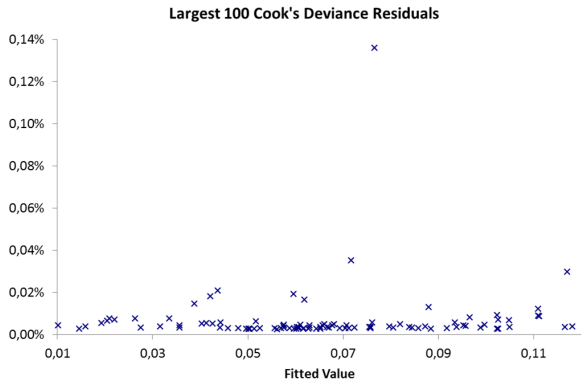


FIGURE 8.12 – 100 plus grandes valeurs de la Déviance de Cook

Courbe d'ajustement
Enfin, nous avons représenté sur les figures 8.13 et 8.14 les courbes d'ajustement (cf. section 6.4) à l'aide desquelles nous pouvons évaluer le caractère prédictif du modèle. Les courbes devraient être relativement proches de celles du modèle initial visible sur les figures 6.9 et 6.10. C'est bien le résultat que nous retrouvons. La courbe bleue des prédictions est très proche de la courbe rouge des observés, permettant ainsi de valider le modèle avec la nouvelle variable.

Le modèle final, une fois la variable WAres rajoutée, est donc validé. C'est ce modèle non contraint qui permettra de calculer la fréquence responsabilité civile matérielle des sinistres conventionnés la plus juste. Il ne sera néanmoins pas mis en production. En effet, les contraintes légales doivent être rajoutées avant de pouvoir offrir un tarif aux clients.

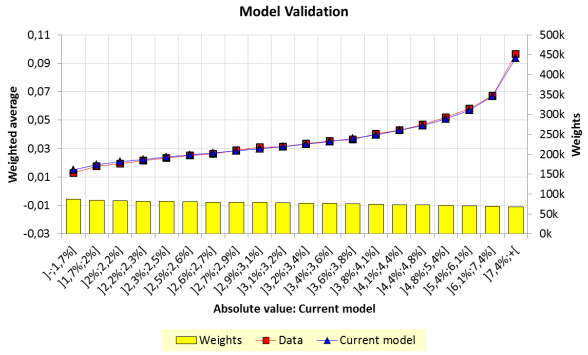


FIGURE 8.13 – Courbe d'ajustement sur l'échantillon de modélisation

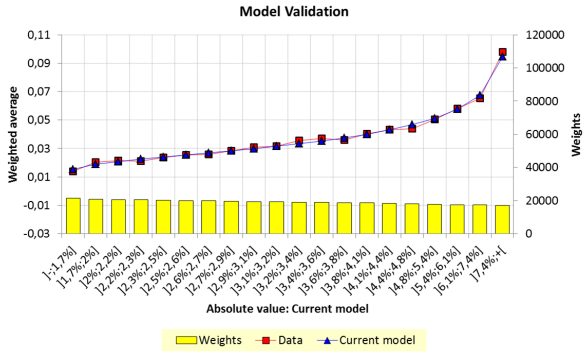


FIGURE 8.14 – Courbe d'ajustement sur l'échantillon de validation

6 Limites et points d'amélioration

La principale difficulté dans cette étude a été la création de multiples classifications afin de déterminer la plus optimale. D'autres classifications auraient pu être créées mais le périmètre a été restreint pour ne pas alourdir l'étude. De plus, un point d'attention a été mis sur l'automatisation du processus de création du véhiculier afin de servir de base pour les différentes entités d'AXA proposant de l'assurance automobile. Aujourd'hui appliqué en France dans l'entité traditionnelle et directe, en Italie et en Corée du Sud, le code est efficace mais un traitement préalable des données est nécessaire, chaque entité ayant ses particularités.

La classification de véhicules construite au cours de ce mémoire est une classification pour un modèle non contraint, i.e. sans les contraintes d'implémentation informatiques ni légales. Elle est donc adaptée à un modèle spécifique qui ne sera pas mis en production, mais permet néanmoins d'avoir une vision du risque suffisamment fine pour estimer les indicateurs techniques ou faire des analyses approfondies de sinistralité. Le processus de création de véhiculier semble trop chronophage pour pouvoir en créer un nouveau spécifique au modèle contraint. La plupart du temps il est d'usage d'utiliser le véhiculier d'un modèle non contraint directement dans le modèle contraint. Cette solution n'est pas idéale mais permet tout de même d'obtenir de meilleurs résultats dans un modèle contraint.

La mise en place d'un code suffisamment générique ainsi que l'allègement de décisions humaines dans le processus de création de la classification permet de réduire le temps consacré à l'étude. Mais il reste cependant assez lourd. D'autres méthodes de machine learning plus automatisées, comme les *random forests*, permettraient de générer des classifications plus rapidement. Elles seraient toutefois plus difficiles à comprendre.

Il aurait été possible de tester un lissage par distance directement sur le nuage de points de véhicules. Cette méthode n'a pas été privilégiée dans la mesure où les classifications géographiques se basent pour la plupart sur le lissage par voisinage.

Cette étude se consacre à la création d'une classification de véhicule pour la garantie responsabilité civile matérielle, plus précisément sur la fréquence des sinistres conventionnés IRSA. Ce périmètre a été choisi au vu du volume d'années polices présentes ainsi que du nombre de sinistre, permettant de créer une méthodologie sur un échantillon suffisant pour être considéré comme robuste. Il pourrait être judicieux d'appliquer cette méthode pour coûts moyens des garanties vol et dommages tous accidents, dans la mesure où elles sont directement liées au véhicule (la clas-

sification en responsabilité civile matérielle tentant plutôt de refléter un signal supplémentaire du comportement de l'assuré en fonction du véhicule qu'il choisi).

Enfin, la principale limite de cette classification concerne les nouveaux véhicules, i.e. ceux qui ne sont pas encore commercialisés ou non présents lors de l'étude dans la base de données SRA. Il est alors nécessaire de créer un nouveau processus afin de classer ces nouveaux véhicules. Ceci nécessite également de nouveaux développements informatiques lorsque la classification est mise en production, i.e. fait partie des variables tarifaires. Aujourd'hui, le processus consiste à rapprocher les nouveaux véhicules avec des véhicules existant en fonction de leurs caractéristiques utilisées dans l'analyse factorielle de données mixtes, et de leur appliquer la classe de véhicule correspondante.

Conclusion

L'étude réalisée pour ce mémoire a eu des objectifs multiples. Le premier était de mettre à jour le modèle de tarification de la garantie Responsabilité Civile Matérielle avec des données plus récentes permettant de prendre en compte toutes les modifications de marché et les tendances de sinistralité plus actuelle. Ce fût également l'occasion de modifier la structure de la modélisation de la garantie à l'aide d'une segmentation en sous-périls - sinistres relevant de la convention IRSA vs sinistres relevant du droit commun ou indemnisés au coût réel. Lorsque le volume est suffisant, cette segmentation s'avère efficace et ajuste au mieux le risque des assurés. Nous avons pu mesurer la valeur ajoutée de cette innovation à l'aide d'indicateurs statistiques par rapport à l'ancien modèle. Le dernier objectif était de créer une classification de véhicule alternative à celle proposée par la SRA plus adaptée au portefeuille de Direct Assurance. Nous avons pu développer une méthodologie la plus générique possible afin qu'elle puisse être implémentée sur d'autres garanties et d'autres portefeuilles. Nous avons pu générer plusieurs classifications et retenir celle donnant les meilleurs résultats statistiques.

Au cours de l'étude, nous avons dû faire face à plusieurs problématiques. Tout d'abord, la représentation spatiale des véhicules : les outils d'analyse de données se sont prêtés à cet exercice, et dans un soucis d'automatisation du processus, l'AFDM a été choisie au détriment de l'ACP, cette dernière requérant un traitement préalable des données (création de classes pour les variables quantitatives). Ensuite, le caractère convexe de la triangulation de Delaunay qui permet de créer une relation de voisinage entre les véhicules nous a poussé à faire un retraitement de la table de contiguïté avec des règles définies. Enfin, l'intégration directe des véhiculiers dans les modèles linéaires généralisés a mis en évidence un effet de sur-apprentissage. Nous avons alors dû revoir la méthode d'ajout des variables dans les modèles afin de s'affranchir de ce phénomène.

Ce modèle final ainsi que la classification des véhicules font désormais partie du modèle de tarification en production de Direct Assurance. La segmentation des périls peut s'avérer intéressante pour d'autres garanties, comme en vol, où les risques de vol total et de vol partiel sont habituellement groupés tout en ayant des caractéristiques différentes. La méthodologie de classification de véhicule peut également être reproduite sur diverses garanties. A ce jour, les résultats sur la garantie vol n'ont montré qu'une amélioration minime.

La mise en production de cette classification de véhicules sur la garantie RCM pose cependant une question de maintenance. En effet, il faut penser à mettre à jour le véhiculier de façon régulière. Par ailleurs, il est nécessaire de développer un algorithme permettant de rattacher de nouveaux véhicules commercialisés depuis l'étude à une classe donnée. Il existe aujourd'hui de nombreuses méthodes de machine learning permettant de créer des classifications de façon plus automatisées. Cependant, les résultats de ces classifications est souvent difficile à interpréter. Il pourrait être intéressant de comparer ces nouvelles méthodes au processus développé dans ce mémoire.

Bibliographie

[1] D. ANDERSON et al. *A practitioner's Guide to Generalized Linear Models*. 2017.

[2] D. ANDERSON et al. *General Insurance Premium Rating Issues Working Party*. 2007.

[3] M. BREVILLIERS. "Construction de la triangulation de Delaunay de segments par un algorithme de flip". Thèse de doct. Université de Haute Alsace, 2008.

[4] R.E. BRUBAKER. "Geographic Rating of Individual Risk Transfer Costs without Territorial Boundaries". In : *Casualty Actuarial Society Forum 1996 Winter Page(s)* (1996), p. 97–127.

[5] S. CHRISTOPHERSON et D.L. WERLAND. "Using a Geographic Information System to Identify Territory Boundaries". In : *Casualty Actuarial Society Forum 1996 Winter Page(s)* (1996), p. 191–211.

[6] D. CLOT. *Cours*. Institut de Sciences Financières et d'Assurance.

[7] P. CONDEMINE. "Etude de la sur-sinistralité des conducteurs âgés". Mém.de mast. ENSAE, 2014.

[8] *Convention d'Indemnisation Directe de l'Assuré et de Recours entre Sociétés d'Assurance Automobile*. 2014.

[9] A.B. DUFOUR. *Cours et TD*. Institut de Sciences Financières et d'Assurance. DOI : `https://pbil.univ-lyon1.fr/R/`.

[10] E. DUPONT. *Convention d'Indemnisation Directe de l'Assuré et de Recours entre Sociétés d'Assurance Automobile*. IBRS, Risque pour les Jeunes Conducteurs dans la Circulation, 2012.

[11] B. ESCOFIER. "Traitement simultané de variables quantitatives et qualitatives en analyse factorielle". In : *Les cahiers de l'analyse des données* 4.2 (1979), p. 137–146.

[12] G. GONNET. "Etude de la tarification et de la segmentation en assurance automobile". Mém.de mast. ISFA, 2010.

[13] E. MICU. *Territorial Ratemaking*. EagleEye Analytics, 2012.

[14] J. PAGES. "Analyse factorielle de données mixtes". In : *Revue de statistique appliquée* 52.4 (2004), p. 93–111.

[15] J. PAGES. "Analyse factorielle de données mixtes : principe et exemple d'application". In : *Laboratoire de mathématiques appliquées - Agrocampus Rennes* (2004).

[16] V. PILLAUD. "Sur le diagramme de Voronoi et la triangulation de Delaunay d'un ensemble de points dans un revêtement du plan euclidien". In : (2006).

[17] R. RAKOTOMALALA. *Analyse des Correspondances Multiples - Principe et pratique de l'ACM*. Cours Université Lumière Lyon 2. DOI : `https://eric.univ-lyon2.fr/~ricco/cours/`.

[18] R. RAKOTOMALALA. *Analyse en Composantes Principales - Principe et pratique de l'ACP*. Cours Université Lumière Lyon 2. DOI : `https://eric.univ-lyon2.fr/~ricco/cours/`.

[19] G. SAPORTA. "Simultaneous analysis of qualitative and quantitative data". In : *Atti della XXXV riunione scientifica ; società italiana di statistica* (1990), p. 63–73.

[20] F. SLOOTMANS, E. DUPONT et P. SILVERANS. *Analyse des facteurs de risque pour les conducteurs de 18 à 24 ans sur la base d'une enquête concernant leur implication dans les accidents*. IBRS, Risque pour les Jeunes Conducteurs dans la Circulation, 2011.

[21] M. TENENHAUS. *Statistique : Méthodes pour décrire, expliquer et prévoir*. Dunod, 2006.

[22] O.A. VASECHKO, M. GRUN-REHOMME et N. BENLAGHA. "Modélisation de la fréquence des sinistres en assurance automobile". In : *Bulletin Français d'Actuariat* 9.18 (2009).

[23] R. VERRALL et M. BOSKOV. "Premium Rating by Geographic Area Using Spatial Models". In : *Cambridge ASTIN Bulletin* 24.1 (1993), p. 131–143.

Annexe A

Figures

Figure évoquée dans la section 1.3 Étude du portefeuille.

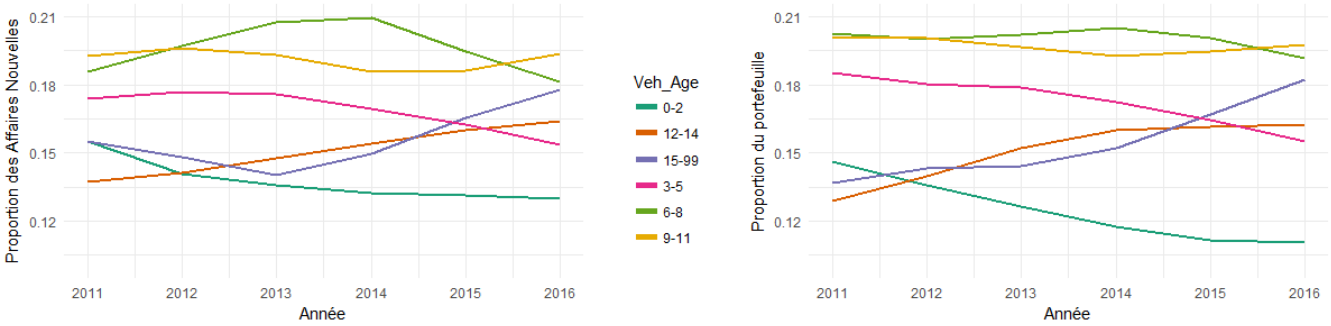


FIGURE A.1 – Proportion des âges de véhicule dans le temps

Figures évoquées dans la section 6.1 Détermination du périmètre de l'étude.

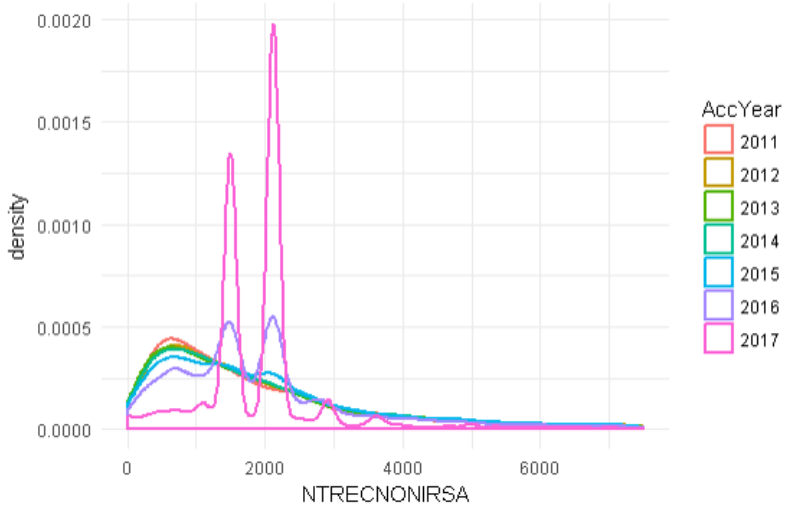


FIGURE A.2 – Distribution du coût moyen RCM par année de survenance

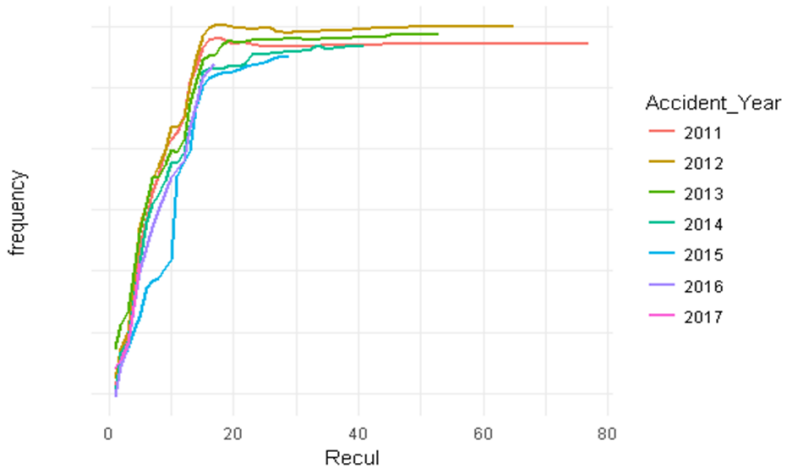


FIGURE A.3 – Déroulé de la fréquence YTD RCM non IRSA par année de survenance

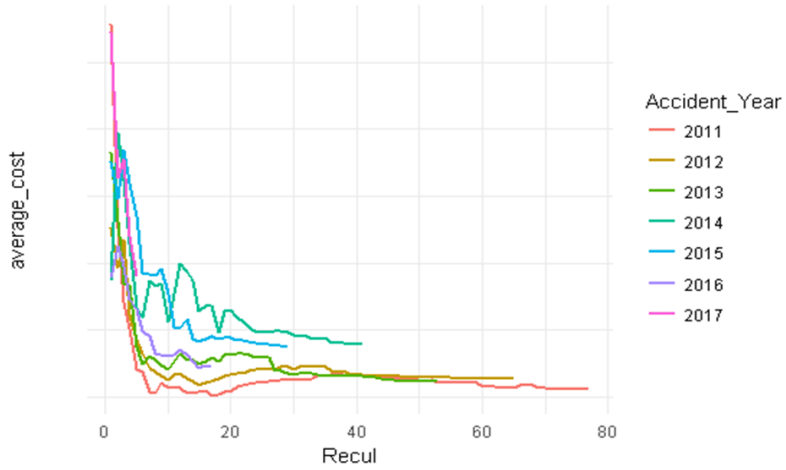


FIGURE A.4 – Déroulé du coût moyen YTD RCM IRSA par année de survenance

Figures évoquées dans la section 6.4 Comparaison des modèles.

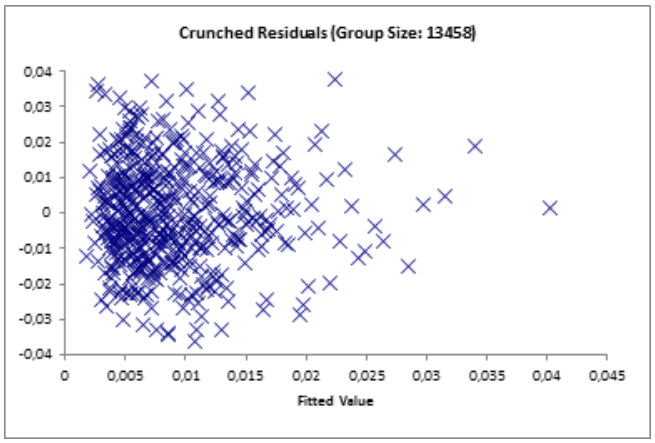


FIGURE A.5 – Modèle Fréquence Non IRSA : Crunched Residuals

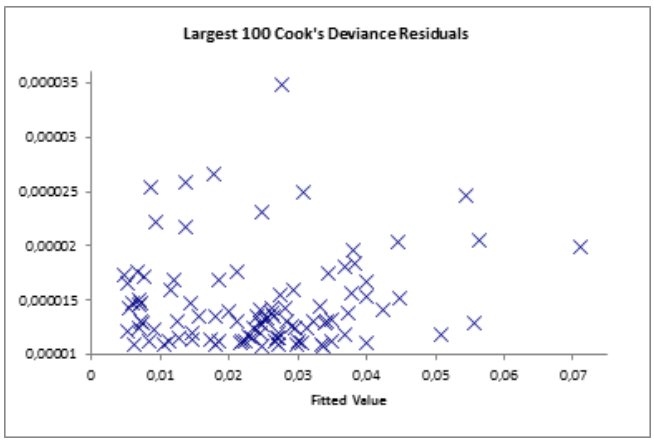


FIGURE A.6 – Modèle Fréquence Non IRSA : 100 plus grandes valeurs de la Déviance de Cook

FIGURE A.7 – Modèle Coût Moyen Non IRSA : Crunched Residuals

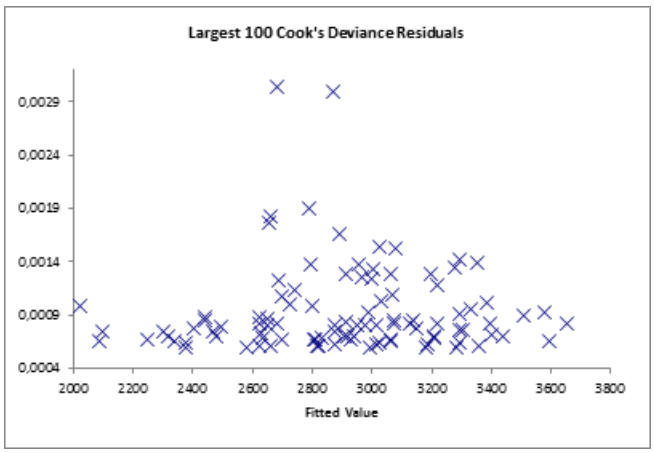


FIGURE A.8 – Modèle Coût Moyen Non IRSA : 100 plus grandes valeurs de la Déviance de Cook

Figures évoquées dans la section 7.1 Résultats de l'AFDM.

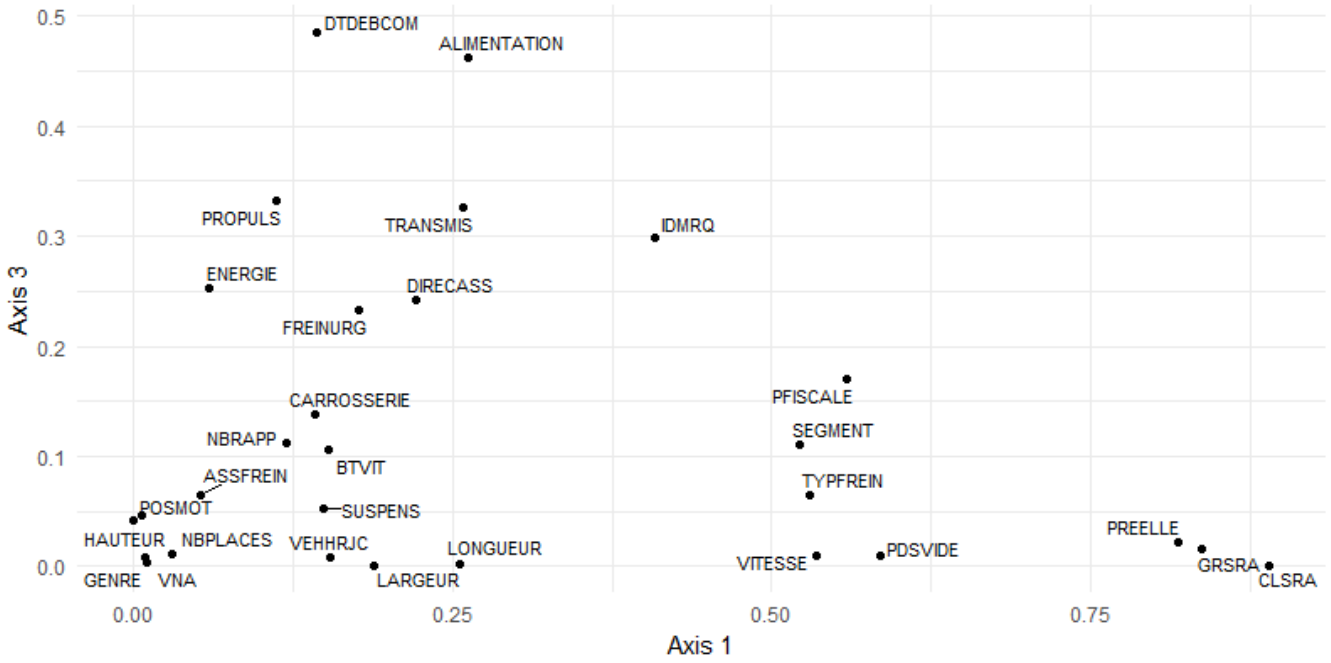


FIGURE A.9 – Nuage des variables sur l'axe 1 et l'axe 3

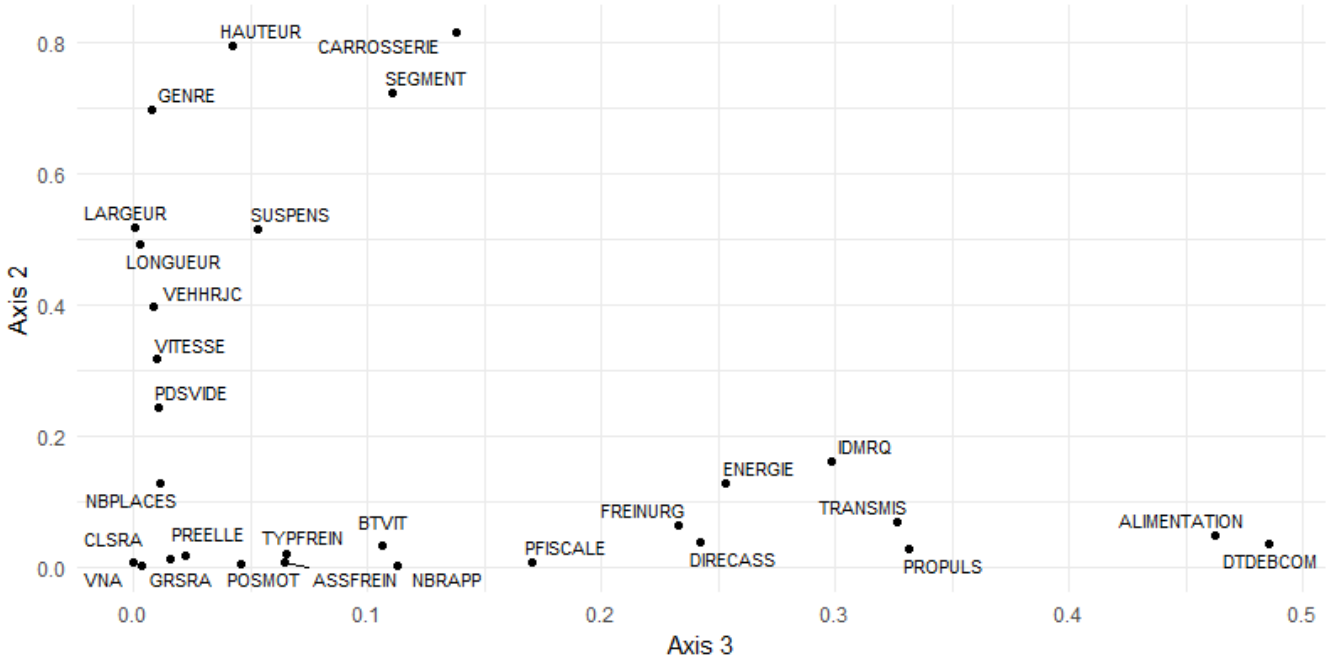


FIGURE A.10 – Nuage des variables sur l'axe 2 et l'axe 3

Axis	inertia	%cum	Axis	inertia	%cum	Axis	inertia	%cum	Axis	inertia	%cum
1	8,1854	7,58%	31	1,0570	56,04%	61	0,8897	83,40%	91	0,1621	99,20%
2	6,3057	13,42%	32	1,0453	57,01%	62	0,8689	84,20%	92	0,1536	99,34%
3	3,6325	16,78%	33	1,0423	57,98%	63	0,8636	85,00%	93	0,1211	99,45%
4	2,6135	19,20%	34	1,0302	58,93%	64	0,8487	85,79%	94	0,1126	99,55%
5	2,4804	21,50%	35	1,0217	59,88%	65	0,8425	86,57%	95	0,1009	99,65%
6	2,1249	23,47%	36	1,0185	60,82%	66	0,7909	87,30%	96	0,0880	99,73%
7	2,0736	25,39%	37	1,0172	61,76%	67	0,7603	88,00%	97	0,0802	99,80%
8	2,0638	27,30%	38	1,0109	62,70%	68	0,7531	88,70%	98	0,0744	99,87%
9	1,9977	29,15%	39	1,0089	63,63%	69	0,7446	89,39%	99	0,0599	99,93%
10	1,8790	30,89%	40	1,0048	64,56%	70	0,7331	90,07%	100	0,0331	99,96%
11	1,7926	32,55%	41	1,0042	65,49%	71	0,7146	90,73%	101	0,0261	99,98%
12	1,7075	34,13%	42	1,0035	66,42%	72	0,6918	91,37%	102	0,0176	100,00%
13	1,6753	35,68%	43	1,0023	67,35%	73	0,6453	91,97%	103	0,0006	100,00%
14	1,5218	37,09%	44	1,0011	68,28%	74	0,6359	92,56%			
15	1,4468	38,43%	45	1,0007	69,20%	75	0,6291	93,14%			
16	1,4017	39,72%	46	1,0003	70,13%	76	0,5908	93,69%			
17	1,3292	40,96%	47	0,9998	71,05%	77	0,5688	94,22%			
18	1,2971	42,16%	48	0,9954	71,98%	78	0,5557	94,73%			
19	1,2709	43,33%	49	0,9905	72,89%	79	0,5425	95,23%			
20	1,2564	44,50%	50	0,9881	73,81%	80	0,5302	95,72%			
21	1,2317	45,64%	51	0,9867	74,72%	81	0,5145	96,20%			
22	1,2024	46,75%	52	0,9829	75,63%	82	0,5051	96,67%			
23	1,1867	47,85%	53	0,9694	76,53%	83	0,4433	97,08%			
24	1,1709	48,93%	54	0,9639	77,42%	84	0,4000	97,45%			
25	1,1457	49,99%	55	0,9560	78,31%	85	0,3585	97,78%			
26	1,1119	51,02%	56	0,9394	79,18%	86	0,3446	98,10%			
27	1,1086	52,05%	57	0,9302	80,04%	87	0,3144	98,39%			
28	1,1007	53,07%	58	0,9268	80,90%	88	0,2756	98,65%			
29	1,0828	54,07%	59	0,9141	81,74%	89	0,2333	98,86%			
30	1,0727	55,06%	60	0,8984	82,57%	90	0,1997	99,05%			

FIGURE A.11 – Tableau de l'inertie totale et cumulée portée par les axes de l'AFDM

	Comp1	Comp2	Comp3
sra_PFISCALE	-0,7473	-0,0779	-0,4126
sra_NBPLACES	-0,1731	0,3558	0,1057
sra_NBRAPP	-0,3458	-0,0155	0,3361
sra_GRSRA	-0,9150	0,1035	-0,1244
sra_CLSRA	-0,9435	-0,0830	-0,0022
sra_DTDEBCOM	-0,3785	0,1873	0,6969
sra_VITESSE	-0,7313	0,5620	-0,0984
sra_PREELLE	-0,9047	0,1271	-0,1482
sra_LONGUEUR	-0,5052	-0,7006	-0,0533
sra_LARGEUR	-0,4333	-0,7187	0,0276
sra_HAUTEUR	0,0166	-0,8907	0,2045
sra_PDSVIDE	-0,7647	-0,4925	0,1018

FIGURE A.12 – Table des corrélations de l'AFDM

FIGURE A.13 – Nuage des individus selon le groupe SRA

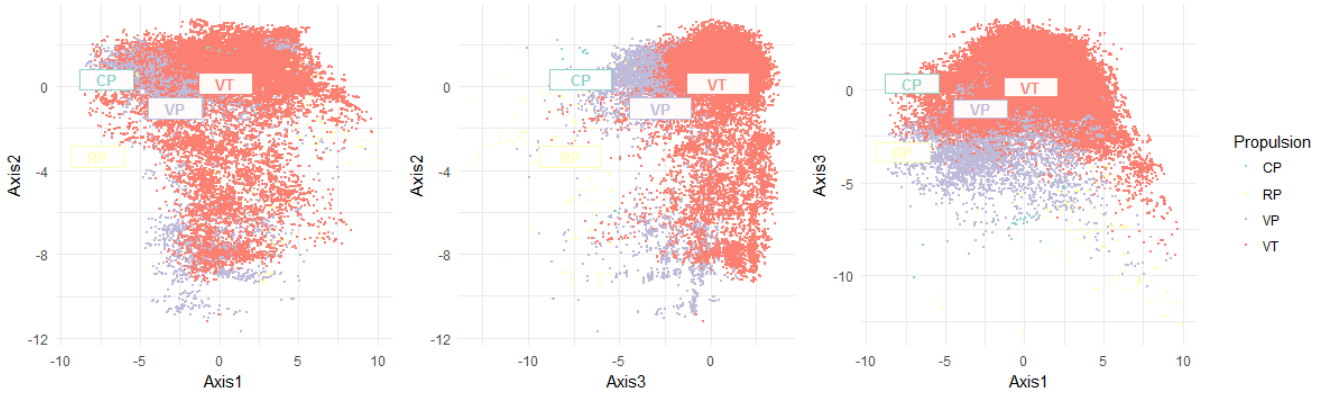


FIGURE A.14 – Nuage des individus selon la propulsion du véhicule

Figures évoquées dans la section 7.2 Etude de la table de contiguïté.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	HC
A	1 301	539	448	328	346	240	180	150	64	32	21	15	7	0	1	0	0	0	0	0	0	0	0
B	539	537	714	460	368	265	214	132	55	17	5	6	1	0	2	0	0	0	0	0	0	0	0
C	448	714	1 573	1 541	936	517	354	278	108	30	21	9	3	0	1	1	0	0	0	0	0	0	0
D	328	460	1 541	2 875	2 881	1 558	797	465	186	62	30	10	2	0	3	0	0	0	0	0	0	0	0
E	346	368	936	2 881	4 781	4 448	1 943	960	459	168	39	16	1	3	1	0	0	0	0	0	0	0	0
F	240	265	517	1 558	4 448	7 349	5 953	2 621	1 183	466	122	33	7	4	4	0	1	0	0	0	0	0	0
G	180	214	354	797	1 943	5 953	9 563	7 929	3 404	1 281	351	96	29	9	12	3	3	0	0	0	0	0	0
H	150	132	278	465	960	2 621	7 929	12 342	8 718	3 276	1 100	430	109	42	14	5	2	0	0	0	0	0	0
I	64	55	108	186	459	1 183	3 404	8 718	13 466	9 560	3 956	1 449	378	134	19	5	3	0	0	0	0	0	0
J	32	17	30	62	168	466	1 281	3 276	9 560	14 437	11 385	4 300	1 498	477	81	28	12	2	0	0	0	0	0
K	21	5	21	30	39	122	351	1 100	3 956	11 385	18 061	12 468	4 850	1 637	477	75	33	7	1	0	0	0	0
L	15	6	9	10	16	33	96	430	1 449	4 300	12 468	20 303	15 233	5 316	1 865	472	151	19	6	0	0	0	0
M	7	1	3	2	1	7	29	109	378	1 498	4 850	15 233	23 332	15 085	5 575	1 660	594	138	47	9	0	0	0
N	0	0	0	0	3	4	9	42	134	477	1 637	5 316	15 085	20 365	13 087	4 716	1 749	538	122	12	0	2	0
O	1	2	1	3	1	4	12	14	19	81	477	1 865	5 575	13 087	15 584	9 925	3 855	1 313	350	50	0	0	0
P	0	0	1	0	0	0	3	5	5	28	75	472	1 660	4 716	9 925	11 158	7 223	2 841	754	177	8	0	1
Q	0	0	0	0	0	1	3	2	3	12	33	151	594	1 749	3 855	7 223	8 631	5 678	1 761	437	5	3	1
R	0	0	0	0	0	0	0	0	0	2	7	19	138	538	1 313	2 841	5 678	5 933	3 267	984	21	16	0
S	0	0	0	0	0	0	0	0	0	0	1	6	47	122	350	754	1 761	3 267	3 193	1 756	61	38	4
T	0	0	0	0	0	0	0	0	0	0	0	0	9	12	50	177	437	984	1 756	1 792	81	52	18
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	5	21	61	81	6	2	1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3	16	38	52	2	0	2
HC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	4	18	1	2	0

FIGURE A.15 – Représentation de la table de contiguïté n'ayant conservé que 99% des liaisons et ayant effectué un retraitement sur la classe de prix.

Figures évoquées dans la section 7.3 Résultat des classifications.

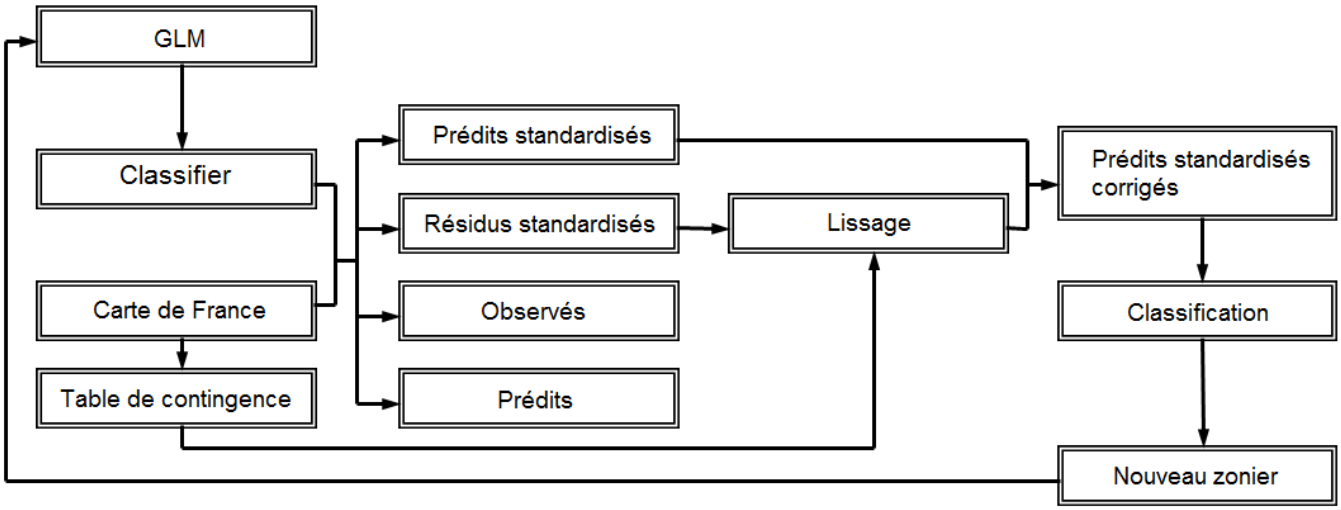


FIGURE A.16 – Processus de création d'un zonier

Figures évoquées dans la section 8.2 Résultats des véhiculiers sur résidus.

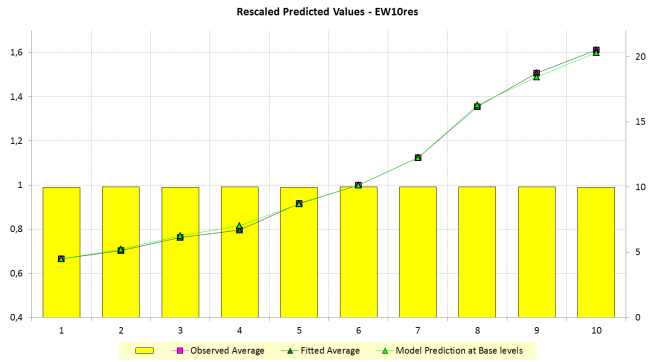


FIGURE A.17 – EWres sur l'échantillon de modélisation, Méthode 1

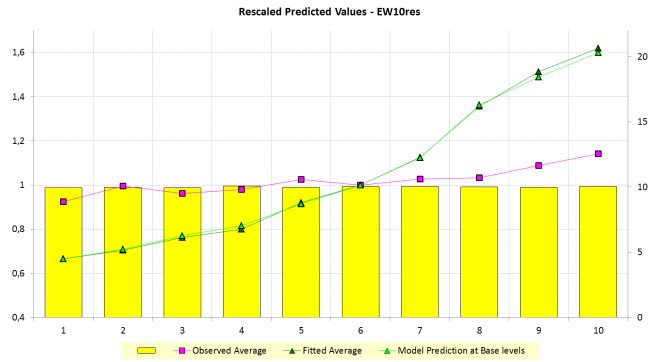


FIGURE A.18 – EWres sur l'échantillon de validation, Méthode 1

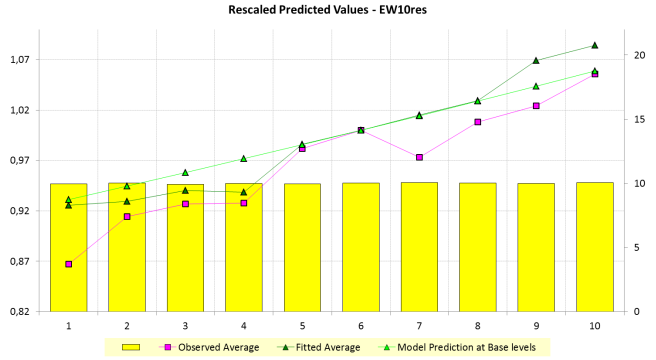


FIGURE A.19 – EWres sur l'échantillon de détermination du lissage, Méthode 2

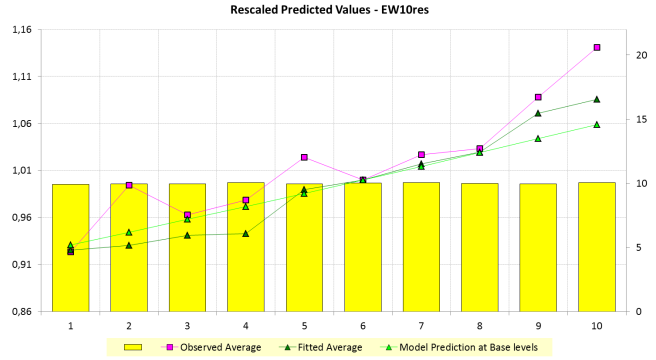


FIGURE A.20 – EWres sur l'échantillon de validation, Méthode 2

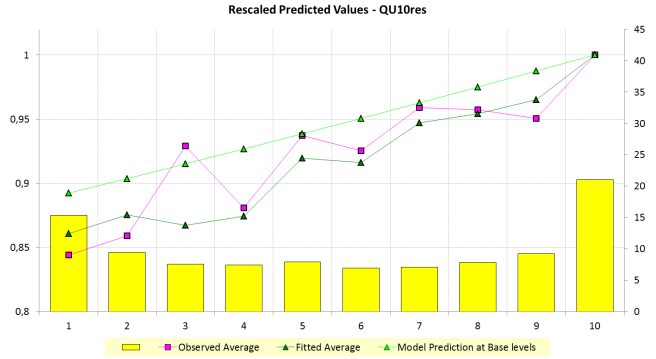


FIGURE A.21 – QUres sur l'échantillon de détermination du lissage, Méthode 2

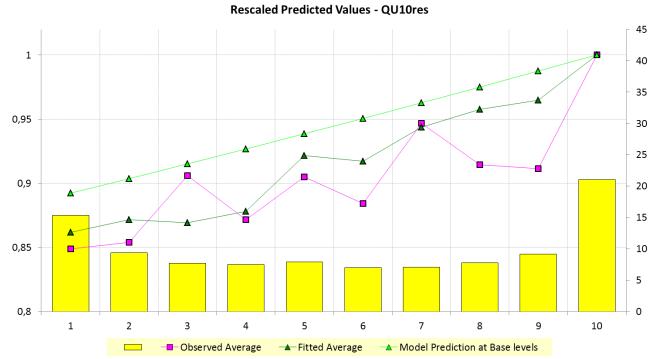


FIGURE A.22 – res sur l'échantillon de validation, Méthode 2

Figures évoquées dans la section 8.3 Véhiculiers sur prédictions et résidus combinés.

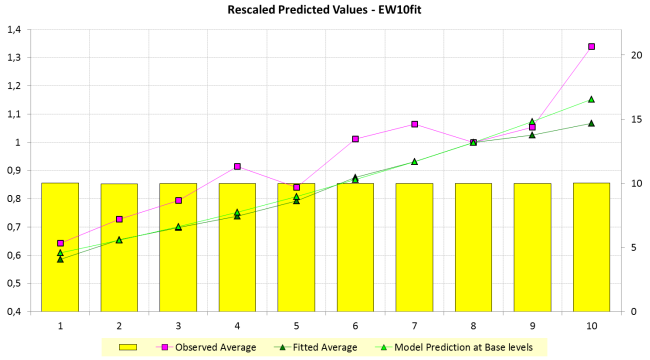


FIGURE A.23 – EWfit sur l'échantillon de modélisation, Méthode 1

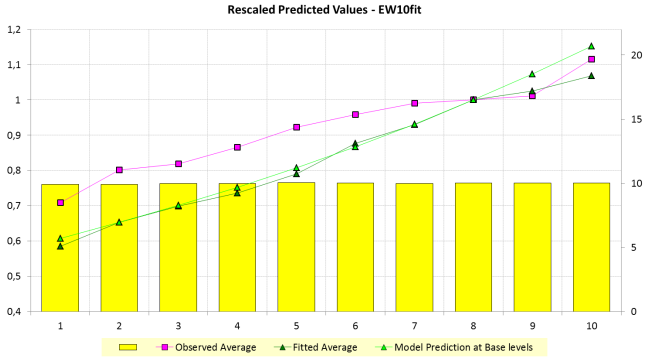


FIGURE A.24 – EWfit sur l'échantillon de validation, Méthode 1

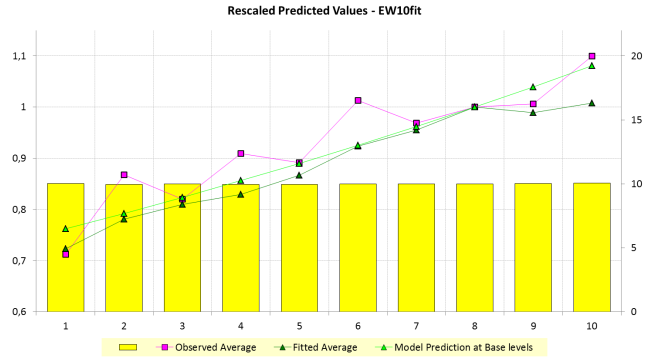


FIGURE A.25 – EWfit sur l'échantillon de détermination du lissage, Méthode 2

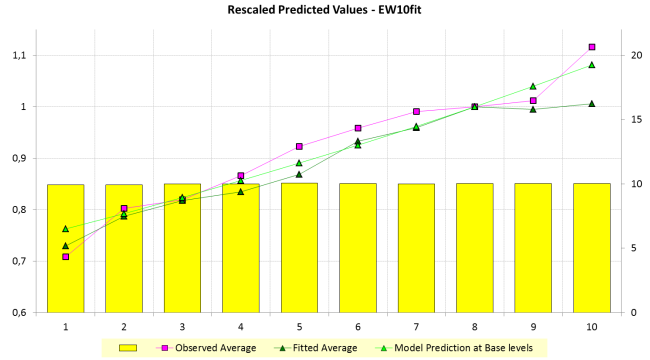


FIGURE A.26 – EWfit sur l'échantillon de validation, Méthode 2

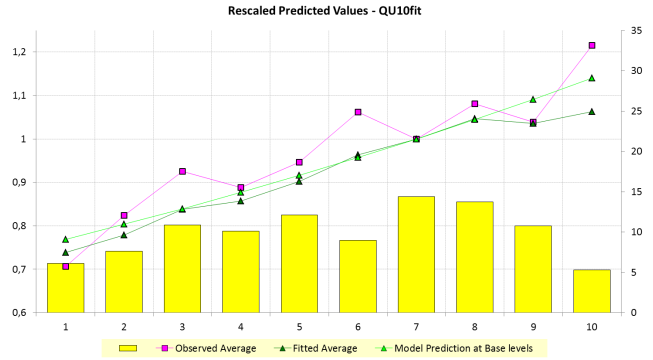


FIGURE A.27 – QUfit sur l'échantillon de détermination du lissage, Méthode 2

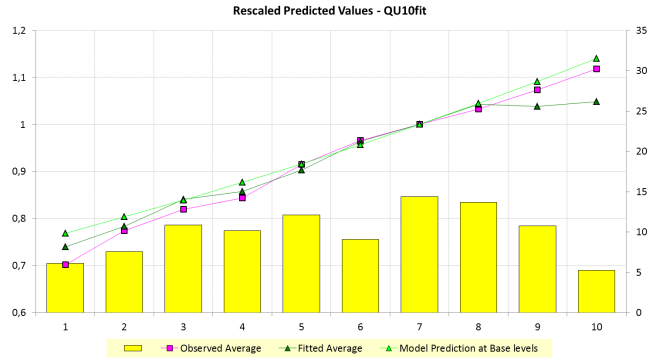


FIGURE A.28 – QUfit sur l'échantillon de validation, Méthode 2

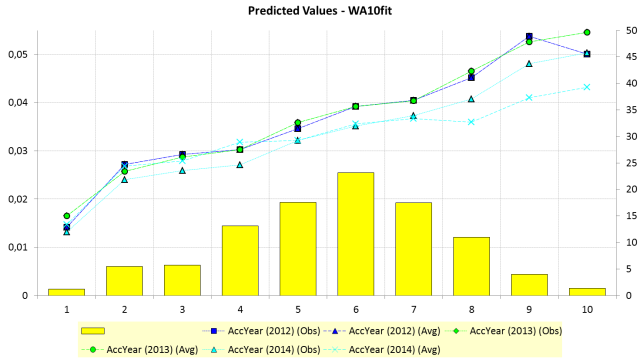


FIGURE A.29 – Test de robustesse du véhiculier WAfit