Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 19/03/2021

Par : **Jean-de-Dieu DELI MADJEON**

Titre : **Modélisation statistique à la création d'un zonier en tarification 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 la filière
M. Pierre PICARD

Membres présents du jury de l'Institut des Actuaires
M. Etienne FLICHY

M. Ying LI

Secrétariat :

Bibliothèque :

Entreprise : Generali IARD
Nom : François-Xavier DUB
Signature : FXD

Directeur du mémoire en entreprise :

Nom : Amin TOUSSI
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

FXD

Signature du candidat

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

RÉSUMÉ

L'assurance automobile représente une part importante de la branche non-vie des compagnies d'assurance. En France, elle est obligatoire pour la garantie Responsabilité Civile, et fait ainsi l'objet d'un marché large et très concurrentiel. Dans ce contexte, les assureurs travaillent davantage à améliorer leurs tarifs afin de répondre le plus fidèlement possible à la demande des assurés. Ce mémoire se propose d'introduire des variables extérieures pour mesurer le risque géographique afin de construire le zonier qui sert à rendre le tarif plus concurrentiel.

Après avoir mené des analyses exploratoires, l'étude se poursuit sur une démarche de modélisation complète de la fréquence des sinistres : c'est le point de départ de l'approche de construction de zonier utilisée dans ce mémoire. La composante géographique du risque, non prise en compte par le modèle précédent, est alors lissée, puis modélisée à l'aide des modèles ensemblistes de machine learning (forêt aléatoire et boosting). Nous terminons sur le zonier construit et sur l'analyse de sa pertinence à répondre à la problématique.

Mots clés : *assurance automobile, concurrence, tarifs, modélisation, zonier, machine learning.*

ABSTRACT

Automobile insurance is an important part of the non-life branch of insurance companies. In France, it is mandatory for Civil Liability coverage and is therefore the subject of a large and highly competitive market. In this context, insurers are working harder to improve their rates in order to respond as faithfully as possible to policyholders' demand. This paper proposes to introduce external variables to measure geographic risk in order to construct the zoning used to make the rate more competitive.

After carrying out exploratory analyses, the study continues with a comprehensive modelling approach to the frequency of claims : This is the starting point for the zoning construction approach used in this paper. The geographical component of the risk - not taken into account by the previous model - was then smoothed and modelled using machine learning models (random forest and boosting). We conclude with the constructed zoning and the analysis of its relevance to the problem.

Keywords : *automobile insurance, competition, pricing, modeling, zoning, machine learning*

NOTE DE SYNTHÈSE

Contexte et problématique

L'assurance automobile représente une part importante de la branche non-vie des compagnies d'assurance. En France, elle est obligatoire pour la garantie Responsabilité Civile, et fait l'objet d'un marché large (22,1 milliards d'euros de chiffre d'affaires en 2018) et très concurrentiel. Dans ce contexte, les assureurs travaillent davantage à améliorer leurs tarifs afin de répondre le plus fidèlement possible à la demande des assurés. Aussi, avec la croissance du volume de données disponible, les méthodes de tarification se sont développées et ont pris de l'importance grâce aux techniques statistiques et économétriques. Ils impliquent une segmentation sophistiquée des clients incluant des critères discriminants du risque, décorrélés des critères usuels de tarification. En conséquence, les nouveaux modèles de tarification en assurance automobile sont plus performants. C'est dans ce contexte que l'introduction des variables extérieures pour mesurer le risque géographique amène à construire le zonier qui sert dès lors à rendre le tarif plus concurrentiel.

L'objectif de ce mémoire est la sophistication du tarif d'assurance automobile pour la garantie Dommages Tout Accident (DTA) de l'Équité à travers la construction d'un zonier. Le zonier est construit en introduisant des variables extérieures pour mesurer le risque géographique.

Données utilisées et méthodologie

Les données utilisées pour mener cette étude sont les données clients fournies par les partenaires dans le cadre du suivi mensuel tel que défini par les conventions signées. Nous avons utilisé les données de cinq partenaires, lesquels faisaient 63% du chiffre d'affaires du

portefeuille d'assurance automobile de l'Équité en 2019.

Les cinq partenaires sont nommés A,B,C,D,E par soucis de confidentialité. Leur choix est fondé sur la représentativité en termes de périmètre géographique couvert et d'exploitabilité des données agrégées. En effet, les formats des données étant plus ou moins différents d'un partenaire à l'autre, il est difficile d'harmoniser les données de plusieurs partenaires et les rendre utilisables dans le cadre d'une modélisation.

L'objectif de la construction du nouveau zonier étant de remplacer l'ancien, celui-ci doit être plus fin que le précédent ; il doit présenter une bonne stabilité et doit par dessus tout garantir une information à la fois précise et robuste sur la sinistralité des assurés.

Le zonier est construit sur la fréquence des sinistres de la garantie DTA du produit d'assurance automobile standard. Les raisons de ce choix sont principalement le fait qu'empiriquement la fréquence des sinistres est plus liée aux caractéristiques géographiques des assurés comparées à leurs coûts. Nous vérifions sur nos données que le coefficient de variation du coût moyen des sinistres par région est de 16.03% alors que celui de la fréquence moyenne des sinistres est de 22.41%.

Pour cette construction, nous avons fait une démarche de modélisation complète de la fréquence des sinistres et avons par la suite comparé le modèle avec zonier à celui sans zonier. L'approche utilisée dans ce mémoire est l'approche classique de construction de zonier en assurance, et se résume suivant ces étapes :

- **Étape 1 :** Modéliser la fréquence des sinistres sans variables géographiques. Nous avons choisi de modéliser par GLM (Modèle Linéaire Généralisé) ;
- **Étape 2 :** Extraire les résidus du GLM ;
- **Étape 3 :** Faire le lissage spatial des résidus du GLM pour corriger les variations importantes au niveau des communes. *Cette étape est facultative selon que les résidus varient énormément par commune ou pas* ;
- **Étape 4 :** Modéliser les résidus lissés par un modèle de machine learning en utilisant comme variables explicatives, des variables géographiques Open Data. Les modèles de Machine Learning utilisés ici sont le Random Forest et le Gradient Boosting (XGBoost) ;
- **Étape 5 :** Faire la prédiction de ces résidus ;
- **Étape 6 :** Découper les résidus prédits en classes pour constituer le zonier. Nous utilisons la méthode des quantiles et des K-means.

La justification de l'utilisation des résidus pour construire le zonier vient du fait que le GLM d'où il sort étant incomplet en termes de variables explicatives, ses résidus ne sont pas des bruits blancs. Ils contiennent l'information des variables absentes de la modélisation et le bruit blanc. Nous voulons ici capter la part de cette information qui provient du risque géographique.

Sinistralité du portefeuille et l'ancien zonier

Le bilan de sinistralité du portefeuille révèle que 98,4% n'ont pas eu de sinistres impliquant la garantie DTA. Il apparaît que la sinistralité du portefeuille est assez hétérogène entre les départements et les régions. Par ailleurs, l'intensité de la sinistralité par département n'est pas la même, selon que l'on parle de la fréquence ou du coût. Cela justifie encore la nécessité d'avoir un zonier fréquence et un zonier coût moyen.

L'ancien zonier a été construit sur la base d'un regroupement autour des grandes régions de France. L'analyse de la sinistralité selon ledit zonier révèle une quasi-absence de discrimination en matière de sinistralité entre quatre des cinq zones définies. Pour précision, les zones sont numérotées de 2 à 6, par ordre décroissant de sinistralité. Les communes ayant les plus faibles fréquences sont donc de la zone 2. Comme on peut le voir sur la figure 1, les zones 2 à 5 - sensées différer en termes de sinistralité - présentent toutes une fréquence de sinistralité autour de 4%. Ce qu'on peut remarquer c'est qu'il semble y avoir une discrimination en termes d'exposition ; la zone 3 sortant du lot avec une forte exposition suivie respectivement des zone 1, 5 et 4. La nécessité d'avoir un nouveau zonier plus discriminant est encore plus renforcée suite à ces constats.

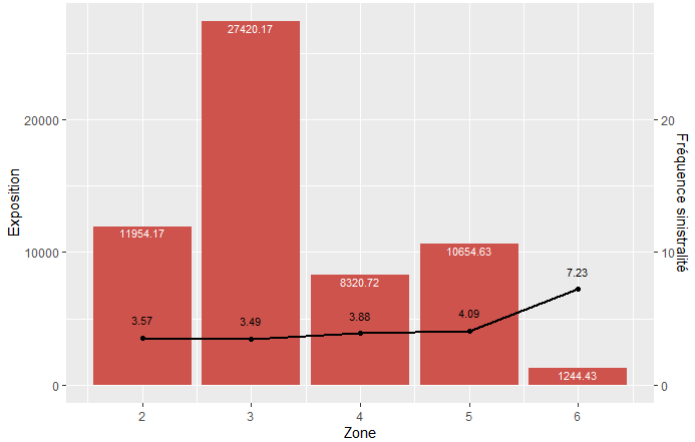


FIGURE 1 – Fréquence de sinistralité selon l'ancien zonier

Modélisation et performance du zonier construit

Au terme de la modélisation de la fréquence, les facteurs discriminants de la sinistralité des assurés se répartissent en trois familles. Dans la première famille, nous avons les caractéristiques de l'assuré : l'âge de son permis, son coefficient bonus-malus, sa profession et le nombre d'accidents qu'il a eu au cours des 36 derniers mois. Dans la deuxième famille, il y a les caractéristiques du véhicule qui sont l'âge du véhicule, et la classe SRA. Le dernier groupe est constitué des caractéristiques du contrat. Il s'agit de l'usage du véhicule et de la présence ou non d'un second conducteur.

Le zonier qui a ensuite été construit, en utilisant les résidus géographiques issus du modèle précédent, a permis au modèle de gagner en pouvoir explicatif. Il a en effet permis de mieux capter le risque de sinistralité du portefeuille, et pourra servir à rendre les primes de la garantie dommage qui en découleront plus concurrentielles. La figure suivante présente la sinistralité actuelle du portefeuille selon le nouveau zonier construit.

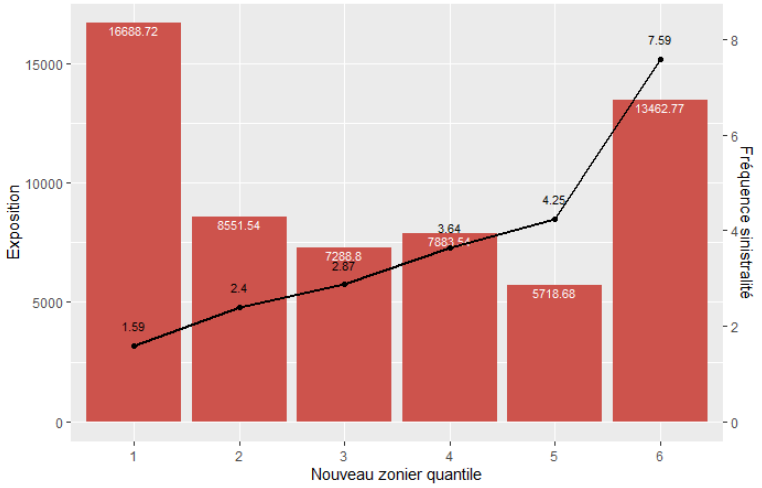


FIGURE 2 – Sinistralité selon le nouveau zonier

Conclusion et perspectives

Bien que le zonier construit affiche de bonnes performances et se présente comme un remplaçant légitime du zonier actuel, différentes voies d'amélioration sont possibles. On pourra : *(i)* ajouter de nouvelles variables géographiques extérieures ; *(ii)* paramétrer plus finement le modèle de lissage spatial si l'on dispose d'une plus grande puissance de calcul ; *(iii)* et réfléchir à intégrer les déplacements du véhicule dans le zonier.

EXECUTIVE SUMMARY

Context and problem

Automobile insurance is an important part of the non-life branch of insurance companies. In France, it is mandatory for Civil Liability coverage and is therefore the subject of a large (22.1 billion euros in revenues in 2018) and highly competitive market. In this context, insurers are working harder to improve their rates in order to respond as faithfully as possible to policyholders' demand. Also, with the growth in the volume of data available, pricing methods have been improved and gained more importance thanks to statistical and econometric techniques. They involve sophisticated segmentation of customers including discriminating risk criteria that are uncorrelated to the usual underwriting criteria. As a result, new pricing models in automobile insurance are more efficient. It is in this context that the introduction of external variables to measure geographic risk leads to the construction of the zoning which is then used to make the rate more competitive.

The objective of this work is to sophisticate the automobile insurance rate for the DTA (Damage Any Accident) coverage of the Équité through the construction of a zoning. The zoning is constructed by introducing external variables to measure geographic risk.

Data used and methodology

To reach this goal, we used data from five partners, which account for 63% of Equité's motor insurance portfolio turnover in 2019.

Those partners were named A,B,C,D,E for confidentiality reasons. The choice of these 5 partners is based on the representativeness in terms of geographical perimeter covered and in terms of usability of the aggregated data. Indeed, the data formats being more or less

different from one partner to another, it is quite difficult to harmonize the data of several partners and make them usable for modeling.

Since the purpose of the new zoning is to replace the old one, it must be thinner than the latter, must be stable and, above all, must provide accurate and robust information on the loss experience of the insured.

The zoning is built on the DTA claim frequency of the standard automobile insurance product. The main reasons for this choice are that empirically the claim frequency is more related to the geographic characteristics of the policyholders compared to their costs. We verify from our data that the coefficient of variation of the average claim cost per region is 16.03% while the coefficient of variation of the average claim frequency is 22.41%.

For this construction, we have carried out a complete modelling process of the frequency of claims and we have compared the models, with zoning and without. The approach used in this study is the classical approach to zoning construction in insurance and is summarized as follows :

- **Step 1 :** Model the frequency of claims without geographical variables. We chose to model by GLM (Generalized Linear Model) ;
- **Step 2 :** Extract the residuals of the GLM ;
- **Step 3 :** Spatial smoothing of the GLM residuals to correct for large variations at the commune level. *This step is optional depending on whether the residuals vary greatly by commune or not* ;
- **Step 4 :** Model the residuals smoothed by a machine learning model using Open Data geographic variables as explanatory variables. The Machine Learning models used here are Random Forest and Gradient Boosting (XgBoost) ;
- **Step 5 :** Make the prediction of these residuals ;
- **Step 6 :** Cut out the predicted residuals in class to form the zoning. We use quantiles and K-means method.

The use of residuals to build the zoning is justified by the fact that the GLM from which they come is incomplete - in terms of explanatory variables - and its residuals are not white noise. They contain the information of the variables absent from the modeling and white noise. We want here to capture the part of this information that comes from the geographic risk.

Portfolio sinistrality and the current used zoning

The portfolio claims record shows that 98.4% had no claims involving DTA coverage. It appears that the portfolio's claims experience is quite heterogeneous between departments and regions. Moreover, the intensity of the loss ratio per department is not the same, depending on whether we are talking about frequency or cost. This further justifies the need for a separate zoning of frequency and cost.

The former zoning system was built on the basis of a grouping around the major regions of France. The analysis of the claims experience according to the said zoning system reveals a virtual absence of discrimination in terms of claims experience between four out of the five zones defined. The zones are precisely numbered from 2 to 6, in descending order of claims. The communes with the lowest frequencies are therefore in zone 2. As it can be seen in figure 3, the zones 2 to 5, which are supposed to differ in terms of claims frequency, have mostly a frequency of about 4%. What can be noticed is that there seems to be a discrimination in terms of exposure ; zone 3 stands out with a high exposure followed by zones 1, 5 and 4 respectively. The need for a new, more discriminating zoning system, is further reinforced by these findings.

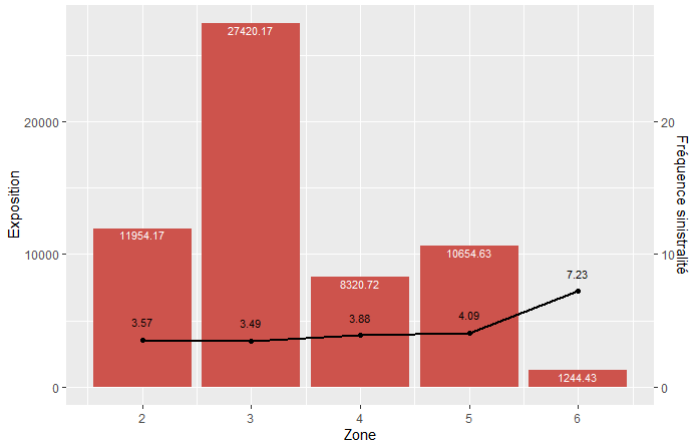


FIGURE 3 – Frequency of claims by former zoning

Modeling results and performance of the new zoning

After modeling the claim frequency, the discriminating factors of the insured's claims experience are divided into three families. In the first family, we have the characteristics of the insured : the age of his license, his bonus-malus coefficient, his profession and the

number of accidents he has had in the last 36 months. In the second family, we have the characteristics of the vehicle which are the age of the vehicle, the use of the vehicle and the SRA class. The last group consists of the contract characteristics. It is the presence indicator of a second driver.

The zoning that was then constructed using the geographic residuals from the previous model allowed the model to gain in explanatory power. In fact, it has made it possible to better capture the portfolio's claim risk and can be used to make the resulting damage coverage rates more competitive. The following figure shows the current loss experience of the portfolio according to the newly constructed zoning.

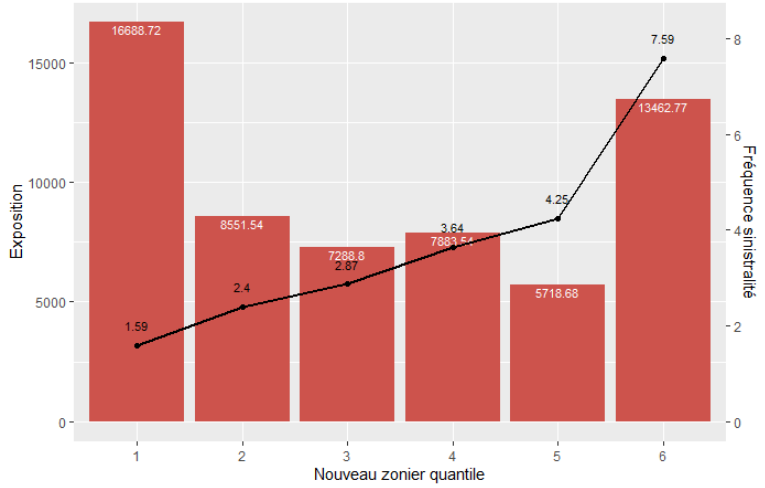


FIGURE 4 – Frequency of claims by the new zoning

Conclusion and perspectives

Although the built zoning variable performs well and is therefore a legitimate replacement for the current zonier, there are several possible ways to improve it. For example, new external geographic variables could be added, the spatial smoothing model could be parameterized more finely if more computing power is available, and consideration could be given to integrating the vehicle's movements into the zoning creation method.

REMERCIEMENTS

Mon stage de fin d'études à Generali IARD s'est bien déroulé grâce au soutien et à la disponibilité de nombreuses personnes à qui, je tiens ici à adresser mes sincères remerciements. Puissiez-vous faire nôtre le succès de ce mémoire. Plus particulièrement :

Je témoigne ma reconnaissance à mon maître de stage M. Amin TOUSSI, Manager Solutions Partenariats Dommages, qui m'a pris en stage dans son équipe et m'a accordé sa disponibilité, son soutien ainsi que ses conseils et orientations pour la rédaction de ce mémoire.

Je remercie également M. Guillaume DURAND, mon deuxième tuteur, et M. Jérémie TOGNI, cadres au sein du pôle assurance automobile, qui ont toujours répondu à mes sollicitations malgré leurs multiples occupations.

Mes remerciements vont également aux cadres, agents, alternants et stagiaires de l'équipe Solutions Partenariats Dommages pour leur chaleureuse collaboration et le climat de convivialité qui a prévalu pendant mon stage.

Pour finir, je remercie mon tuteur académique, M. Pierre PICARD, pour son suivi et les conseils qu'il a pu m'apporter pour la réalisation des travaux présentés dans ce mémoire.

TABLE DES MATIÈRES

INTRODUCTION

L'assurance automobile représente une part importante de la branche non-vie des compagnies d'assurance. En France, elle est obligatoire [1] pour la garantie Responsabilité Civile et fait ainsi l'objet d'un marché large (22.1 milliards d'euros de chiffre d'affaires en 2018 [2]) et très concurrentiel. La montée en puissance croissante des bancassureurs constitue en effet un facteur important faisant monter la concurrence sur ce marché. Dans ce contexte, les assureurs travaillent davantage à améliorer leurs tarifs afin de répondre le plus fidèlement possible à la demande des assurés. Face à cela et avec la croissance des volumes de données utilisées, les méthodes de tarification se sont développées et ont pris de l'importance grâce aux techniques statistiques et économétriques.

Ces techniques impliquent une segmentation sophistiquée des clients incluant des critères discriminants du risque, décorrélés des critères usuels de tarification. En conséquence, les nouveaux modèles de tarification en assurance automobile sont plus performants et imposent à la fois des contraintes sur la structure du risque modélisé et sur les interactions entre les variables explicatives du risque. C'est dans ce contexte que l'introduction des variables extérieures pour mesurer le risque géographique amène à construire le zonier qui sert dès lors à sophistiquer le tarif et à le rendre plus concurrentiel.

Les travaux présentés dans ce mémoire ont été réalisés dans le cadre de mon stage de fin d'études effectué à l'Équité, filiale de Generali France, stage qui s'est déroulé sur la période du 18 mai au 13 novembre 2020. Le portefeuille d'assurance automobile de l'Équité représente près de la moitié de son chiffre d'affaires (1 milliard d'euros en 2019) pour plus de 60 partenaires. La filiale s'inscrit depuis quelques mois dans une logique de mise à jour de ses zoniers dont le zonier pour la tarification automobile. En effet, le zonier automobile

1. loi du 27 février 1958
2. chiffres de la Fédération Française de l'Assurance

actuel est vieux ; il est peu détaillé et est fondé sur une méthodologie qui reprend juste la distinction entre les grandes villes de France. Nous nous proposons donc dans ce mémoire de construire un zonier qui soit plus représentatif du marché actuel de l'assurance automobile en France en général et du portefeuille de la filiale en particulier.

Cette étude commence par la présentation du cadre théorique et conceptuel du milieu de l'assurance automobile en France. Il est ensuite présenté les données de l'étude ainsi que les traitements effectués sur celles-ci. La démarche méthodologique retenue pour la construction du zonier est ensuite présentée. L'analyse exploratoire sera par ailleurs menée afin de mettre en évidence les variables pertinentes pour la suite. L'étude se poursuivra sur la modélisation de la fréquence des sinistres en prélude à l'analyse spatiale des résidus qui serviront à la construction du zonier. Nous terminerons sur l'analyse de la performance du nouveau zonier comparé à l'ancien et à sa pertinence pour la tarification.

Nous tenons à signaler que pour des raisons de confidentialité, les résultats affichés ont été modifiés mais cela ne modifie pas les conclusions des études réalisées.

CHAPITRE 1

ENVIRONNEMENT ET CONTEXTE DE L'ÉTUDE

L'assurance est une opération par laquelle une personne appelée « l'assuré » confie la gestion de son risque à une entité appelée « l'assureur » pour qu'en cas de survenance d'un sinistre il puisse bénéficier du secours de ce dernier. Le principe de l'assurance est simple : l'assureur et l'assuré nouent une relation en signant un contrat par lequel l'assuré s'engage à verser des primes ponctuelles ou régulières et l'assureur s'engage, lorsqu'un sinistre assuré survient, à verser une indemnisation à l'assuré. L'assurance automobile, objet du présent mémoire, est l'une des principales composantes de l'assurance IARD (Incendies, Accidents et Risques Divers) qui permet de protéger les biens. Ce chapitre préliminaire se propose de présenter le cadre conceptuel et méthodologique de l'étude. Il s'agit ici de présenter les généralités sur le marché de l'assurance automobile en France ainsi que la revue de la littérature sur la modélisation tarifaire et la construction de zonier en assurance automobile.

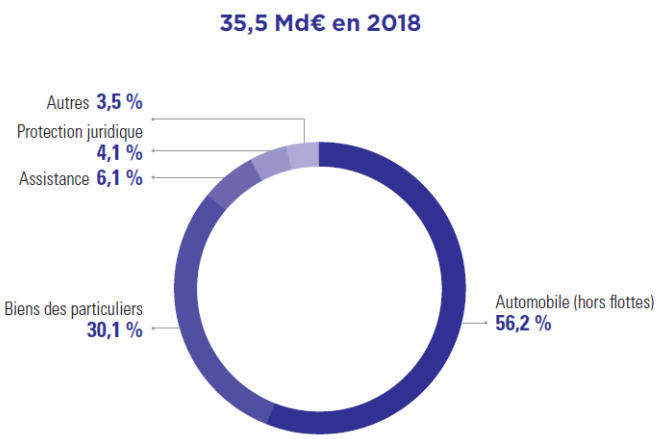
1.1 L'assurance automobile en France et à l'Équité

L'assurance automobile en France est une assurance de principe indemnitaire destinée aux véhicules terrestres à moteur assurés en France et circulant sur le territoire français

comme dans l'Espace économique européen ou dans la zone carte verte [1].

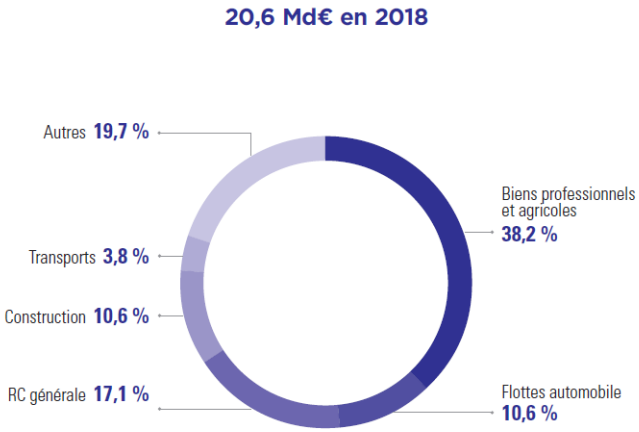
1.1.1 Quelques chiffres clés du marché

Assurer son véhicule est obligatoire en France depuis 1958 pour la garantie Responsabilité civile, ce qui fait de l'assurance automobile un marché important. On note en effet d'après les chiffres de la FFA (Fédération Française de l'Assurance) que pour l'année 2018 il y a eu plus de 49 millions de véhicules de particuliers assurés et plus de 4 millions de véhicules assurés en contrats flottes. Le chiffre d'affaires de l'assurance automobile en 2018 s'établit quant à lui à 22.1 milliards d'euro et représente 39% de l'ensemble des cotisations des assurances de biens et de responsabilité et 56.2% si on se restreint aux particuliers (cf. figures 1.1 et 1.2).



Source : FFA, données clés 2018

FIGURE 1.1 – Assurance IARD : cotisation des particuliers



Source : FFA, données clés 2018

FIGURE 1.2 – Assurance IARD : cotisation des professionnels

En termes de charge sinistre, l'assurance automobile représente 44.8% des sinistres survenus en 2018 en assurance IARD. On peut voir à ce niveau la particularité que présente ce grand marché : le meilleur acteur sur le marché est celui qui saura tirer son épingle du jeu malgré le contexte tendu de sinistralité. Par ailleurs, en matière de garantie, l'assurance automobile en France présente une situation encore plus tendue avec 65% des sinistres dommage enregistrés en assurance IARD. Au sein de l'assurance automobile, 64% des cotisations sont versées pour cette garantie. On note dès lors que la garantie dommage des contrats d'assurance automobile nécessite une attention particulière comme nous le verrons dans la suite de cette étude.

1. La carte verte (ou Carte internationale d'assurance automobile) est un certificat d'assurance automobile délivré par les assureurs européens créée en 1949

1.1.2 Caractéristiques des produits d'assurance automobile

1.1.2.1 Les types de produit

Les produits d'assurance automobile en France se distinguent selon la nature du risque couvert et selon le contenu en termes de garanties couvertes. La nature du risque couvert dépend directement du véhicule assuré et du profil de risque du conducteur. C'est ainsi qu'on distingue selon le critère choisi les 4 roues des 2 roues, les véhicules modestes des véhicules haut de gamme et des véhicules de collection, le risque standard et le risque aggravé. Avec la montée en puissance des trottinettes et objets connexes, une nouvelle gamme de produits d'assurance auto se développe, les produits nouvelles mobilités.

Les produits d'assurance automobile standard sont les produits classiques d'assurance auto pour les profils de risques courants couverts. Les assurances fonctionnent sur le principe de mutualisation des primes et des risques, ce qui permet de compenser les légères différences entre assurés. Passées certaines limites, un assuré potentiel peut mettre en danger cet équilibre. L'assureur considérera alors qu'il entre dans la catégorie du risque aggravé.

le Risque Aggravé est un produit adressé aux personnes ayant un profil atypique (avec par exemple un antécédent d'assurance ou une sinistralité trop élevée). Il est nécessaire de créer un modèle mathématique adapté spécialement pour ce produit car les variables tarifaires utilisées ne sont pas forcément les mêmes que pour le produit standard. Par exemple, l'assureur peut rentrer des données sur le motif de la résiliation, le nombre de mois d'interruption d'assurance, la durée de suspension du permis, etc.

FIGURE 1.3 – Caractéristiques du produit risque aggravé en assurance automobile

Le contenu des produits d'assurance automobile est principalement dépendant des ga-

ranties et options souscrites. Ces garanties et options se combinent au travers des formules pour donner le produit tel qu'inscrit dans le contrat. Pour schématiser simplement, il existe trois principales formules d'assurance auto en France : l'assurance au tiers, l'assurance au tiers plus ou tiers confort (assurance au tiers + autres garanties) et l'assurance tous risques (toutes les principales garanties), auxquelles vous pouvez rajouter les assurances spécifiques comme l'assurance au kilomètre. Il existe alors une multitude de garanties vous permettant de personnaliser votre assurance auto au point de quasiment bénéficier d'une couverture sur-mesure.

1.1.2.2 Les garanties proposées

Comme présenté plus tôt dans cette étude, en France, seule la garantie Responsabillité Civile (RC) est obligatoire. Les autres formules sont optionnelles et rien ne contraint les Français à souscrire à une couverture plus que nécessaire. Pourtant, c'est bien la formule Tous Risques qui séduit le plus de Français au moment d'assurer leur véhicule, afin d'avoir une couverture maximale de leurs risques. Nous présentons ici les quelques grandes garanties en assurance automobile, mais n'oublions pas qu'il existe d'autres garanties et options.

La garantie Responsabilité Civile, une assurance auto obligatoire :

Seule garantie obligatoire en automobile, la RC (appelée parfois assurance au tiers) permet l'indemnisation des dommages causés aux tiers par la faute du conducteur du véhicule ou d'un de ses passagers. L'indemnisation concerne blessures ou décès d'un piéton, d'un passager, ou d'un occupant d'un autre véhicule ou encore les dégâts aux biens matériels autre que le véhicule de l'assuré. La responsabilité civile ne permet pas d'indemniser le conducteur responsable d'un accident de ses propres dommages, mais ses passagers seront indemnisés, quel que soit le lien qu'ils ont avec lui. L'assuré peut aussi choisir une couverture plus étendue des risques. Son contrat comportera alors, en plus de la responsabilité civile, d'autres garanties proposées par l'assureur, comme c'est le cas par exemple dans les formules de contrat dit tous risques.

La garantie Dommage Tout Accident (DTA) :

La garantie DTA est recommandée pour les propriétaires de voitures neuves ou récentes dont la valeur financière est importante. Cette garantie communément appelée « tous risques » permet de dédommager l'assuré de l'ensemble des dommages qui peuvent être causés à

son véhicule : le vandalisme, le délit de fuite subi par l'assuré, la collision avec un animal sauvage ou les accidents dont il est responsable que celui-ci ait impliqué un tiers ou non. Pour prétendre au remboursement des dommages subis par son véhicule, l'assuré doit néanmoins s'acquitter d'une franchise qui dépend d'un certain nombre de facteurs.

La garantie Vol et Incendie (VI) :

Cette garantie permet de recevoir une indemnité dans la limite de la valeur du véhicule le jour de l'incendie ou du vol (et des dommages dus à une tentative de vol ou de forçage de la serrure, de modification des branchements électriques, etc.). En principe, la garantie incendie inclut aussi l'indemnisation des conséquences d'une explosion, de la chute de la foudre ou d'un incendie spontané, criminel ou causé par l'incendie d'un objet voisin. Pour cette garantie l'assureur peut aussi exiger des mesures de préventions (pose d'alarme, garage pour mettre la voiture, mise en place d'un coupe-circuit interdisant le démarrage, gravage sur toutes les glaces du numéro d'immatriculation ou de série, accompagné d'une inscription du véhicule sur un fichier accessible à la police, etc.).

La garantie Bris de Glace (BDG) :

Elle couvre les dommages subis par le pare-brise et peut aussi s'étendre aux glaces latérales, aux vitres de toit ouvrant, à la lunette arrière, aux blocs optiques de phares et aux rétroviseurs. Dans le cadre des assurances pour les véhicules à deux roues, le terme utilisé est « bris d'optique ». Cette garantie a souvent une franchise et celle-ci peut varier selon la valeur du véhicule et selon qu'il s'agit seulement de réparer un impact ou de changer une partie vitrée.

Les garanties DTA, VI et BDG constituent ce qu'on appelle garanties dommages facultatives.

La garantie Protection du Conducteur (GPC) :

Cette garantie couvre les dommages corporels du conducteur lors d'un accident où celui-ci est responsable ou dans lequel aucun responsable n'est désigné. C'est cette garantie qui permet d'assurer le conducteur à l'image de la RC pour les tiers. Elle prend en charge, selon les contrats d'assurance :

- les frais médicaux, chirurgicaux, pharmaceutiques, d'hospitalisation et de prothèses ;
- le préjudice financier lié à un arrêt de travail ou à une incapacité permanente ;

- le préjudice des ayants droit consécutif au décès.

1.1.3 L'offre Équité en assurance automobile

Afin d'apporter des solutions répondant aux spécificités et aux évolutions du marché, des partenariats sont organisés autour de différents domaines d'expertise avec une gamme de produits étendue à toutes les branches (cf. figure 1.4).

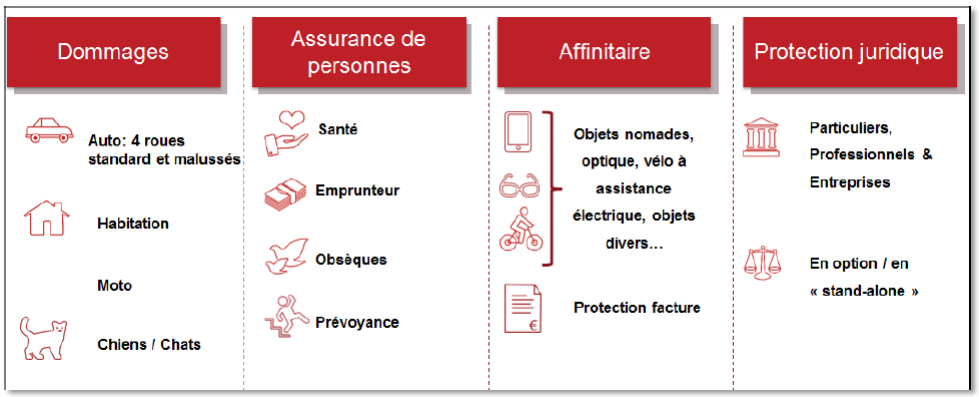


FIGURE 1.4 – L'offre d'assurance de l'Équité.

Aux Partenariats, l'activité automobile représente 39% du chiffre d'affaire et reste une branche à développer. Les nouvelles inventions dans le marché de l'automobile créent de vraies opportunités de développement pour les assureurs. Avec un chiffre d'affaire à hauteur de 400 millions d'euros en 2019, le pôle auto de l'Équité comme les autres assureurs est à l'affût de nouvelles technologies dans le marché qui lui permettraient de créer de nouvelles offres et d'agrandir sa part de marché. Étant donné les produits d'assurance auto de Generali déjà présents sur le marché, l'Équité explore les niches de produit que Generali ne peut pas aller chercher.

L'Équité dispose de quatre principaux atouts qui lui permettent de faire de sa mission une réussite. Elle s'appuie sur l'esprit d'innovation de ses équipes, la réactivité qui se manifeste à travers un time-to-market court, la souplesse de son modèle qui est un modèle en marque blanche et en délégation totale. Vue son activité portée par le partenariat, l'Équité se maintient sur le marché en proposant les garanties et les options les plus communes que l'on peut trouver dans une assurance automobile en France et qui répondent à ses contraintes, des particuliers aux professionnels :

- Le produit **Risque Standard** : il s'agit du produit d'assurance auto classique ;

- Le produit **NVEI (Nouveaux Véhicules Electriques Individuels)** : ce produit est destiné aux trottinettes électriques, hoverboards, rollers électriques, gyropodes, gyroroues et autres engins électriques de déplacement personnel ;
- Les produits **2 roues** : Tous les produits concernant l'assurance des motos, cyclomoteurs, etc ;
- Le produit **véhicule haut de gamme** : ce produit est adressé aux détenteurs de véhicules dont la valeur dépasse un certain seuil ;
- Le produit **véhicule de collection** ;
- Le produit **Risque Aggravé**.

Le contenu des produits proposés varie avec le partenaire car le business modèle de la filiale est de proposer des produits sur mesure. Cela signifie que pour chaque garantie hors RC, l'ÉQUITE pourra proposer un niveau de franchise personnalisé. Certaines garanties restent non intégrables dans les propositions d'assurance, telles que la garantie perte financière, spécificité des véhicules acquis en location longue et/ou avec option d'achat. Par ailleurs, à l'image de la position de leader du groupe dans le domaine, l'Équité continue sa prise de position sur le marché des nouvelles mobilités. C'est dans ce contexte que le produit NVEI a été lancé.

1.2 Modélisation tarifaire en assurance automobile

La tarification d'un produit d'assurance est un processus complexe et rigoureux. Pour évaluer le montant des cotisations de l'assuré, les compagnies d'assurance se basent sur de nombreux critères, liés à la fois au véhicule et à son conducteur. Certains profils d'individus ou certaines marques de véhicule sont considérés à « risques » par les assureurs et subissent une tarification adaptée à leurs profils ; c'est le cas pour les jeunes conducteurs ou jeunes assurés. Nous présentons ici les grandes étapes de la tarification d'un produit d'assurance automobile.

1.2.1 Fondement mathématique de la tarification

La principale caractéristique qui différencie le fonctionnement de l'assurance d'une quelconque production économique est l'inversion du cycle de production. Contrairement à la

situation classique où le producteur d'un bien connaît le coût de production et peut en conséquence proposer un prix de vente pour son bien en adéquation, l'assureur demande une prime d'assurance à l'assureur sans connaître le montant réel des sinistres que l'assuré est susceptible de subir. La constitution de provisions techniques et de marges de solvabilité sont les conséquences directes de l'inversion du cycle de production. Cette spécificité a deux conséquences :

- La première est la nécessité de mettre en place des outils mathématiques sophistiqués afin d'évaluer le montant de la prime à demander à l'assuré pour le protéger du risque et éviter les pertes pour l'assureur. Ces outils sont principalement statistiques et probabilistes et utilisent les données historiques pour cerner la variabilité des risques ;
- La seconde est que l'assurance est extrêmement dépendante des données connues par l'assureur d'une part et l'assuré d'autre part sur le risque couvert par un contrat. Des asymétries d'informations entre les deux protagonistes sont régulièrement constatées et expliquent une partie des règles qui encadrent l'activité d'assurance.

En assurance, deux principes fondamentaux apparaissent antagonistes. Le premier de ces principes est la mutualisation des risques. Pour faire face à une grande variabilité de la réalisation des risques, il y a besoin de les mutualiser, c'est-à-dire en considérer un grand nombre afin de réduire le risque moyen. Le second principe est celui de la segmentation pour avoir des ensembles de risques homogènes. Ce besoin de segmentation vient du fait qu'en général les risques ne sont pas homogènes et qu'il y a besoin de les regrouper afin de pouvoir appliquer une prime différenciée à chacun des groupes ayant un risque homogène.

La recherche de modèles probabilistes permettant une segmentation appropriée des risques est une activité marquante des mathématiques modernes de l'assurance. Il faut cependant noter que trop de segmentation nuit au principe de mutualisation et conduit à une situation problématique puisque les primes de certains assurés peuvent devenir trop importantes. Dans cette situation, le rôle de l'assurance n'est pas de déterminer un équilibre entre ces deux objectifs antagonistes. Le rôle de l'actuaire est, pour un risque donné, d'évaluer de la façon la plus précise possible sa loi de probabilité [2].

Le modèle mathématique développé dans ce contexte par les actuaires permet d'estimer le montant de la prime pure que doit payer l'assuré. La prime commerciale, qui est la prime réellement payée par les assurés est une transformation de la prime pure après le passage

2. C. Dutang ; cours d'actuariat non vie à l'Ensae ; 2019.

par plusieurs chargements. Ces chargements sont les chargements de l'assureur(couvrent ses coûts administratifs), la commission du courtier (finance les coûts de distribution et de gestion supportés par le courtier) et le chargement de sécurité qui couvre le risque de mauvaise tarification.

L'enjeu est alors de déterminer le coût probable des sinistres. Plusieurs approches sont possibles. D'un côté les modèles individuels supposent qu'à chaque police d'assurance correspond un risque individuel et de l'autre les modèles collectifs mettent en avant la segmentation pour définir des classes homogènes de risque et au final répartir les primes grâce à la mutualisation. Au sein des modèles collectifs, il y a les approches de tarification a priori et la tarification a posteriori fondée sur la théorie de la crédibilité. Nous présentons ici l'approche couramment utilisée qui est l'approche fréquence/coût [3] des modèles de tarification a priori.

Rappelons ici que le principe de la mutualisation des risques sur lequel repose la tarification en assurance est fondée sur la loi des grands nombres.

Théorème : loi (faible) des grands nombres

Soit (X_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ. La moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converge en probabilité vers l'espérance : pour tout $\varepsilon > 0$, on a $\lim_{n\to+\infty} \mathbf{P}\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0$.

La charge totale des sinistres est une variable aléatoire car elle n'est pas connue à l'avance. Notons la S et le montant des primes pures perçues est noté π. Le principe de la prime pure montre [4] que $\pi = \mathbb{E}[S]$. Selon le modèle général pour l'approche fréquence/coût, on définit la variable aléatoire S comme la somme

$$S = \begin{cases} \sum_{k=1}^{N} B_k & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

où N est le nombre de sinistre pour une période donnée et les B_k représentent les montants de sinistres. On suppose que N est une variable discrète (dans $\mathbb{N}$) représentant la fréquence tandis que les B_k sont des variables continues positives représentant la sévérité (dans $\mathbb{R}_+$). Un couple de lois pour le nombre et le montant des sinistres définit un modèle fréquence/coût.

On fait les hypothèses :

3. il y a aussi l'approche indemnitaire et l'approche forfaitaire
4. demontrée annexe A page 79

H1 : les sévérités (coût) sont indépendantes et identiquement distribuées : $(B_k)_k \overset{\text{i.i.d.}}{\sim} B$.
H2 : La fréquence des sinistres est indépendante de leurs coûts : $\forall k, N \perp B_k$. Autrement dit la fréquence n'a pas d'influence sur le coût et les montants des sinistres ont le même comportement aléatoire.

On montre que $\forall\ x \geq 0$, en notant $p_N(k)$ la fonction de masse de probabilité de N, la fonction de répartition de S est donnée par :

$$\begin{aligned} F_S(x) &= \sum_{k=0}^{\infty} P(S \leq x \mid N = k) p_N(k) = \mathbb{1}_{x \geq 0} p_N(0) + P\left(\sum_{j=1}^{k} B_j \leq x \mid N = k\right) p_N(k) \\ &= p_N(0) + \sum_{k=1}^{\infty} p_N(k) P\left(B_1 + \cdots + B_k \leq x\right) \end{aligned}$$

Par dérivation par rapport à x, la densité (impropre si $p_N(0) > 0$) vaut :

$$f_X(x) = \sum_{k=1}^{\infty} p_N(k) f_{B_1 + \cdots + B_k}(x)$$

Il est nécessaire d'estimer l'espérance de S donnant la sinistralité attendue afin de déterminer quel montant de prime pure l'assureur doit demander à ses assurés. La variance de S permet de déterminer le niveau de risque pris par l'assureur. On montre (cf. annexe A page 78) que, sous H1 et H2,

$$\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[B] \text{ et } \mathbb{V}ar[S] = \mathbb{E}[N]\mathbb{V}ar[B] + \mathbb{V}ar[N](\mathbb{E}[B])^2$$

L'espérance de S s'écrit au final comme un produit de deux espérances. C'est pourquoi, la connaissance de la loi de N et de la loi de B suffit à déterminer la prime pure ; d'où le nom de l'approche « fréquence × coût». L'avantage d'une telle approche de modélisation de la prime est qu'elle laisse ouverte la possibilité que les facteurs déterminants de la fréquence des sinistres soient différents de ceux de leurs coûts.

1.2.2 Mise en oeuvre pratique de la modélisation tarifaire

L'étape clé de la tarification auto consiste à trouver le meilleur modèle pour estimer les éléments de la formule $\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[B]$ dans le cas de l'approche « fréquence × coût». C'est à ce niveau qu'intervient les modèles d'économétrie et de machine learning qui offrent

des solutions variées et ayant chacune leurs spécificités.

La modélisation repose sur l'exploitation d'une base de données. La base de données est constituée des risques assurés. Les risques représentent la liste des contrats et leurs caractéristiques vus à chaque période d'observation. Pour chaque risque nous avons l'information sur le sinistre survenu s'il y en a eu. En tarification auto, les caractéristiques les plus utilisées sont :

- **les caractéristique de l'assuré :** âge, ancienneté du permis, CRM etc ;
- **les caractéristiques du véhicule :** âge du véhicule, prix, caractéristiques techniques ;
- **les caractéristiques du contrat :** sinistralité antérieure, usage du véhicule, type de parking, etc.

Le modèle de tarification choisi devra alors estimer pour chaque risque, l'espérance du nombre de sinistres et l'espérance du coût des sinistres. Empiriquement, le modèle estimera la fréquence moyenne et le coût moyen. Nous présentons dans le cadre de ce travail le modèle linéaire généralisé (GLM) qui est le modèle le plus couramment utilisé pour la tarification en assurance. Les modèles de machine learing (Arbre de décision, forêt aléatoire, boosting, etc) sont aussi de plus en plus utilisés et peuvent donner dans certains cas de meilleures résultats quand les hypothèses paramétriques sur lesquelles repose le GLM sont peu vérifiables.

Le GLM

Le modèle linéaire généralisé (GLM) comme son nom l'indique est une généralisation du modèle linéaire classique. Dans le cas du GLM, on ne modélise pas la variable à expliquer directement mais plutôt une fonction de l'espérance de cette variable (appelée fonction de lien). Pour un risque i donné, soit Y_i la variable expliquée (la fréquence ou le coût des sinistres) et X_i le vecteur des variables explicatives. Un GLM est caractérisé par trois hypothèses :

1. Une loi de probabilité : les $(Y_i)_i$ sont indépendants et suivent une loi de la famille exponentielle $\mathcal{F}_{\text{exp}}(\theta_i, \phi_i, a, b, c)$ [5] ;
2. Une fonction déterministe : le vecteur de variables explicatives X_i donne le prédicteur linéaire $\eta_i = X_i^T \beta$ où β est le vecteur des coefficients du modèle ;

5. cf. annexe A page 78 pour la définition

3. Une fonction lien $g : \mathbb{R} \mapsto \bar{\mathbb{Y}}$ monotone, différentiable et inversible telle que $\mathbb{E}(Y_i) = g^{-1}(\eta_i)$. $\bar{\mathbb{Y}}$ est le support de la variable expliquée Y. La fonction de lien peut être la fonction identité (modèle linéaire classique), puissance, inverse, logarithme, etc.

Le modèle s'écrit :

$$g(\mathbb{E}[Y_i|X_i]) = \eta_i = X_i^T\beta = g(b'(\theta_i))$$

Si $\theta_i = \eta_i$, alors la fonction lien est dite canonique. C'est à dire $\theta = g(b'(\theta)) \Leftrightarrow g(x) = (b')^{-1}(x)$. Même si la théorie légitime l'utilisation de la fonction de lien canonique, toute fonction monotone, différentiable et définie sur le bon ensemble suffit. Les fonctions identité et logarithme sont privilégiées en pratique. L'effet des coefficients du modèle dépend en effet de la fonction de lien choisie. La fonction de lien identité propose un modèle « additif », et la fonction logarithmique, un modèle « multiplicatif », ce qui en pratique est utile pour arbitrer et calibrer les coefficients.

L'estimation du vecteur de paramètres β et du paramètre ϕ se fait par la méthode du maximum de vraisemblance. Reprenant les notations de l'annexe A sur la définition de la famille exponentielle, la log vraisemblance s'écrit :

$$\ln \mathcal{L}(\beta, \phi) = \sum_{i=1}^{n} \frac{\theta_i(\beta) y_i - b(\theta_i(\beta))}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i)$$

avec $\theta_i = (b')^{-1}(g^{-1}(\eta_i))$. Les équations du score dans le cas général sont

$$\forall j = 1, \ldots, d, \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i) V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} = 0$$

Dans le cas d'un lien canonique, les équations se simplifient

$$\forall j = 1, \ldots, d, \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i)^2} x_{ij} = 0$$

Il n'y a pas de solution explicite à ces équations ; les logiciels utilisent des algorithmes numériques pour estimer les paramètres (à l'instar de l'algorithme de Newton-Raphson).

La principale problématique des modèles linéaires généralisés correspond aux choix effectués, notamment ceux de la loi et des variables sélectionnées. Si les hypothèses faites ne sont valides, la modélisation risque d'être mauvaise.

1.2.3 Une particularité de l'assurance auto : le Coefficient de Réduction-Majoration (CRM)

Le principe de marché libre permet à l'assureur de fixer librement le montant des primes qu'il souhaite appliquer. L'évolution de cette prime avec le temps est cependant encadrée par le code des assurances. En effet, l'assureur est tenu de multiplier ce montant par un coefficient de réduction-majoration (CRM), communément dénommé « bonus-malus ». Le bonus-malus est un système de modulation de la prime d'assurance en fonction du comportement de l'assuré. Plus il est bas, moins la cotisation est élevée et inversement en cas de malus. Le CRM initial est fixé à 1 et est plafonné à 3.5. Un assuré bénéficie d'une réduction de 5% de son CRM pour chaque année sans accident comportant une part de responsabilité de celui-ci. La réduction maximale est fixée à 50%. Ainsi, 13 ans sont nécessaires pour atteindre un bonus de 0,5. Lorsque l'assuré est déclaré responsable lors d'un sinistre, son CRM est majoré de 25% par accident et de 12,5% lorsqu'il n'est que partiellement responsable.

1.3 L'importance du zonier en assurance automobile

La zone géographique est parmi les critères de tarification les plus universellement employés en assurance dommage. En règle générale, les automobilistes qui circulent habituellement dans des zones à faible concentration urbaine ont moins de sinistre que les autres. La zone dans laquelle l'automobiliste est amené à circuler va donc influer sur le tarif. Ces constats sont à la base de la consécration par les assureurs d'un temps précieux à l'élaboration d'un zonier pour la tarification des produits d'assurance auto. En plus, le contexte concurrentiel dans lequel évolue le marché de l'assurance automobile oblige les acteurs à rechercher constamment la meilleure segmentation. Les autres variables étant presque les mêmes en terme de contenu, le zonier devient l'une des variables sur laquelle l'assureur est sûr de créer la différence s'il réussit sa construction. Pour le choix de la zone de tarification, on retient généralement le lieu de stationnement habituel du véhicule.

La question qui se pose à ce niveau est de savoir si l'introduction d'une telle variable dans le modèle de tarification n'engendrerait pas de problèmes méthodologiques. En effet, la localisation étant en général corrélée avec d'autres variables explicatives de la sinistralité des assurés, et que l'effet de la localisation sur la sinistralité est indirect, on peut se demander si un tel modèle ne souffrirait pas d'une lacune méthodologique. Ce qu'il faut noter c'est

que le zonier est un proxy pour diverses variables inobservables, qui peuvent correspondre soit à de l'information cachée soit à des comportements cachés et hétérogènes (c'est-à-dire à de l'antisélection et à de l'aléa moral). La prise en compte de l'antisélection et de l'aléa moral est en effet la problématique essentielle de l'assureur.

Si une bonne segmentation permet en général de traiter correctement le problème de l'antisélection, l'assureur prend généralement en compte l'aléa moral dans les caractéristiques de son contrat à travers les franchises et d'autres caractéristiques explicitement écrites dans le contrat. Le zonier construit en assurance auto permet de rajouter une sophistication à la segmentation faite afin de traiter plus finement les comportements d'antisélection et d'aléa moral qui seraient imputables à la zone géographique.

CHAPITRE 2

DONNÉES DE L'ÉTUDE ET MÉTHODOLOGIE

Ce chapitre présente le support des résultats produits dans ce mémoire : les données et la méthodologie qui ont permis d'avoir ces résultats.

2.1 Les données et leurs traitements

Les données utilisées pour mener cette étude sont les données clients fournies par les partenaires dans le cadre du suivi mensuel tel que défini par les conventions signées. Nous avons utilisé les données de cinq partenaires, lesquels font 63% du chiffre d'affaire du portefeuille d'assurance automobile de l'Équité en 2019.

La compréhension de la complexité de l'entrepôt des données et leurs traitements a constitué presque la moitié de la durée du stage. En effet, afin d'avoir les résultats, il fallait disposer d'une base de données ayant en ligne les images des contrats agrégés par risque [1] et en colonne les caractéristiques de ces contrats. La démarche adoptée pour y arriver se résume alors dans les points suivants :

- **Etape 1 :** Comprendre la structure globale de l'entrepôt des données utilisées par l'équipe ;

1. un changement intervenant sur le contrat pendant sa durée de vie en fait un nouveau risque

- **Etape 2 :** Comprendre la structure et les formats des bases envoyées par chaque partenaire ;
- **Etape 3 :** Importer les bases de données des différents partenaires ;
- **Etape 4 :** Vérifier la validité des données, choisir les variables, harmoniser les bases et faire les retraitements nécessaires ;
- **Etape 5 :** Fusionner les bases images et bases sinistres afin d'obtenir une table image-sinistres ;
- **Etape 6 :** Agréger la table finale fusionnée par assuré par image afin d'avoir une base au format requis pour la modélisation. L'aggrégation consiste à calculer pour chaque image du contrat d'un assuré en une année donnée, le nombre de sinistres survenus.

2.2 Validation de la base de données finale

Avant de se lancer dans les analyses, la base de données finale a été validée. La validation a consisté à sortir les statistiques élémentaires sur la répartition de certaines variables clés ainsi que de la sinistralité et à concilier avec les chiffres décrivant le portefeuille global. Au final, la base de données utilisées pour l'étude est constituée de 1172978 lignes de risques d'assurance automobile comportant une exposition entre le 1^{er} janvier 2017 et le 31 mars 2020. La répartition des contrats selon l'année de début d'exposition du risque et par type de produit est présentée dans le tableaux 2.1 et la figure 2.1. Nous comptons dans le portefeuille pour 72.42% de produit d'assurance auto standard.

Année déb image	Effectif	Proportion (%)	Expo	Proportion expo(%)
2017	410637	35.00	172915.70	37.05
2018	305815	26.07	142711.94	30.58
2019	355887	30.34	139590.62	29.91
2020	100639	8.59	11440.95	2.46

TABLE 2.1 – Répartition de l'échantillon par année de début d'exposition

Dans la suite, comme évoqué plus tôt dans ce rapport, nous avons choisi de construire le zonier pour la garantie DTA. La figure 2.2 présente la répartition du portefeuille de l'étude selon la formule de garantie souscrite. La garantie DTA est souscrite dans la formule Tous Risques. Il apparaît que 19.54% des assurés du portefeuille des 5 partenaires ont souscrit à la formule Tous Risques. En terme de sinistralité, la figure 2.3 présente la garantie DTA

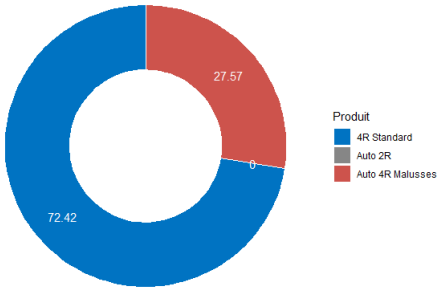


FIGURE 2.1 – Répartition du portefeuille par type de produit

comme la troisième garantie la plus sinistrée après la RC matérielle et la garantie Bris de glace. Dès lors, l'importance de la garantie DTA devient grande et justifie une étude plus poussée de sa répartition au sein du portefeuille en terme de risque.

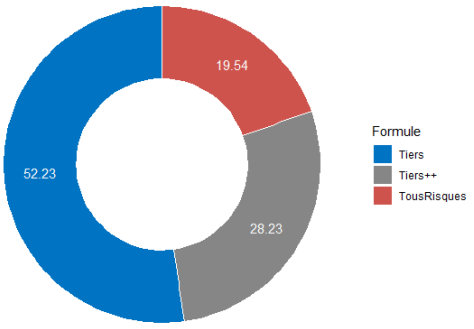


FIGURE 2.2 – Répartition du portefeuille selon la formule

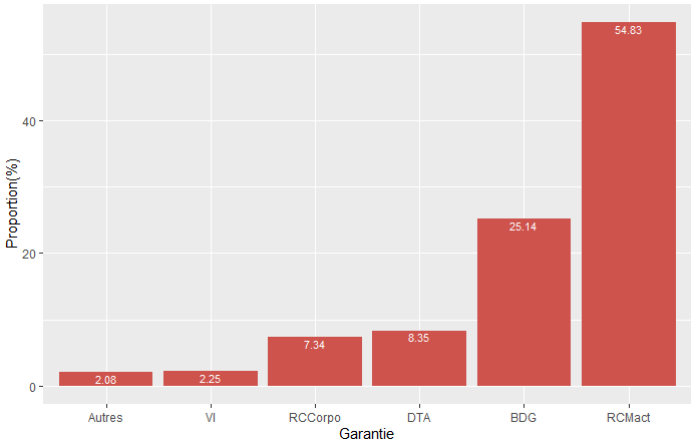


FIGURE 2.3 – Répartition des sinistres selon la garantie

2.3 Méthodologie

Pour rappel, cette étude a pour objectif de construire un nouveau zonier fréquence plus fin que le précédent, présentant une bonne stabilité et qui puisse garantir une information à la fois précise et robuste. Afin d'arriver à cette fin, des analyses exploratoires seront faites au préalable. Cette étape permettra d'avoir des idées sur les variables à conserver dans la modélisation et dans quel sens elles agiront sur la fréquence des sinistres. Les outils d'analyse exploratoires utilisés ici sont les outils de statistique descriptive univariée et bivariée. Des tests statistiques (comparaison de moyennes, indépendance et corrélation) sont utilisés au cours de l'analyse bivariée pour confirmer la significativité des résultats. L'ensemble des

travaux a été fait sur SAS pour le traitement des données et R pour les analyses. Le périmètre considéré est le produit standard d'assurance automobile quatre roues.

L'étude est menée sur les données de 5 partenaires que nous nommons A,B,C,D,E par soucis de confidentialité. Le choix de ces 5 partenaires est fondé sur la représentativité en terme de périmètre géographique couvert et en terme d'exploitabilité des données agrégées. En effet, les formats des données étant plus ou moins différents d'un partenaire à l'autre, il est assez difficile d'harmoniser les données de plusieurs partenaires et les rendre utilisables dans le cadre d'une modélisation.

2.3.1 Approche de construction du zonier

Pour construire un zonier, il faut répondre clairement à certaines questions :

- Pour quelle(s) garantie(s) le zonier doit-il être construit ?
- Quelle est la maille qui nous intéresse et qui nous convient vu les données disponibles ?
- Le zonier est construit sur la fréquence, les coûts, ou les deux ?

Dans le cas de cette étude, nous construisons un zonier pour la garantie DTA. En fonction des résultats obtenus, l'étude sera étendue en interne aux autres garanties en construisant un zonier pour chacune d'elles. La priorité accordée à la garantie DTA vient du fait qu'elle mobilise presque qu'autant de fonds (primes et indemnisation) que la garantie RC. Par ailleurs, elle intègre moins de forfaits à prendre en compte dans le retraitement des données.

Le zonier est construit sur la fréquence des sinistres. Les raisons de ce choix sont principalement le fait qu'empiriquement la fréquence des sinistres est la plus liée aux caractéristiques géographiques des assurés comparée aux coûts des sinistres. Nous vérifions sur nos données que le coefficient de variation du coût moyen des sinistres par région est de 16.03% alors que celui de la fréquence moyenne des sinistres est de 22.41%. Les figures 2.4 et 2.9 suivante montrent en effet comment la fréquence moyenne des sinistres varient beaucoup en terme d'amplitude entre les régions que le coût moyen.

Pour cette construction, il faudra faire une démarche de modélisation complète de la fréquence des sinistres et comparer le modèle avec ou sans zonier (nouveau zonier construit vs zonier actuellement utilisé). L'approche utilisée dans ce mémoire est l'approche classiquement utilisée de construction de zonier en assurance et se résume suivant ces étapes :

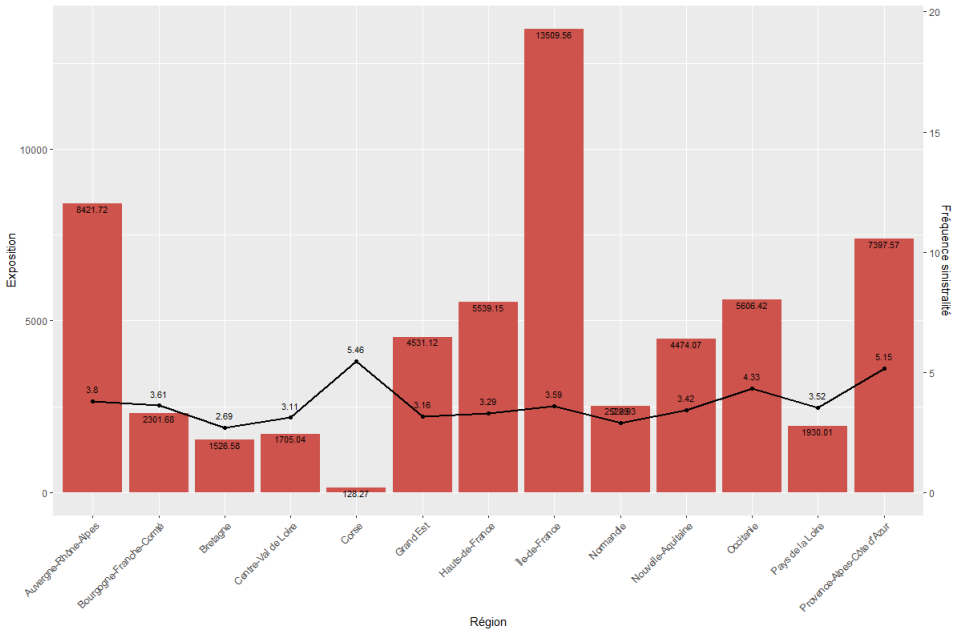


FIGURE 2.4 – Fréquence des sinistres par région

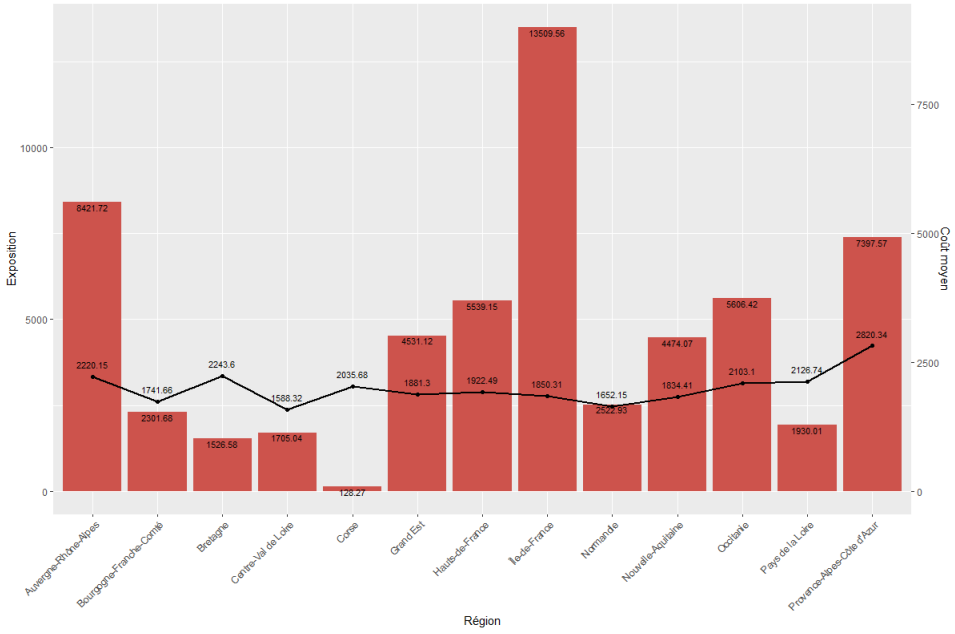


FIGURE 2.5 – Coût moyen des sinistres par région

- **Étape 1 :** Modéliser la fréquence des sinistres sans variables géographiques. Nous avons choisi de modéliser par GLM ;
- **Étape 2 :** Extraire les résidus du GLM ;
- **Étape 3 :** Faire le lissage spatial des résidus du GLM pour corriger les variations importante au niveau des communes. *Cette étape est facultative selon que les résidus varient énormément par commune ou pas* ;
- **Étape 4 :** Modéliser les résidus lissés par un modèle de machine learning en utilisant comme variables explicatives, des variables géographiques Open Data. Les modèles de Machine Learning utilisés ici sont le Random Forest et le Gradient Boosting (XGBoost) ;
- **Étape 5 :** Faire la prédiction de ces résidus ;
- **Étape 6 :** Découper les résidus prédits en classe pour constituer le zonier. Nous utilisons la méthode des quantiles et des K-means ;

La justification de l'utilisation des résidus pour construire le zonier vient du fait que le GLM d'où il sort étant incomplet en termes de variables explicatives, ses résidus ne sont pas bruits blancs. Ils contiennent l'information des variables absentes de la modélisation et le bruit blanc. Nous voulons ici capter la part de cette information qui provient du risque géographique. En effet, la modélisation de la fréquence peut se découper de cette manière :

Fréquence prédite = Fréquence observée +/- Bruit

où **Fréquence observée = Fréquence observée non géographique + Fréquence observée géographique**.

Il faut cependant se rassurer que les résidus extraits du GLM ne contiennent plus aucune autres informations sur la sinistralité que le risque géographique. C'est pour cela que la phase de modélisation de la fréquence non géographique nécessite une attention particulière. Le GLM doit donc avoir été bien calibré et bien estimé. Après la construction du zonier, nous le testons en l'introduisant dans le GLM pour voir s'il améliore les résultats en terme de précision (RMSE, vraisemblance) et de robustesse (GINI).

2.3.2 Rappel sur les lois de fréquence utilisées

Pour modéliser la fréquence des sinistres, il faut avoir une idée des lois de probabilité auxquelles s'ajustent mieux les données. Les lois de probabilité utilisées pour ce type de

modélisation sont les lois de comptage. En assurance, les lois de comptage standards utilisées sont la loi de Poisson, la loi binomiale et la loi binomiale négative que nous présentons dans cette section.

L'exhaustivité de l'utilisation de ces trois lois pour modéliser la fréquence vient des valeurs possibles que peuvent prendre la moyenne et la variance du nombre de sinistres. En effet, selon que la variance du nombre de sinistre est inférieure, égale ou supérieure à la moyenne du nombre de sinistre, nous nous retrouvons dans les conditions d'utilisation de l'une ou l'autre de ces trois lois. Il est évidemment possible d'aller plus loin dans la modélisation en ajustant des formes plus poussées de modèles. Ces modèles seront fondés sur ces trois lois, selon qu'il y a ou non une masse importante d'assurés n'ayant pas eu de sinistre (lois zéro-inflatées et lois zéro-modifiées).

2.3.2.1 La loi binomiale

Pour introduire la loi binonimale, il est habituel d'évoquer préalablement l'expérience (ou l'épreuve) de Bernoulli. L'expérience de Bernoulli est une question de choix arbitraire entre deux valeurs, 0 (échec) et 1 (succès). Il faut juste définir ce qui est considéré comme succès. Au sens probabiliste, le succès est une réponse positive à une question qui n'implique aucun jugement de valeur. Dans notre cas, le succès est la survenance d'un sinistre chez l'assuré. Une variable aléatoire binaire qui prend les valeurs 0 et 1 décrivant l'expérience de Bernoulli est appelée variable de Bernoulli. On note p la probabilité que cette variable prenne la valeur 1 (donc le succès). Le nombre k de succès après la répétition de n épreuves de Bernoulli indépendantes toutes de probabilité p suit la loi binomiale $\mathcal{B}(n,p)$ avec $p \in]0,1[, n \in \mathbb{N}$. En notant X la variable aléatoire de loi binomiale $\mathcal{B}(n,p)$, la probabilité d'obtenir k succès est donnée par :

$$\mathbb{P}(X=k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \text{ pour } k = 0, 1 \ldots, n$$

Sa moyenne et sa variance sont données par : $\mathbb{E}[X] = np$ et $\mathbb{V}ar[X] = np(1-p)$. On a donc $\mathbb{V}ar[X] < \mathbb{E}[X]$ pour la loi binomiale.

En reprenant les résultats de la page 12, si la fréquence des sinistres suit la loi binomiale, la fonction de répartition du coût des sinistres devient :

$$F_S(x) = (1-p)^n + \sum_{k=1}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} F_{B_1+\cdots+B_k}(x)$$

2.3.2.2 La loi de Poisson

Parfois appelée loi des petits nombres, la loi de Poisson est une loi utilisée pour modéliser les évènements rares (comme les accidents dans notre cas). On peut présenter cette distribution comme étant une approximation de la loi binomiale lorsque l'effectif n tend vers l'infini (en pratique, plusieurs dizaines) et la probabilité d'occurrence p tend vers 0 (en pratique, $p < 0.1$). Si le nombre moyen d'occurrences dans un intervalle de temps fixé est λ, en notant X la variable aléatoire comptant le nombre d'occurences, alors la probabilité qu'il existe exactement k occurrences (k étant un entier naturel, k = 0, 1, 2...) est donnée par :

$$P(X = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$$

On dit alors que X suit la loi de Poisson de paramètre λ, noté $X \sim \mathcal{P}(\lambda)$. Sa moyenne et sa variance sont données par : $\mathbb{E}[X] = \mathbb{V}ar[X] = \lambda$.

En reprenant les résultats de la page 12, si la fréquence des sinistres suit la loi de Poisson, la fonction de répartition du coût des sinistres devient :

$$F_S(x) = e^{-\lambda} + \sum_{k \geq 1} \frac{\lambda^k}{k!} e^{-\lambda} F_{B_1+\cdots+B_k}(x)$$

Dans le cadre du GLM, on considère le modèle de Poisson avec des variables explicatives numériques $x_i^{(1)}, \ldots, x_i^{(p)}$ pour l'assuré i. Soit N_i une suite de v.a. indépendantes de loi de Poisson $\mathcal{P}(e_i \lambda_i)$ où λ_i est le paramètre d'intérêt et e_i l'exposition. Le GLM avec la fonction de lien g défini comme à la page 14 suppose une forme paramétrique du paramètre de Poisson :

$$\lambda_i = h\left(\beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)}\right) = h\left(\eta_i(\beta)\right)$$

où $h = g^{-1}$ et $\beta = (\beta_0, \ldots, \beta_p)$ est le vecteur de paramètres à estimer. Typiquement on choisit comme fonction $h = \exp$; et donc la fonction de lien canonique du modèle Poissonien est $g = \log$.

La moyenne du modèle tient bien compte des expositions. On a :

$$\mathbb{E}[N_i] = e_i \exp(\eta_i(\beta)) \Leftrightarrow \log(\mathbb{E}[N_i]) = 1 \times \log(e_i) + \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)}$$

L'exposition doit donc être incorporée au modèle comme un décalage (offset) de coefficient 1 non estimé.

2.3.2.3 La loi binomiale négative

La loi Binomiale Négative apparaît dans l'étude du nombre d'événements pouvant se réaliser dans le temps, événements qui n'ont pas la même probabilité de se réaliser (comme par exemple les mélanges de lois de Poisson et de lois Gamma). Cette loi est très utile en assurance pour l'étude du nombre de sinistres par police durant un temps fixé dans des portefeuilles à risques hétérogènes. Cette loi décrit la variable aléatoire représentant le nombre d'échecs (avant l'obtention du nombre donné r de succès) suite à la répétition d'épreuves de Bernoulli de probabilité p. En notant X cette v.a, la probabilité d'observer k échecs est donnée par :

$$P(X = k) = \begin{pmatrix} r+k-1 \\ k \end{pmatrix} (1-p)^k p^r, k \in \mathbb{N}$$

Sa moyenne et sa variance sont données par : $\mathbb{E}[X] = r\frac{1-p}{p}$ et $\mathbb{V}ar[X] = r\frac{1-p}{p^2}$. On a donc $\mathbb{V}ar[X] > \mathbb{E}[X]$ pour la loi binomiale négative.

En reprenant les résultats de la page 12, si la fréquence des sinistres suit la loi binomiale négative, la fonction de répartition du coût des sinistres devient :

$$F_S(x) = p^r + \sum_{k \geq 0} \begin{pmatrix} r+k-1 \\ k \end{pmatrix} (1-p)^k p^r F_{B_1+\cdots+B_k}(x)$$

Dans le cadre du GLM, nous considérons une suite de v.a indépendantes de loi $\mathcal{BN}(r, p_i)$. Les variables explicatives et l'espérance sont liés comme précédemment par

$$\mathbb{E}[N_i] = \frac{r(1-p_i)}{p_i} = \mu_i = h\left(\beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)}\right).$$

2.3.3 Méthodes utilisées

Nous présentons ici les principales méthodes statistiques et d'apprentissages utilisées dans ce travail en dehors du GLM déjà présenté plus tôt dans ce rapport.

2.3.3.1 Le lissage spatial

Le lissage spatial est utilisé sur les résidus avant leur modélisation par les variables externes. Il permet d'éviter de grands sauts entre des communes voisines. Il est aussi utilisé pour l'interpolation dans les communes où il y a pas de données. Pour une commune donnée, les résidus lissés seront issus d'un mélange entre les données de cette commune et une moyenne pondérée des résidus de toutes les autres commune, la pondération étant plus importante pour les communes les plus proches. Plusieurs méthodes de lissage existent, selon la définition que l'on donne du voisinage et de la distance choisie.

Toutes les méthodes d'interpolation sont basées sur le fait que les données spatiales sont auto-corrélées. L'idée se résume dans cette phrase de Tobler : « tout est relié à tout le reste, mais les choses proches partagent plus de caractéristiques que des choses distantes ». Les techniques d'interpolation et de lissage spatial peuvent être séparées en deux principales catégories : les approches déterministes et géostatistiques. Les méthodes déterministes n'essayent pas de capturer la structure spatiale des données. Elles utilisent seulement des équations mathématiques prédéfinies pour prédire des valeurs à des positions où aucun échantillon n'est disponible. Au contraire, les méthodes géostatistiques cherchent à ajuster un modèle spatial aux données. Cela permet de générer une valeur prédite à des positions non échantillonnées (comme les méthodes déterministes) et de fournir aux utilisateurs une estimation de la précision de cette prédiction. Les méthodes déterministes regroupent les approches TIN, IDW et d'analyses de tendance de surface. Les approches géostatistiques regroupent le krigeage et ses dérivés. Nous utilisons dans notre étude le lissage par la méthode du krigeage.

Puisqu'on peut calculer la variance d'estimation pour tout estimateur linéaire, le krigeage est la méthode de lissage qui choisit celui qui assure la variance d'estimation minimale. En effet, le krigeage est, en géostatistique, la méthode d'estimation linéaire garantissant le minimum de variance. Il réalise l'interpolation spatiale d'une variable régionalisée (les résidus moyens par commune dans notre cas) par calcul de l'espérance mathématique d'une variable aléatoire, utilisant l'interprétation et la modélisation du variogramme expérimental.

C'est le meilleur estimateur linéaire non-biaisé ; il se fonde sur une méthode objective. Il tient compte non seulement de la distance entre les données et le point d'estimation, mais également des distances entre les données deux-à-deux.

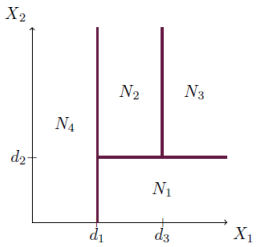
2.3.3.2 L'arbre de décision : CART

Nous présentons l'arbre de décision ici parce qu'elle est à la base des méthodes de machine learning que nous avons utilisées pour modéliser les résidus. Le Random Forest et le Gradient Boosting sont en effet des méthodes d'agrégation d'arbres de décisions.

L'acronyme CART signifie Classification And Régression Trees. Il désigne une méthode statistique, introduite par Breiman et al dans le but de diviser les données d'origine à l'aide de règles déterministes. Ces arbres binaires offrent une manière puissante et conviviale de fournir des résultats dans les problèmes de classification et de régression. Le modèle CART est un outil non paramétrique flexible, et qui est basée sur un algorithme itératif et récursif.

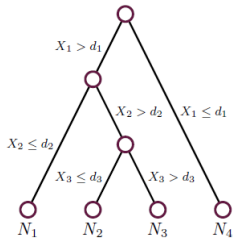
Supposons que nous disposons de $X = \left(X^{(1)}, \cdots, X^{(p)}\right)$ le vecteur de variables explicatives avec $X \in \mathbb{R}^{n \times p}$ et $Y = (Y_1, \cdots Y_n)$ la variable à expliquer pouvant être numérique dans le cas d'une régression ou catégorielle dans le cas d'une classification. L'enjeu de la modélisation consiste alors à estimer la fonction de régression $m(x) = E[Y|X = x]$.

Dans un problème de régression, l'algorithme consiste à faire successivement un partitionnement de l'espace des variables explicatives en plusieurs régions au sein desquelles la variable est modélisée par sa moyenne. Les figures 2.6 et 2.7 illustrent le fonctionnement de l'arbre de décision dans le cas de deux variables explicatives.



Source : B. Le Boucher (2016)

FIGURE 2.6 – Partitionnement de l'espace



Source : B. Le Boucher (2016)

FIGURE 2.7 – Arbre de décision correspondant

La question fondamentale dans la construction des arbres est de savoir quelle variable choisir et quel découpage effectué à chaque noeud. Il faut aussi savoir à quel moment s'arrêter pour éviter le sur-apprentissage. La construction d'un arbre CART se fait alors en 3 étapes

à savoir :

- La construction de l'arbre maximal ;
- L'élagage ;
- La sélection finale.

Le lecteur pourra se référer à l'annexe A page 78 pour la procédure théorique de construction de l'arbre.

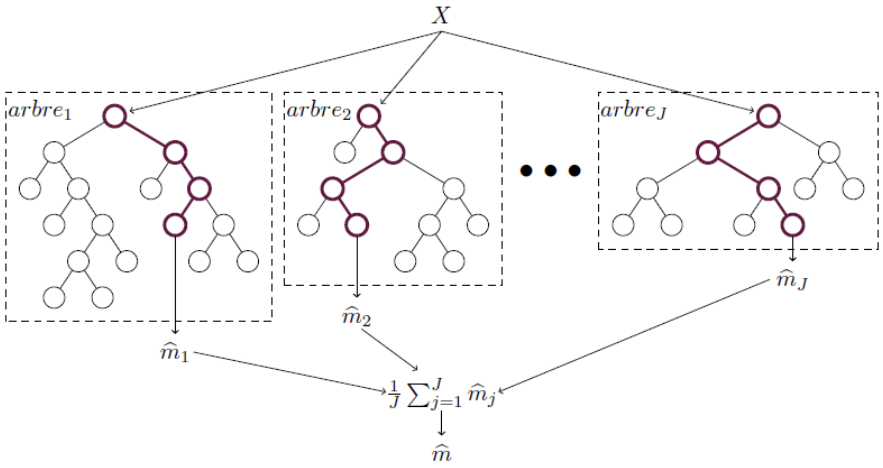
L'un des principaux avantages du CART c'est qu'il est une méthode non paramétrique ; ce qui lui permet d'être utilisé sans besoin d'hypothèse sur la distribution des données. Il permet également de faire une bonne sélection de variables, ce qui lui confère une large applicabilité et une facilité d'interprétation. Par ailleurs, il traite aussi bien le problème des valeurs manquantes en prédiction et résiste aux valeurs aberrantes. Cependant, l'un des défauts majeurs des arbres de décision est leur instabilité. Autrement dit, de petites perturbations de l'échantillon d'apprentissage peuvent engendrer de grandes modifications de la prédiction obtenue. C'est d'ailleurs pour cette raison que nous avons utilisé pour notre étude les méthodes d'agrégation qui résolvent bien ce problème.

2.3.3.3 Le Random Forest

Le Random Forest (ou Forêt Aléatoire en français) est une spécification des méthodes de bagging (contraction de bootstrap aggregating) utilisées en machine learning. Le bagging est une méthode d'agrégation qui repose sur la construction aléatoire d'une famille de modèles. Dans le cas du Random Forest proposé par Breiman (2001), le modèle sous-jacent est l'arbre de décision et on introduit l'aléa dans la sélection des prédicteurs à chaque noeud. Ce tirage aléatoire des variables explicatives à chaque noeud permet d'aboutir à des arbres plus petits et non corrélés. En reprenant les notations de la section précédente sur les arbres de décision, chaque arbre j permet la construction d'un estimateur de la fonction de régression $\widehat{m}_j(x)$ de sorte qu'au final, l'estimateur du Random Forest est obtenu en agrégeant les estimateurs de chaque arbre par :

$$\widehat{m}_J(x) = \frac{1}{J}\sum_{j=1}^{J}\widehat{m}_j(x)$$

Cette agrégation des prédictions par la moyenne de plusieurs modèles permet de réduire la variance et donc de réduire l'erreur de prédiction. La figure 2.8 donne une illustration du fonctionnement du random forest



Source : B. Le Boucher (2016)

FIGURE 2.8 – Illustration du Random Forest à J arbres

L'algorithme des forêts d'arbres de décisions effectue donc un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. La proposition de Breiman vise à corriger plusieurs inconvénients connus de la méthode initiale, comme la sensibilité des arbres uniques à l'ordre des prédicteurs, en calculant un ensemble d'arbres partiellement indépendants. Le principal revers du Random Forest est que l'on perd l'aspect visuel des arbres de décision uniques.

L'algorithme Random Forest permet à la fois de construire un prédicteur et de calculer une estimation de l'erreur de généralisation : l'erreur OOB (Out-of-Bag en anglais). Le calcul de cette erreur utilise le fait que les arbres sont construits sur des estimateurs agrégés construits chacun sur un échantillon aléatoire et qu'ils n'exploitent donc pas toutes les observations. En notant J_i, le nombre d'arbres construits sur un échantillon bootstrap ne contenant pas l'observation i, l'erreur OOB est donnée par :

$$err_{OOB} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{J_i}\sum_{j=1;j \text{ construit sans i}}^{J_i} \widehat{m}_j(x)\right)^2.$$

2.3.3.4 Le Gradient Boosting : XGBoost

XGBoost signifie eXtreme Gradient Boosting. Comme son nom l'indique, c'est un algorithme innovant de Gradient Boosting développé par Tianqi Chen et Carlos Guestrin en 2014. Les algorithmes de boosting sont récents, apparus en 1996 pour l'Adaboost. Le boosting regroupe un ensemble de méthodes dont l'objectif est d'améliorer les techniques de classification et de régression. Nous l'utilisons ici sur l'arbre de décision : il est alors appelé

Gradient Tree Boosting. Il consiste en l'ajout progressif de fonction d'apprentissage faible (à haut biais et faible variance) pour former au final une méthode performante. L'algorithme utilisé capte progressivement l'information de l'arbre, en ajoutant à chaque itération un arbre estimé sur la part de l'information restant à expliquer (les individus mal prédis). A chaque itération, l'algorithme pondère l'importance des observations mal estimées lors de l'étape précédente. Ainsi, chaque modèle est une version adaptative du précédent qui va donner un poids plus important lors de la prochaine estimation aux observations mal prédites et laisse inchangé le poids d'un individu bien prédit.

Pour comprendre le lien avec le gradient, il est nécessaire de rappeler le fonctionnement de l'algorithme de descente du gradient. Il s'agit d'un algorithme itératif utilisé en optimisation pour obtenir le minimum d'une fonction. Dans le cas du Gradient Boosting, il s'agit de minimiser à chaque itération la fonction de perte qui est ici l'erreur de la régression pondérée par le poids déterminé par l'algorithme.

Reprenant les notations introduites pour décrire les arbres de décision, le gradient boosting cherche à minimiser la fonction objectif suivante qui est la fonction de perte pénalisée :

$$\begin{aligned}
&\mathcal{L}(\phi) = \textstyle\sum_i l\left(\hat{y}_i, y_i\right) + \sum_j \Omega\left(m_j\right) \\
&\text{où } \Omega(m) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2 \\
&\text{et } \hat{y}_i = \phi\left(\mathbf{x}_i\right) = \textstyle\sum_{j=1}^{J} m_j\left(\mathbf{x}_i\right)
\end{aligned}$$

Ici l est une fonction de perte différentiable et convexe qui mesure la différence entre la prédiction $\widehat{y_i}$ et la réponse y_i. Le second terme Ω pénalise la complexité du modèle. T est le nombre de feuille de l'arbre m, w est le vecteur de poids de l'arbre m, λ et γ sont des paramètres de pénalisation. L'ajout de la pénalisation permet de lisser les poids du modèle final pour éviter le sur-apprentissage (Chen et Guestrin, 2016). C'est la résolution de ce problème d'optimisation qui utilise l'algorithme de la descente du gradient. Soit $\hat{y}_i^{(t)}$, la prédiction de l'observation i à l'itération t. Le principe de l'algorithme est alors d'ajouter l'arbre m_t pour minimiser la fonction objectif suivante à l'itération t :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + m_t\left(x_i\right)\right) + \Omega\left(m_t\right)$$

Ceci veut dire qu'à chaque itération l'algorithme ajoute l'arbre qui améliore le modèle conformément à la fonction de perte définie précédemment.

La principale différence entre le XGBoost et d'autres implémentations de la méthode du Gradient Boosting réside dans le fait que le XGBoost est optimisé dans le code pour rendre rapides les différents calculs nécessaires à l'application d'un Gradient Boosting. Il traite en effet les données en plusieurs blocs compressés permettant de les trier beaucoup plus rapidement ainsi que de les traiter en parallèle. Son avantage réside aussi dans le fait qu'il offre une panoplie d'hyperparamètres permettant de contrôler finement le modèle implémenté.

2.3.4 Outils de comparaison de modèle

Le problème de la sélection ou de la comparaison de modèles se pose dès lors qu'on cherche à sortir d'une étape de modélisation avec le modèle s'ajustant le mieux aux données en présence. Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle. En effet, les outils disponibles varient beaucoup avec les modèles testés, chaque modèle ayant ses avantages et inconvénients pour un critère donné. Nous utiliserons ici trois types d'outils de comparaisons de modèles :

- ajustement du modèle : vraisemblance, deviance, AIC, BIC ;
- capacité de prédiction du modèle (précision) : RMSE ;
- qualité de classement des prédictions(robustesse) : L'indice de GINI.

2.3.4.1 Mesures d'ajustement du modèle

Les mesures d'ajustement ou d'adéquation sont utilisées pour valider l'ajustement du modèle aux données. Ils permettent de voir si globalement le modèle est le bon modèle pour les données en présence.

La vraisemblance

Dans le cadre du GLM que nous utilisons, les mesures d'ajustement sont toutes basées sur la vraisemblance. D'ailleurs, l'estimation des paramètres du GLM est fait par la maximisation de la vraisemblance. La vraisemblance décrit en effet la plausibilité des valeurs estimées des paramètres du modèle, étant donné l'observation des réalisations de la variable aléatoire dont on a fait l'hypothèse avant estimation. La vraisemblance et log vraisemblance du GLM pour un échantillon de taille n s'écrivent :

$$\mathcal{L}(\beta,\phi)=\prod_{i=1}^{n}\exp\left(\frac{\theta_i(\beta)y_i-b\left(\theta_i(\beta)\right)}{a\left(\phi_i\right)}+\sum_{i=1}^{n}c\left(y_i,\phi_i\right)\right)$$

$$\mathcal{LL}(\beta,\phi)=\ln\mathcal{L}(\beta,\phi)=\sum_{i=1}^{n}\frac{\theta_i(\beta)y_i-b\left(\theta_i(\beta)\right)}{a\left(\phi_i\right)}+\sum_{i=1}^{n}c\left(y_i,\phi_i\right)$$

La vraisemblance étant une probabilité, elle varie de 0 (modèle nulle) à 1(modèle théoriquement parfait). La log vraisemblance quant à elle varie de $-\infty$ à 0. Le meilleure modèle est celui qui a la plus grande vraisemblance et log vraisemblance. On teste la significativité globale du modèle par le test du rapport des vraisemblances. Les hypothèses du test sont :

- **H0** : tous les β sont nuls sauf la constante (modèle contraint) ;
- **H1** : Le modèle considéré est complet, les β sont non nuls.

La statistique du test est donnée par :

$$S_{RV}=2\left(\mathcal{LL}_{n,1}\left(\widehat{\beta},\phi\right)-\mathcal{LL}_{n,0}\right)\xrightarrow{\mathcal{L}}\chi^2(q-1)\text{ sous }H0$$

où $\mathcal{LL}_{n,j}\left(\widehat{\beta},\phi\right)$ désigne la log vraisemblance du modèle sous l'hypothèse $H_j, j=0,1$, $\widehat{\beta}$ est l'EMV de β et q est le nombre de paramètres du modèle complet (y compris la constante). Sous le seuil de significativité α, l'hypothèse H_0 est rejetée si la statistique du test est supérieure au quantile d'ordre $1-\alpha$ de la loi $\chi^2(q-1)$ ou si la p-valeur est inférieure au seuil.

La vraisemblance étant une fonction croissante du nombre de paramètres du modèle, elle sélectionne toujours le modèle le plus complexe. C'est ainsi qu'on va plus loin dans la comparaison de modèle en proposant d'autres mesures d'ajustement : la déviance (par analogie avec le terme anglais *deviance*), et les critère d'information d'Akaike (AIC) et bayésien (BIC).

La déviance

Le modèle le plus simple appelé modèle nul correspond au modèle avec un seul paramètre commun à toutes les observations $y=(y_1,\cdots,y_n)$. Toute la variabilité des observations $y_1,\cdots,y_n$ est expliquée par la composante aléatoire du GLM. A l'opposé le modèle le plus complexe appelé modèle saturé correspond au modèle avec autant de paramètre que d'observations. Toute la variabilité des observations $y_1,\cdots,y_n$ est expliquée par la composante déterministe du GLM.

Notons $\mathcal{LL}(\mu, \phi, \boldsymbol{y})$ la log-vraisemblance en fonction du paramètre de moyenne μ plutôt que du paramètre canonique θ. La déviance pour un modèle ayant q variables explicatives est donnée par le ratio de log-vraisemblance suivant :

$$D(\hat{\mu}, \phi; \boldsymbol{y}) = 2\mathcal{LL}(\boldsymbol{y}, \phi, \boldsymbol{y}) - 2\mathcal{LL}(\hat{\mu}, \phi, \boldsymbol{y})$$

Pour la famille exponentielle, on obtient :

$$D(\hat{\mu}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{2y_i \left(\theta\left(\hat{\mu}_i\right) - \theta\left(y_i\right)\right) - b\left(\theta\left(\hat{\mu}_i\right)\right) + b\left(\theta\left(y_i\right)\right)}{a(\phi)}$$

où $\hat{\mu}$ le paramètre calibré ayant les q variables explicatives, $\theta(\hat{\mu})$ le paramètre canonique pour $\hat{\mu}, \theta(y)$ le paramètre canonique pour le modèle saturé. Plus le modèle est explicatif plus la déviance est petite.

Les critères d'information : AIC et BIC

Les critères d'information permettent de pénaliser les modèles avec un grand nombre de variables afin d'avoir un compromis entre le biais (qui diminue lorsque le nombre de paramètres augmente) et la parcimonie (décrire les données avec le moins de variables explicatives possible). Pour le modèle avec q paramètres estimés (y compris la constante) sur n observations, ils sont donnés par les formules :

$$AIC(\hat{\beta}) = -2\ \mathcal{LL}(\widehat{\beta}) + 2q \text{ et } BIC(\widehat{\beta}) = -2\mathcal{LL}(\hat{\beta}) + q \log n$$

Le meilleur modèle pour un critère d'information est celui qui minimise le critère d'information choisi.

On a : $\log n > 2$ (pour $n \geq 8$). Le BIC aura tendance a choisir des modèles plus parcimonieux comparé au AIC. Burnham et Anderson défendent aussi l'idée que l'AIC possède certains avantages théoriques sur le BIC : d'abord parce que l'AIC est dérivé des principes de la théorie de l'information, au contraire du BIC, ensuite parce que les hypothèses sous-tendant la dérivation du BIC sont discutables. Yang (2005) a également proposé des comparaisons poussées entre AIC et BIC dans le contexte des régressions. En particulier, l'AIC est asymptotiquement optimal lorsque l'on souhaite sélectionner le modèle avec l'erreur quadratique moyenne la plus faible (si l'on fait l'hypothèse que le modèle générant les données n'est pas parmi les candidats, ce qui est en fait presque toujours le cas en pratique) ;

ce n'est pas le cas du BIC. Yang montre également que la vitesse de convergence de l'AIC vers l'optimum est, dans un certain sens, la meilleure possible.

2.3.4.2 Mesures de précision du modèle

La précision des modèles ou leur pouvoir prédictif est évaluée par les mésures appelées indicateurs d'écart. Ce sont entre autres la somme des carrés des résidus (SCR), le carré moyen des erreurs (MSE), l'erreur quadratique moyenne - abrégée RMSE pour *Root Mean Square Error* en anglais -, l'erreur absolue moyenne (MAE) et l'erreur absolue moyenne en pourcentage (MAPE). Nous utilisons dans le cadre de notre étude, le RMSE. En notant y la variable modélisée et $\hat{y}$ sa valeur prédite, le RMSE du modèle est défini par la formule :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_l\right)^2}$$

Comme sa définition le laisse voir, le meilleur modèle selon ce critère est celui qui a le plus petit RMSE. En effet, le modèle minimisant le RMSE est celui dont les estimations sont les plus proches des valeurs observées ; ce qui prend en compte à la fois le biais et la variance. Il faut quand même garder en tête qu'on ne peut comparer que les RMSE de modèles dont les variables expliquées sont exprimées dans la même unité et la même échelle. Il s'agit d'une estimation de l'écart type des erreurs, et l'écart type de la fréquence par exemple ne saurait être de la même grandeur que l'écart type du coût des sinistres. On peut cependant corriger ce problème en divisant le RMSE par l'écart type de la variable y.

2.3.4.3 Mesures de robustesse du modèle : l'indice de Gini

Traditionnellement utilisé pour mesurer des inégalités, l'indice de Gini sert aussi à mesurer la robustesse d'un modèle statistique. Il s'agit ici d'utiliser l'indice de Gini pour évaluer et comparer le pouvoir de segmentation des modèles. En général, les sinistres constituent un phénomène rare ; une petite partie de la population assurée concentre la totalité des sinistres. Le modèle qui modélise un tel phénomène dans l'objectif de segmenter les risques (sous le contrôle global de la mutualisation) devrait donc pouvoir mieux répartir ce risque entre les différents segments qu'il devra dégager.

L'indice de Gini varie entre 0 et 1. Il est égal à 0 dans une situation d'égalité parfaite où toutes les classes ont le même risque ; ce qui voudrait dire le modèle n'est pas du tout

discriminant. La situation la plus inégalitaire correspond à la valeur 1 de l'indice. Il s'agit alors de la situation où le modèle est très discriminant. Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé. On comprend alors que plus l'indice de Gini se rapproche de 0, plus les risques sont mutualisés et donc moins segmentés. Le meilleur modèle est alors le modèle avec le plus grand indice de Gini.

La courbe de Lorentz permet de mieux visualiser et calculer l'indice de Gini. Dans notre cas comme en assurance de manière générale, les proportions cumulées de l'exposition sont représentées en abscisse et en ordonnée se trouvent les pourcentages cumulés des valeurs prédites par le modèle. Si la répartition est parfaitement égalitaire, la courbe est alors confondue avec la première bissectrice. Sinon, l'aire entre la courbe et la première bissectrice est d'autant plus grande que la classification du modèle est bonne.

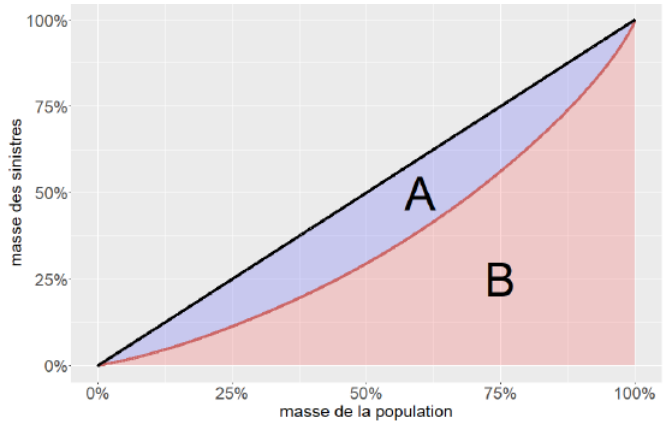


FIGURE 2.9 – Illustration de la courbe de Lorentz

On a, en considérant les notations sur la figure, l'indice de Gini, donné par la formule :

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

2.3.5 La classification par la méthode des k-means

La méthode des k-means constitue l'un des algorithmes de classification le plus répandus. Il permet d'analyser un jeu de données caractérisées par un ensemble de de variables, afin de regrouper les données "similaires" en groupes (ou clusters) de sorte à minimiser une certaine fonction. En considérant la distance d'un point à la moyenne des points de son cluster, la fonction à minimiser est la somme des carrés de ces distances.

La similarité entre deux points de données peut être mesurée grâce à la "distance" séparant les variables qui les décrivent. Ainsi deux données très similaires sont deux données

dont les caractéristiques sont très proches. Cette définition permet de formuler le problème de classification comme la recherche de k "données prototypes", autour desquelles peuvent être regroupées les autres données. Ces données prototypes sont appelées centroïdes. En pratique l'algorithme associe chaque donnée à son centroïde le plus proche, afin de créer des clusters. D'autre part, les moyennes des variables qui décrivent un cluster, définissent la position de leur centroïde dans l'espace des variables : ceci est à l'origine du nom de cet algorithme (k-moyennes ou k-means en anglais).

Il existe un algorithme classique pour le problème. Il est très utilisé en pratique et considéré comme efficace même si son optimalité n'a pas été prouvée. Il se présente comme suit :

- Choisir k points qui représentent la position moyenne des partitions initiales (au hasard par exemple) ;
- Répéter jusqu'à ce qu'il y ait convergence :
  - Affecter chaque observation à la partition la plus proche (c-à-d. effectuer une partition de Voronoï selon les moyennes) ;
  - Mettre à jour la moyenne de chaque cluster.

Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données selon une métrique donnée : on dit alors qu'il a convergé. Dans le cadre de cette étude, le choix du nombre optimal de classes est fait en regardant l'évolution de deux métriques selon le nombre de classes : la métrique silhouette et la méthode du gap statistic.

Pour chaque point à classer, le coefficient de silhouette est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation). La métrique silhouette est alors la moyenne des coefficients de silhouette de tous les points classés.

Dans la méthode du Gap Statistic de Tibshirani et al. (2001), le choix du nombre de classe k est basé sur la comparaison de la variation du logarithme de la distorsion empirique pour le problème de clustering considéré et de celle obtenue pour des données uniformément distribuées.

Comme tout algorithme, la méthode des k-means présente des avantages et des inconvénients : il est simple, rapide et facile à comprendre ; cependant il ne permet pas de trouver des groupes formant des données avec une structure complexe.

CHAPITRE 3

ANALYSES STATISTIQUES DU PORTEFEUILLE

Les travaux de construction du zonier ont été faits en se restreignant au produit d'assurance auto standard souscrit pour la formule Tous risques. Cela nous permet de nous restreindre aux contrats exposés uniquement à la sinistralité sur la garantie DTA. Ces contrats comptent 2232 sinistres DTA sur la période d'étude. La fréquence de sinistralité du portefeuille est de 3.74%. Rappelons ici que la fréquence est définie par la formule :

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

Nous rappelons aussi qu'en assurance, l'exposition représente le nombre d'année-assurance, c'est-à-dire la durée de temps au cours de l'année où le contrat est exposé au risque assuré. Une exposition de 1 signifie que le contrat est exposé toute l'année d'assurance souscrite. Il faut aussi dire qu'en terme des systèmes de gestions, certaines expositions peuvent être légèrement supérieures à 1 sans que cela ne soit une anomalie en soit.

Le bilan de sinistralité (présenté dans le tableau 3.1 suivant) du portefeuille révèle que 98.4% n'ont pas eu de sinistres impliquant la garantie DTA. Les fréquences présentent donc une masse en 0.

Ces chiffres sont faibles mais représentent bien la tendance globale des contrats gérés par la compagnie pour ce qui concerne les produits d'assurance automobile standard restreints à

Nb. sinistres	Eff. polices	Proportion (%)	Exposition	Proportion expo (%)
0	133928	98.40	58128.67	97.54
1	2126	1.56	1423.13	2.39
2	48	0.04	40.22	0.07
3	2	0.00	1.35	0.00
4	1	0.00	0.76	0.00

TABLE 3.1 – Répartition du nombre de sinistres

la garantie DTA. Il faut aussi ajouter que la sinistralité DTA est peu fréquente mais présente un coût particulièrement élevé après la garantie Responsabilité (RC). Dans la suite de la modélisation, nous regroupons les deux, trois et quatre sinistres. Les contrats du portefeuille sur lequel la modélisation sera faite ont été distribués par cinq partenaires dont la répartition est donnée par la figure suivante.

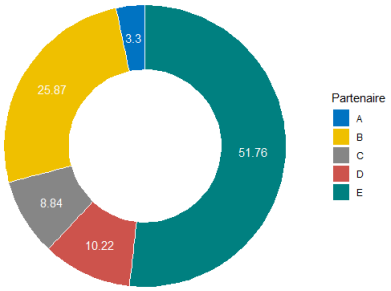


FIGURE 3.1 – Origine des contrats du portefeuille

3.1 Analyse géographique de la sinistralité

L'objectif principal de cette étude est de mettre en évidence l'importance et la nécessité d'une bonne segmentation géographique pour avoir un tarif d'assurance automobile concurrentiel pour la garantie DTA proposée par l'Équité.

On pouvait déjà voir d'après la figure 2.4 page 21 que la fréquence des sinistres varie beaucoup avec la région du lieu de garage du véhicule. La figure 3.2 va à un niveau plus fin et représente la fréquence de sinistralité et le coût moyen des sinistres survenus par département.

Il apparaît que la sinistralité du portefeuille est assez hétérogène entre les départements. On peut voir que l'intensité de la sinistralité par département n'est pas la même, selon que l'on parle de la fréquence ou du coût. Cela justifie encore la nécessité d'avoir un zonier

fréquence et un zonier coût moyen. On peut ainsi noter sur la carte que les sinistres sont plus coûteux en moyenne sur les côtes Nord-Est, Sud et Est. En terme de fréquence, on peut noter une grande hétérogénéité entre les départements si bien qu'il est difficile de sortir une tendance claire. Il faut remarquer quand même que les côtes Est et Sud ressortent encore avec des fréquences élevées, suivi de quelques départements des régions du centre, du pays de la Loire, d'île-de-France et de Picardie.

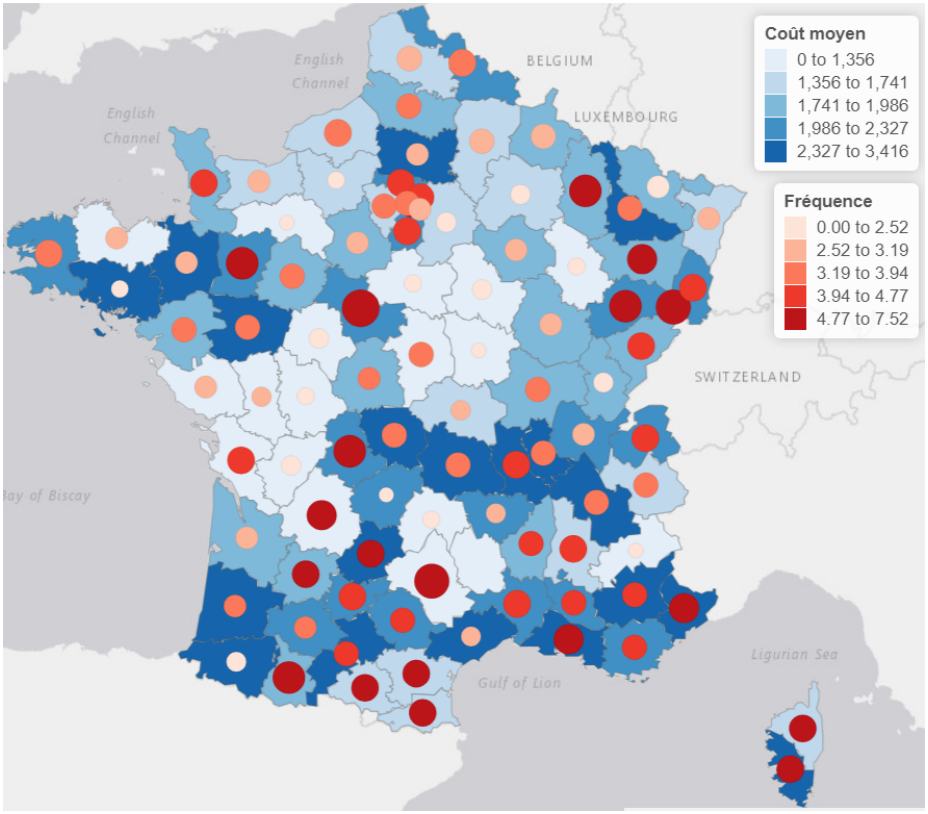


FIGURE 3.2 – Carte de la sinistralité par département

L'ancien zonier a été construit sur la base d'un regroupement autour des grandes régions de France. L'analyse de la sinistralité selon ledit zonier révèle une quasi-absence de discrimination en termes de sinistralité entre quatre des cinq zones définies. Pour précision, les zones sont numérotés de 2 à 6, par ordre décroissant de sinistralité. Les communes ayant les plus faibles fréquences sont donc de la zone 2. Comme on peut le voir dans la figure 3.3, les zones 2, 3, 4 et 5 sensés différer en terme de sinistralité présentent toutes une fréquence de sinistralité autour de 4%. Ce qu'on peut remarquer c'est qu'il semble y avoir une discrimination en terme d'exposition ; la zone 3 sortant du lot avec une forte exposition suivie respectivement des zone 1, 5 et 4. La nécessité d'avoir un nouveau zonier plus discriminant est encore plus renforcée face à ces constats.

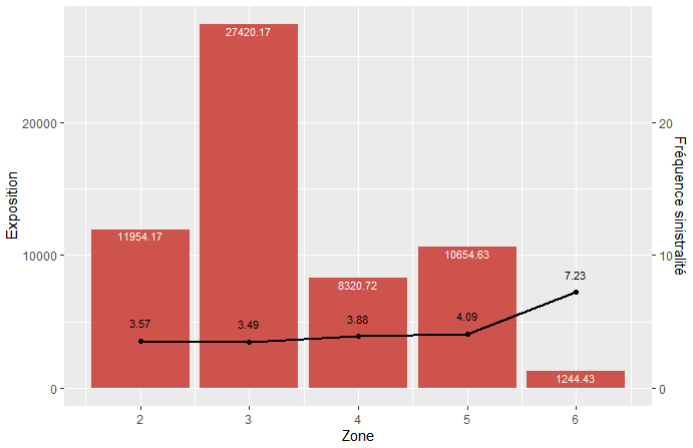


FIGURE 3.3 – Fréquence de sinistralité selon l'ancien zonier

Regardons à présent l'adéquation entre la sinistralité de l'ancien zonier et les tarifs pratiqués. Les deux cartes qui suivent présentent l'ancien zonier (les communes) et le loss ratio par département. Une bonne adéquation voudrait que les zones classées à haute sinistralité soient aussi les zones où les tarifs pratiqués sont élevés de sorte que les loss ratio par département soient comparables.

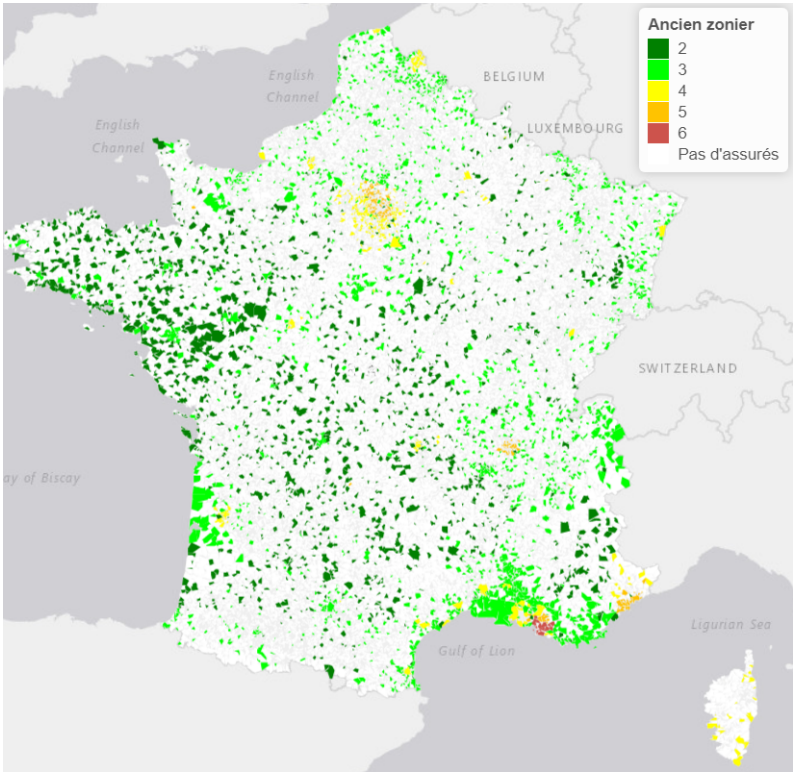


FIGURE 3.4 – Zonier actuel utilisé

Pour rappel, le loss ratio, encore appelé ratio de sinistralité ou S/P est un indicateur important pour le pilotage d'un portefeuille d'assurance. Il permet d'apprécier la rentabilité

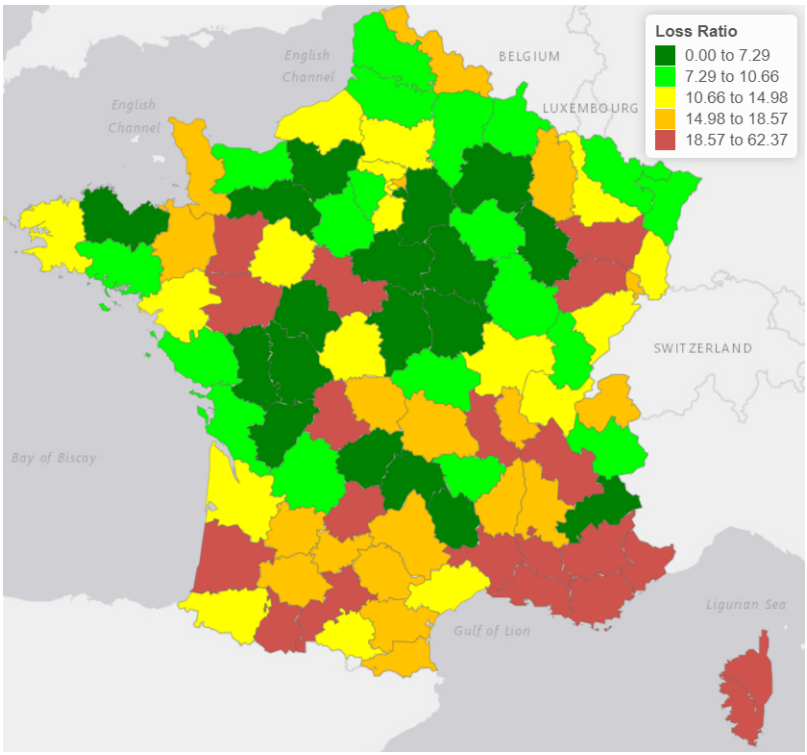


FIGURE 3.5 – Loss Ratio de la garantie DTA par département

du portefeuille en faisant le rapport de la charge des sinistres sur la prime encaissée. L'idéal pour l'assureur est donc d'avoir un loss ratio le plus faible possible ; ce qui voudrait dire qu'il arrive à dégager une marge de bénéfice avec le tarif pratiqué. La figure 3.5 montre que le loss ratio varie avec le département et que globalement le Sud de la France est moins rentable que le Nord. En confrontant avec le zonier actuel on peut voir par exemple que la partie Sud-Ouest présente un loss ratio qui n'est pas en adéquation avec sa classification dans le zonier. En effet, le zonier classe cette zone parmi celle à faible sinistralité alors que le loss ratio indique qu'il s'agit de l'une des zones les moins rentables du portefeuille.

3.2 Caractéristiques des risques du portefeuille

La modélisation de la sinistralité du portefeuille en vue de la construction du zonier nécessite de disposer d'informations conséquentes sur les contrats d'assurance. L'importance de disposer de suffisamment de variables décrivant les contrats tient au fait que cela permet d'être sûr de modéliser fidèlement le risque non géographique avant l'extraction des résidus pour le zonier. Le tableau 3.2 présente les variables explicatives retenues ainsi que les différentes modalités qu'elles peuvent prendre.

Variable		Description	Modalités/amplitude
Caractéristiques de l'assuré	Age du conducteur	Différence entre la date de naissance et la date d'effet du contrat	18-110 ans
	Ancienneté permis	Différence entre la date de peris et la date d'effet du contrat	0-70 ans
	CRM	Coefficient de Réduction-Majoration (bonus-malus)	0.5-1.0
	Ancienneté CRM50	Nombre d'années à CRM 50	0-50
	Profession	Profession du conducteur principal	cf fig
	Situation Familiale	Situation familiale du conducteur principale	Célibataire;Concubin;Divorcé;Marié;Veuf
Caractéristiques du véhicule	Age du véhicule	Différence entre la date de mise en circulation du véhicule et la date d'effet du contrat	0-60 ans
	Classe SRA	Classe définie par le SRA (valeur du véhicule)	A-Z;HC
	Groupe SRA	Groupe défini par le SRA (puissance du véhicule)	20-50
	Lieu de garage	Type de lieu de garage du véhicule	Garage_Box;Parking Extérieur;Voie Publique
	Mode d'acquisition	Mode d'acquisision du véhicule	Comptant_Crédit;Leasing
	Usage	Usage du véhicule	Privé;Prive_Trajet_Travail ;Tout déplacement
Caractéristiques du contrat	Fractionnement du paiement	Mode fractionnement de la prime	Annuel;Semestriel;Trimestriel;Mensuel
	Second conducteur	Présence de second conducteur	Oui;Non
Autres	Sinistres Matériels	Nombre de sinistres matériels au cours des 36 derniers mois	0-7
	Sinistres Corporels	Nombre de sinistres corporels au cours des 36 derniers mois	0-2
	Autres Sinistres	Nombre des autres sinistres au cours des 36 derniers mois	0-4
	Partenaire	Courtier distributeur du produit	A;B;C;D;E

TABLE 3.2 – Liste des variables retenues pour le modèle de fréquence

La justification de l'utilisation du CRM comme variable tarifaire pour la fréquence vient du fait qu'au plan empirique - constat fait par les professionnels du métier - son niveau ne reflète pas toujours le risque réel de l'assuré. L'expérience a en effet montré que les assurés après avoir duré à des niveaux bas du CRM, assez longtemps, baisse de vigilance et devienne moins attentif au volant. En associant ainsi le niveau de CRM à l'ancienneté au CRM50, on arrive à capter certains niches de risques et ainsi sortir de meilleurs tarifs.

Le groupe et la classe SRA (Sécurité et Réparation Automobiles) sont des caractéristiques importantes du véhicule. Ils permettent d'agréger en deux variables à la fois les caractéristiques techniques et le prix du véhicule. Ils sont obtenus à partir du code SRA (antérieurement code GTA). Pour retrouver un code SRA, l'assureur a besoin d'informations sur la voiture (marque, modèle, version, année de première mise en circulation, etc), mais surtout du CNIT (Code National d'Identification du Type) ou anciennement Type Mines. Noté de 20 à 50, le groupe SRA représente la puissance de l'automobile (20 = faible puissance, 50 = forte puissance), mais également sa dangerosité.

Il existe deux types de classes dans le fichier SRA : la classe de prix et la classe de

réparation. En assurance automobile, la classe SRA utilisée est la classe de prix. Notée de A à Z (A = faible valeur, Z = forte valeur, HC = Hors Classe), elle est établie par la valeur à neuf TTC du véhicule (hors option, hors remise). Elle est utile à l'assureur en cas de vol ou de remplacement du véhicule suite à un sinistre.

3.2.1 Caractéristiques du conducteur

Les caractéristiques présentées ici sont les caractéristiques de l'assuré telles qu'indiquées dans le tableau 3.2. Le portefeuille étudié est constitué d'assurés âgés en moyenne de 46 ans (cf figure 4.19 en annexe). La courbe de la fréquence par âge (cf figure 3.6) suggère que la discrétisation de l'âge pourrait constituer des groupes plus homogènes de risques. Les principaux critères de découpage sont la sinistralité des classes et l'équilibre des effectifs par classe. Nous avons ainsi choisi de faire quatre classes d'âge : 18-29, 30-59, 60-69 et 70 ans et plus. La figure 3.7 présente les résultats du découpage et montre que le découpage a bien pris en compte les niveaux de risque.

La distribution de l'âge du permis s'inscrit dans le même cadre d'évolution que celle de l'âge. Il va de 0 à 70 ans d'âge avec une moyenne de 24 ans. Le risque de sinistralité est élevé pour les jeunes conducteurs puis se temporise à partir de 10 ans avant de remonter après 45 ans d'expérience de conduite, dû notamment à l'âge déjà élevé du conducteur. Le découpage de l'âge du permis s'est fait sur le même principe que celui de l'âge du conducteur. (cf figures 4.21 et 4.22 en annexe).

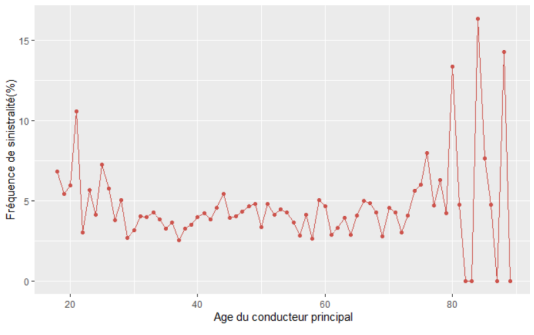


FIGURE 3.6 – Sinistralité par âge du conducteur

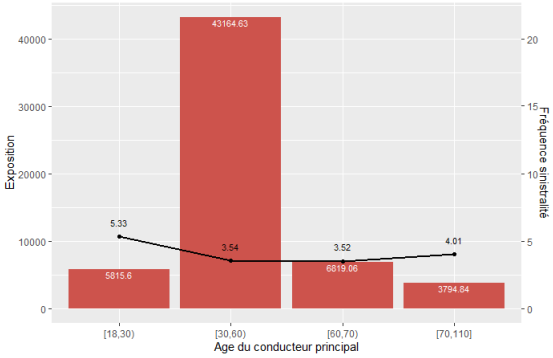


FIGURE 3.7 – Découpage de l'âge en classe

En ce qui concerne la qualité et l'expérience pratique des conducteurs, le coefficient bonus-malus moyen du portefeuille est de 0.62 et trois conducteurs sur 4 ont un coefficient inférieur ou égal à 0.72. On peut dire à ce niveau qu'ils sont relativement bons conducteurs,

d'autant plus que 50% des conducteurs du portefeuille ont un CRM de 0.50. L'information apportée par le niveau du CRM est renforcée par le nombre d'années au CRM 0.5, le coefficient le plus bas. Cela permet d'aller plus loin dans la discrimination en distinguant parmi les assurés ayant un CRM de 50 lesquels sont meilleurs conducteurs. Les résultats montrent que l'ancienneté des assurés au CRM 50 ne permet de réduire leur sinistralité que dans les premières années ; il semble y avoir un relâchement après 3 ans. Comme présenté sur la figure 3.8, le CRM a aussi été découpé par classe en suivant le même principe que celui de l'âge.

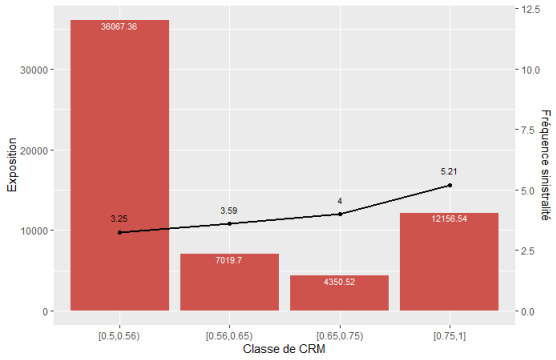


FIGURE 3.8 – Sinistralité par classe de CRM

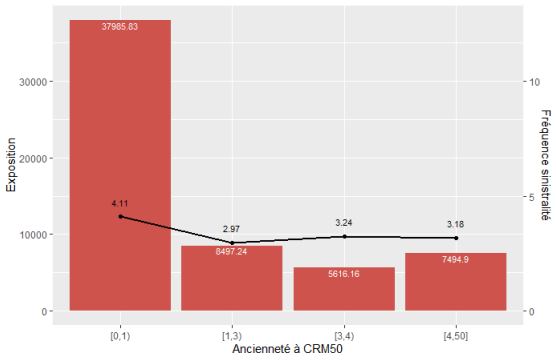


FIGURE 3.9 – Sinistralité par ancienneté au CRM 50

Les résultats pour la profession et la situation familiale sont présentés en annexe (cf figures 4.23 et 4.24). Le portefeuille est constitué à 66% de salariés et les étudiants, représentant 1% du portefeuille ont la fréquence la plus élevée (4.65%). Pour la situation familiale, 87% des risques couverts sont détenus par les célibataires (54%) et les mariés (33%). Ce sont les mariés qui présentent la plus faible fréquence de sinistralité.

3.2.2 Caractéristiques du véhicule et du contrat

92% des véhicules assurés ont été acquis au comptant ou à crédit. En moyenne, ils ont 6 ans d'âge (cf. figure 4.20 en annexe). Le plus vieux véhicule est un véhicule de 77 ans, classe D, groupe 28. Les véhicules acquis par Leasing sont les plus sinistrés (4.76% de fréquence). L'analyse de la sinistralité montre une décroissance linéaire entre la fréquence des sinistres DTA et l'âge du véhicule (figure 3.10) si bien que les véhicules de plus de 15 ans ont une fréquence de sinistralité de 1.75% alors que les véhicules neufs de moins de 5 ans ont enregistré une fréquence de 4.56%.

L'analyse de la sinistralité selon l'usage du véhicule montre aussi que ce sont les véhicules

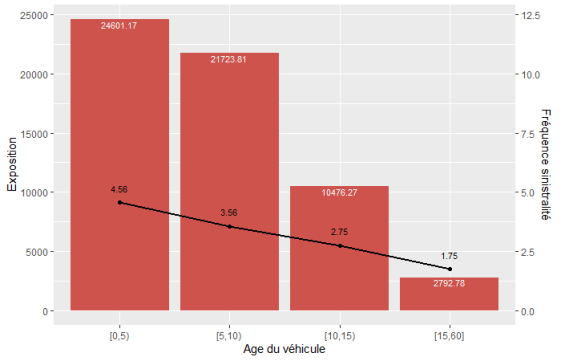


FIGURE 3.10 – Sinistralité selon l'âge du véhicule

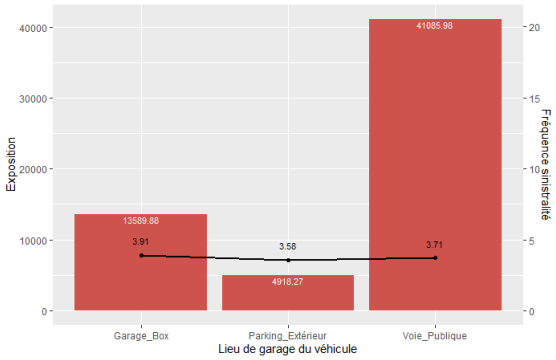


FIGURE 3.11 – Sinistralité selon le type de garage

utilisés pour tout type de déplacement qui ont enregistré plus de sinistres (4.37% de fréquence) et que les parkings extérieurs constituent un facteur de réduction du risque d'accidents des assurés. Concernant la classification SRA du véhicule, c'est la classe qui présente une certaine discrimination en terme de sinistralité. Comme le montre la figure 3.12, plus on monte en classe, plus le risque d'accident augmente. Le groupe SRA quant à lui apparaît peu discriminant pour la fréquence de sinistralité des véhicules assurés même si les véhicules des groupes 37 à 39 sont ceux ayant connu le plus d'accidents.

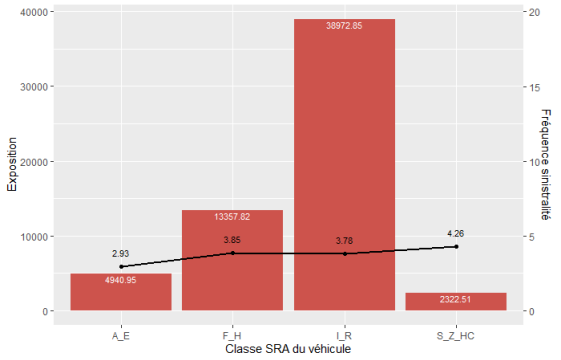


FIGURE 3.12 – Sinistralité selon la classe SRA du véhicule

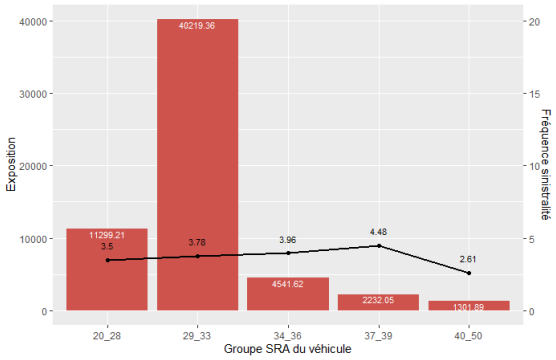


FIGURE 3.13 – Sinistralité selon le groupe SRA du véhicule

Sur la même base, nous mettons en évidence une sinistralité différente selon l'origine du contrat et le mode de fractionnement de la prime. Cela s'explique par le fait que les partenaires avec lesquelles la compagnie traite n'ont pas les mêmes conditions de souscription, ni les mêmes politiques de distribution. La figure 4.27 en annexe présente en dernier l'effet des antécédents d'assurance sur le risque de l'assuré. Il ressort que la sinistralité présente de l'assuré s'inscrit dans la logique de sa sinistralité passée.

3.2.3 Effet des caractéristiques géographiques sur la sinistralité

Nous présentons ici les variables extérieures choisies pour modéliser le risque géographique de la sinistralité. Le tableau 3.3 en donne le récapitulatif.

Variable	Source/Description	Modalités/amplitude
Région du lieu de garage	Insee	
Département du lieu de garage	Insee	
Grille de densité	Insee; Grille de densité de la commune	4 variables de densité; une variable typologie de densité
Type de commune	Insee	Urbaine/Rurale
Statut de la commune	Insee; Statut de la commune au sein de l'unité urbaine	Rurale;Ville-centre;Banlieue;Ville isolée
Nombre d'accident corporels	data.gouv.fr; survenus en 2017 et 2018	
Population active travaillant sur place à 4R	Insee	
Population active travaillant dans une autre commune à 4R	Insee	

TABLE 3.3 – Liste des variables géographiques extérieures utilisées

Le statut de la commune

Lorsqu'une unité urbaine est constituée d'une seule commune, on la désigne sous le terme de ville isolée. Lorsqu'une unité urbaine est constituée de plusieurs communes, on la désigne sous le terme d'agglomération multicommunale. Les communes qui la composent sont soit ville-centre, soit banlieue. Si une commune représente plus de 50% de la population de l'agglomération multicommunale, elle est seule ville-centre. Sinon, toutes les communes qui ont une population supérieure à 50% de celle de la commune la plus peuplée, ainsi que cette dernière, sont villes-centres. Les communes urbaines qui ne sont pas villes-centres constituent la banlieue de l'agglomération multicommunale. [1]

La grille de densité

Les communes étant de superficies très variables, certaines d'entre elles peuvent apparaître comme peu densément peuplées ou au contraire densément peuplées, alors même que leurs populations sont de tailles comparables. Pour prendre en compte la population communale et sa répartition dans l'espace, la grille communale de densité de l'Insee s'appuie sur la distribution de la population à l'intérieur de la commune en découpant le territoire en carreaux de 1 kilomètre de côté. Elle repère ainsi des zones agglomérées. C'est l'importance

1. cf. `https://www.insee.fr/fr/information/2115018` pour plus de détails

de ces zones agglomérées au sein des communes qui va permettre de les caractériser (et non la densité communale habituelle).

Cette classification reprend les travaux d'Eurostat, en introduisant une catégorie supplémentaire pour tenir compte des espaces faiblement peuplés, plus fréquents en France que dans d'autres pays européens. Ainsi, on distingue parmi les communes peu denses, des communes très peu denses. La grille communale permet ainsi de distinguer quatre catégories de communes :

- les communes densément peuplées ;
- les communes de catégorie intermédiaire ;
- les communes peu denses ;
- les communes très peu denses [2].

On peut voir sur les figures 2.4, 3.14 et 3.15 que la sinistralité est bien dépendante des caractéristiques géographiques des communes. Les régions Provence-Alpes-Côte d'Azur, la Corse, l'Occitanie sont les régions les plus sinistrées. La Brétagne est quant à elle la région comptant le moins de sinistres. Les communes qui sont des ville-centres ont aussi un risque plus élevé comparées aux autres communes. Concernant la densité, il apparaît que les communes denses présentent une sinistralité plus élevée.

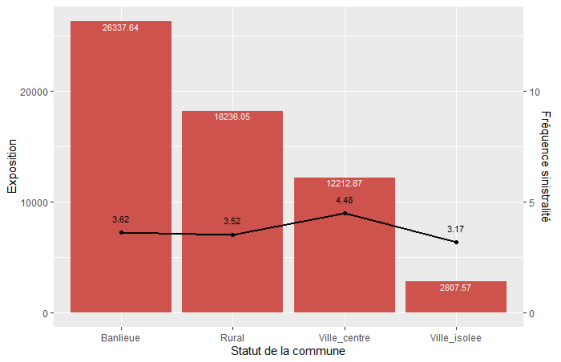


FIGURE 3.14 – Sinistralité selon le statut de la commune

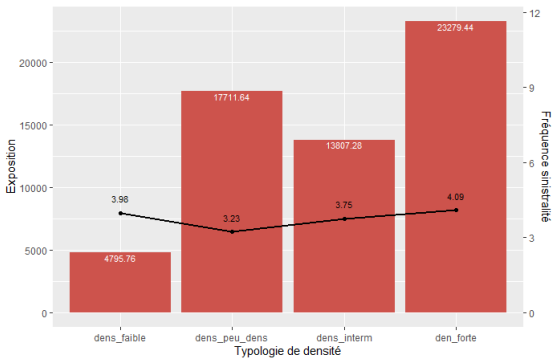


FIGURE 3.15 – Sinistralité selon la typologie de densité de la commune

2. cf. `https://www.insee.fr/fr/information/2114627` pour plus de détails

CHAPITRE 4

MODÉLISATION ET CONSTRUCTION DU ZONIER

Cette section est le coeur de notre étude. Ici, nous commençons par la modélisation de la fréquence des sinistres. Cette modélisation donnera les résidus qui seront ensuite lissés avant d'être modélisés par les variables extérieurs. La dernière étape sera alors de découper les résidus modélisés pour constituer le zonier fréquence.

4.1 Validation de l'approche fréquence × coût

Toute la méthodologie du zonier fréquence s'appuie ici sur l'approche de tarification fréquence × coût ; laquelle tient sur des hypothèses indispensables à sa validité. En définissant un zonier fréquence à part, nous comprenons donc que l'approche fréquence × coût est utilisable sur nos données ; c'est-à-dire la fréquence et le coût des sinistres sont modélisables séparément et en toute indépendance. La figure 4.1 présente le nuage de point de la charge et la fréquence. Elle permet de se faire une première idée de l'existence ou non de dépendance entre la fréquence des sinistres et leurs coûts.

Il ressort à la vue de la figure que la répartition des points est uniforme et qu'il ne dégage aucune tendance apparente ; ce qui nous conforte dans l'idée d'une indépendance entre la fréquence et le coût des sinistres.

Nous confirmons cela par le calcul des coefficients de corrélations (corrélation de Pearson

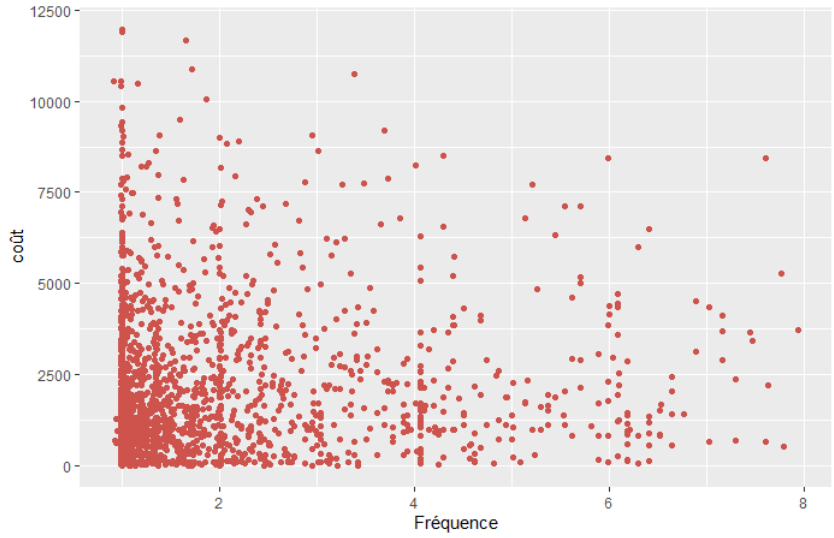


FIGURE 4.1 – Nuage des points : fréquence × coût

et Spearman). Le coefficient de corrélation de Pearson mesure la corrélation linéaire entre deux variables quantitatives. Il est donné par la formule :

$$r^{Pearson}(X,Y) = \frac{\mathbb{C}ov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X - \mathbb{E}[X]]^2}\sqrt{\mathbb{E}[Y - \mathbb{E}[Y]]^2}}$$

La corrélation de Spearman est la version non paramétrique de la corrélation de Pearson. Elle est plus robuste et permet de mieux mesurer le lien entre les variables en cas de lien non linéaire, et même en présence de valeurs atypiques. Elle est calculée par la même formule que celle de Pearson, à la différence que les valeurs de X et Y sont remplacées par leurs rangs. Les coefficients de corrélation varient de -1 à 1. Une valeur nulle témoigne d'une absence de corrélation. Une corrélation positive signifie que les deux variables varient dans le même sens et une corrélation négative signifient qu'elles varient dans le sens inverse.

Même si la nullité de ces coefficients de corrélation ne permet pas de conclure à une indépendance, elle nous rapproche d'une conclusion de ce genre. En effet, l'indépendance implique un coefficient de corrélation de Pearson nulle, mais la réciproque n'est pas toujours vraie si les deux variables ne sont pas gaussiennes. Nous obtenons 0.063 pour la corrélation de Pearson et 0.089 pour celle de Spearman ; preuves supplémentaires d'une grande probabilité d'indépendance entre la fréquence et le coût des sinistres.

Nous achevons l'analyse de l'indépendance entre la fréquence et le coût des sinistres par l'utilisation des copules. Les copules sont des outils statistiques qui modélisent la dépendance entre des variables aléatoires, permettant ainsi de considérer de manière distincte la structure

de dépendance décrite par la fonction de distribution conjointe et le comportement marginal des variables considérées. Pour notre cas ici nous utilisons la copule bivariée, le but étant de séparer dépendance et comportement marginal de la distribution bivariée de la fréquence et du coût des sinistres.

Une copule bivariée est une fonction $\mathrm{C} : [0,1]^2 \to [0,1]$ qui vérifie les propriétés suivantes :

- Pour tout $u \in [0,1], C(u,0) = C(0,u) = 0$ et $C(u,1) = C(1,u) = u$: les distributions marginales sont uniformes.
- C est 2 -croissante : $C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \geq 0$, pour tout $(u_1,u_2,v_1,v_2) \in [0,1]^4$ tels que $0 \leq u_1 \leq v_1 \leq 1$ et $0 \leq u_2 \leq v_2 \leq 1$

La copule produit, non paramétrique, est la plus adaptée pour notre étude puisqu'elle caractérise l'indépendance entre deux variables. Elle est définie par $C^{\perp}(u,v) = uv$.

Une copule peut être représentée graphiquement grâce à ses lignes de niveaux. L'analyse de la dépendance consiste alors à comparer les lignes de niveau de la copule empirique estimée sur la fréquence et la charge et les lignes de niveau de la copule théorique d'indépendance. La copule empirique est estimée par :

$$C_{X,Y}^{emp}(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{\frac{\mathrm{rg}X_i}{n} \leq u; \frac{\mathrm{rg}Y_i}{n} \leq v\right\}}.$$

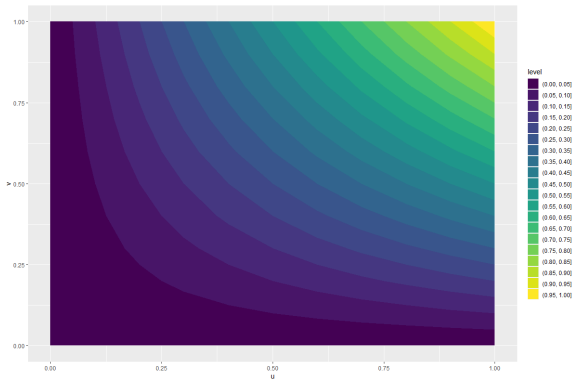


FIGURE 4.2 – Lignes de niveau de la copule d'indépendance

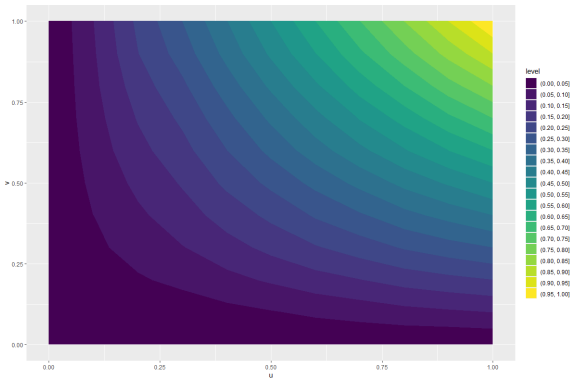


FIGURE 4.3 – Lignes de niveau de la copule empirique estimée

Il apparaît sur les deux figures que les lignes de niveau de la copule empirique sont assez similaires aux lignes de niveau de la copule théorique d'indépendance. L'indépendance entre la fréquence et le coût des sinistres est alors confirmée. L'utilisation de l'approche fréquence $\times$ coût est désormais justifiée et le zonier fréquence peut donc être construit.

4.2 Indépendance entre les variables explicatives

L'une des hypothèses fondamentales à la convergence et à la bonne estimation des modèles économétriques est l'absence de colinéarité entre les variables explicatives. La variance des résultats d'une modélisation avec des variables dépendantes est plus grande que la variance d'une modélisation où ce n'est pas le cas. Comme présenté dans le chapitre précédent, les variables explicatives retenues pour la modélisation de la fréquence sont toutes qualitatives. L'usage en assurance est d'utiliser ce type de variable car elles permettent ensuite de définir les différentes classes et segments au sein desquels le risque assuré est relativement homogène.

Le test d'indépendance du Khi-deux est l'outil statistique utilisé pour évaluer l'indépendance entre deux variables qualitatives. Ce test permet de vérifier l'absence de lien statistique entre deux variables qualitatives X et Y. Les deux sont dites indépendantes lorsqu'il n'existe aucun lien statistique entre elles, dit autrement, la connaissance de X ne permet en aucune manière de se prononcer sur Y.

L'hypothèse nulle (H0) de ce test est la suivante : les deux variables X et Y sont indépendantes. X et Y doivent prendre un nombre fini de modalités, supposons I pour X et J pour Y. Soit N la taille de l'échantillon. Le test consiste à comparer les effectifs réels des croisements des modalités des deux variables qualitatives avec les effectifs théoriques qu'on devrait obtenir dans le cas d'indépendance de ces deux variables. Pour cela, on construit un indice mesurant l'écart constaté entre les effectifs réels et les effectifs théoriques. Cet indice est la distance de Khi-deux défini définie par :

$$\chi = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où O_{ij} est l'effectif observé de données pour lesquelles $X = i$ et $Y = j$, $E_{ij} = \frac{O_{i+} \times O_{+j}}{N}$ désigne l'effectif théorique d'indépendance avec $O_{i+} = \sum_{j=1}^{J} O_{ij}$ (nombre d'observations pour lesquelles $X = i$) et $O_{+j} = \sum_{i=1}^{I} O_{ij}$ (nombre d'observations pour lesquelles $Y = j$). On montre que, sous l'hypothèse nulle, la statistique de test χ suit asymptotiquement une loi du χ^2 à $(I–1)(J–1)$ degrés de liberté. Au niveau de significativité α, on rejette H0 si χ est supérieure au quantile d'ordre $1 - \alpha$ de la loi du χ^2 à $(I–1)(J–1)$ degrés de liberté. On peut aussi calculer la p-value et rejeter H0 si p-value $< \alpha$.

Ce test est renforcé par le calcul du V de Cramer donné par la formule :

$$V = \sqrt{\frac{\chi}{N \times [\min(I, J) - 1]}}$$

La statistique du χ indique s'il existe ou non une liaison entre les deux variables et le V de Cramer indique si cette liaison est forte ou pas. Le V de Cramer varie de 0 (indépendance) à 1 (dépendance parfaite). Plus il est proche de 1, plus la liaison entre les deux variables étudiées est forte ; plus il est proche de 0, moins les variables étudiées sont dépendantes. Deux variables sont considérées fortement corrélées lorsque le V de Cramer est supérieur à 0.3. La figure suivante présente la matrice du V de Cramer entre les variables explicatives candidates pour le modèle de fréquence.

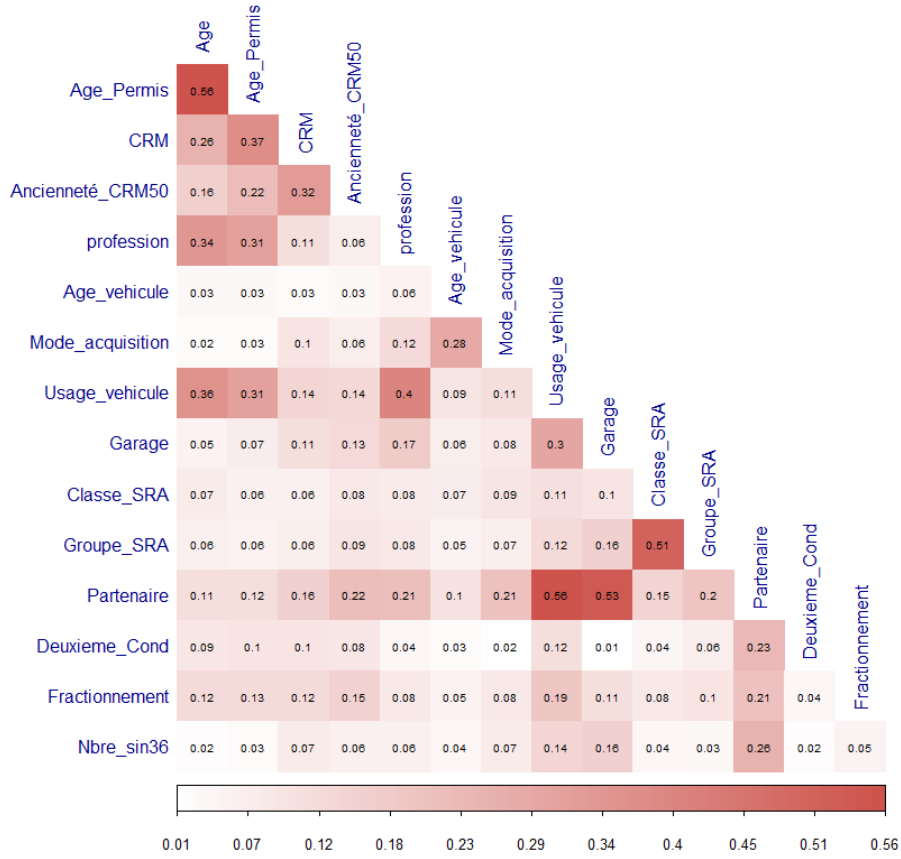


FIGURE 4.4 – Intensité des liaisons : Matrice du V de Cramer

Nous pouvons mettre en évidence dans ce tableau quelques liaisons fortes entres les variables en présence. Il y a le lien évident entre l'âge du conducteur et l'âge du permis avec un V de Cramer de 0.56. Il s'agit même d'un lien linéaire car l'âge du permis s'incrémente de la même manière que l'âge du conducteur. En plus on ne peut pas avoir 15 ans d'âge de permis avant 30 ans et peu d'assurés âgés de plus de 40 ans ont un permis de moins de 5

ans. (cf figure 4.28 en annexe) Seule l'une des deux variables sera donc conservée dans le modèle. Nous laissons la méthode stepwise de sélection de variables nous suggérer laquelle garder.

Nous remarquons aussi le lien évident qui se dégage entre la classe et le groupe SRA (V de Cramer de 0.51). Comme présentée dans le chapitre précédent, la classe SRA du véhicule est basée sur le prix de celui-ci. Le groupe quant à lui est définit la dangerosité du véhicule, et est calculé notamment grâce à la puissance fiscale, la vitesse maximale et la cylindrée. Or, ces trois éléments déterminent pour les constructeurs automobiles une grosse partie du prix de vente. Ces deux variables sont donc logiquement liées par construction. Nous laisserons aussi la procédure stepwise décider laquelle des deux variables apporte plus d'information au modèle.

Il y a ensuite la liaison forte entre le partenaire et l'usage du véhicule d'une part et entre le partenaire et le type de lieu de garage d'autre part (V de Cramer respectivement de 0.56 et 0.53). Cela s'explique beaucoup par le fait que le partenaire majoritaire dans le portefeuille apporte beaucoup de déséquilibre dans la répartition des usages du véhicule, si bien qu'un lien fictif se crée (cf. figures 4.29 et 4.30 et en annexe). Par ailleurs, les partenaires ont des politiques de souscription construites sur la base des autres caractéristiques du risque couvert. C'est d'ailleurs pour cela qu'on peut voir dans la matrice que le V de Cramer est particulièrement élevé entre le partenaire et les autres variables. La variable partenaire sera tout simplement supprimée dans la suite de la modélisation. Cela convient encore mieux à la politique de l'Équité qui se veut équitable dans la pratique de ces tarifs avec tous ses partenaires. En dernier lieu, nous devons aussi garder en tête que le zonier construit devra être utilisé de la même manière dans la modélisation faite par chaque partenaire. C'est à ce niveau que chaque zone se verra attribué un coefficient tarifaire différent en fonction de la politique de souscription du partenaire.

La liaison entre le CRM et l'anciennement au CRM50 (V de Cramer de 0.32) est aussi logique car elle s'articule autour de la première classe de CRM. L'ancienneté au CRM50 des assurés qui ont un CRM supérieur à 50 est presque nulle. Les deux variables seront conservées dans le modèle.

Il reste le lien entre la profession et l'usage du véhicule (V de Cramer de 0.4), qui s'explique aussi facilement par la figure 4.31 en annexe. D'un côté, les retraités, et les personnes sans emploi assurent majoritairement leur véhicule pour un usage privé. De l'autre, les sa-

lariés souscrivent à l'assurance pour un usage privé et pour le trajet du travail. La meilleure méthode pour garder l'information des deux variable est de les croiser dans le modèle. Cette méthode est d'ailleurs utilisée en interne pour les modèles utilisés.

4.3 Démarche pratique de validation des travaux

Nous tenons à préciser ici que les travaux ont été faits sur l'échantillon global de l'étude. En effet, étant donné la finalité du modèle qui est de construire un zonier, nous avons d'abord validé la démarche dans une logique de division de l'échantillon en échantillon d'apprentissage et test. L'échantillon d'apprentissage représente 70% de l'échantillon global. Pour cela, nous avons calibré le GLM sur l'échantillon d'apprentissage puis nous avons validé ses propriétés prédictives sur l'échantillon test. Le GLM ayant été ainsi validé, nous l'avons relancé sur l'échantillon global afin d'avoir les résidus et les utiliser comme input dans les modèles de machine learning utilisés pour construire le zonier.

Les modèles de machine learning ont ensuite été validés suivant le même procédé que celui du GLM ; c'est à dire après division des résidus par commune en échantillon test et d'apprentissage. C'est après validation du modèle final de machine learning (Random Forest) utilisé pour construire le zonier que celui ci a été relancé avec l'échantillon global de toutes les communes présentes dans les données. Ce sont finalement ces résidus prédits par communes qui ont été utilisés dans le krigeage pour le lissage et l'interpolation avant découpage en classe pour constitution du zonier. Le krigeage a été fait sur l'échantillon global car tous les modèles qui nous y ont conduit ont été bien validés. Nous avons également choisi de faire le krigeage sur l'échantillon global parce qu'étant une méthode d'interpolation, en utilisant toutes les données, nous arrivons à la construction d'un zonier plus exhaustif et plus fiable.

4.4 Calibrage de la loi du nombre de sinistres

4.4.1 Comparaison des modèles estimés

Disposant de la loi binomiale, la loi de poisson et la loi binomiale négative, nous avons utilisé le critère basé sur le rapport variance/moyenne pour faire premier tri. Le rapport variance/moyenne donne comme valeur 1.04, ce qui soupçonne une sur-dispersion, éliminant directement la possibilité que les fréquences soient distribuées selon la loi binomiale (pour

laquelle la variance est plus petite que la moyenne).

La surdispersion renvoie beaucoup plus à la loi binomiale négative mais étant donnée la faible valeur du rapport nous testons l'hypothèse de surdispersion par la méthode du modèle linéaire. La méthode consiste à choisir une variable définissant des classes de risque et d'estimer un modèle linéaire de la variance sur la moyenne avec constante nulle. On teste ensuite l'égalité à 1 du coefficient de la moyenne. Nous choisisons les variables âge et département pour les classes de risque. En regardant sur les figures 4.33 et 4.34 (cf annexe), la surdispersion apparaît légèrement mais semble négligeable. Le test de surdispersion du nombre de sinistre par âge stipule qu'on peut, au seuil de 5% de se tromper, accepter que la moyenne et la variance sont égale, et donc la surdispersion est rejetée. Par contre en considérant les départements comme classe de risque, le test met en évidence la présence de surdispersion au seuil de 5%. Le modèle le plus approprié pour les fréquence est alors le modèle binomial négatif confirmé d'ailleurs par les tests d'adéquation du Khi-deux. Les tableaux 4.5 et 4.6 en annexe présente les résultats d'adéquation aux deux lois. Nous testons tout de même le modèle de poisson.

Pour prendre en compte le maximum d'information apportée par les vaiables explicatives tout en essayant d'être parcimonieux, nous lançons d'abord le modèle avec toutes les variables décrites plus haut. Pour chacun des deux lois que l'on se propose de tester, la sélection des variables pertinentes pour modéliser la fréquence est faite par une procédure stepwise basée sur l'AIC. Les deux modèles ont mis en évidence les mêmes variables pertinentes, ils sont donc comparables. Ces caractéristiques pertinentes pour expliquer le risque non géographique de la fréquence des sinistres sont : l'âge du permis, le CRM, la profession, l'âge du véhicule, l'usage du véhicule, la classe SRA, la présence d'un deuxième conducteur, le nombre de sinistre au cours des 36 derniers mois. Le tableau 4.1 donne le récapitulatif des deux modèles finaux à comparer.

Table 4.1 – Résumé des modèles de fréquence testés

Modèle	Lien	AIC	BIC	log-vraisemblance	deviance	RMSE	Gini
Poisson	log	21609	21854.53	-10779.5	17172.29	0.12	0.51
Bin neg	identité	21586.01	21843.36	-10768	15444	0.12	0.51

Nous pouvons dire que le modèle binomial négatif est le modèle le plus approprié car il minimise les critères *AIC* et *BIC*, la log-vraisemblance et la déviance. Le modèle est globalement significatif (p-value= 4.686e-10).

Dans la suite de ce mémoire, l'indice de Gini et le RMSE ne seront plus utilisés pour comparer les modèles GLM sur la fréquence. En effet, les valeurs de la fréquence variant très peu et sur un très petit échelle, l'indice de Gini et le RMSE varie très peu avec le modèle de prédiction utilisé. Ceci rend leur utilisation assez peu discriminante comme critère de sélection de modèle.

4.4.2 Validation du modèle choisi

Avant d'extraire les résidus du GLM afin de construire le zonier, nous procédons d'abord à la validation dudit modèle. La validation du modèle se fait en étudiant les résidus et les fréquences prédites. Les résidus représentent l'estimation du bruit non observable lors d'une modélisation. Ils mesurent donc l'écart entre la fréquence observée de sinistres et la fréquence prédite et sont définis par :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

pour un assuré i, en notant y la fréquence observée et $\hat{y}_i$ la fréquence prédite par le modèle.

Dans un premier temps, nous cherchons à valider les hypothèses fondamentales sur lesquelles repose le GLM. Ce sont la linéarité de $g\left(\mathbb{E}\left[Y|X\right]\right)$ en fonction des prédicteurs et l'homoscédasticité des résidus. On peut valider ces hypothèses graphiquement en représentant la valeur des résidus en fonction des valeurs prédites. Si l'hypothèse de linéarité n'est pas respectée on peut voir apparaître dans le graphique une tendance. Le non respect de l'hypothèse d'homoscédasticité est quant à lui identifié par une forme conique du graphique.

Nous étudions dans un premier temps les résidus groupés, appelés aussi « crunshed residuals ». Le but est de regrouper en paquets d'expositions homogènes les résidus du modèle. Ces paquets sont faits à partir des valeurs prédites triées dans un ordre croissant, afin que ceux-ci soient homogènes en termes de risque prédit.

Nous n'observons pas de tendance particulière dans les graphes ci-dessus. Malgré quelques valeurs s'écartant du nuage de résidus pour les classes de risques inférieures qui correspondent aux fréquences prédites faibles, les graphiques semblent confirmer la bonne adéquation du modèle. Sur le graphique de droite, les résidus sont représentés en fonction de leurs classes d'expositions parce que cela permet de réordonner le nuage de point et de mieux mettre en évidence les formes de la représentation. Cette représentation est plus lisible et nous voyons bien que nos résidus sont répartis de manière homogène. On peut ainsi valider l'homosce-

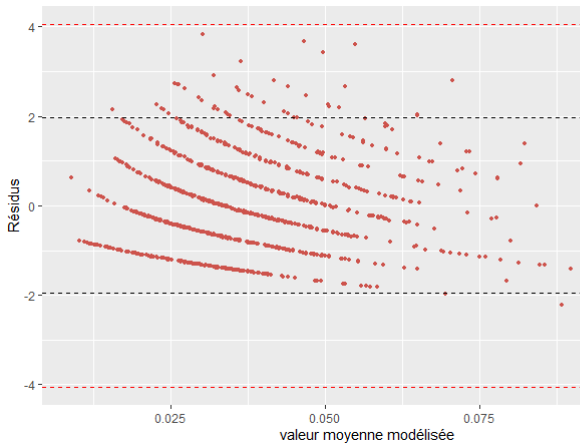


FIGURE 4.5 – Résidus en fonction des valeurs moyennes prédites

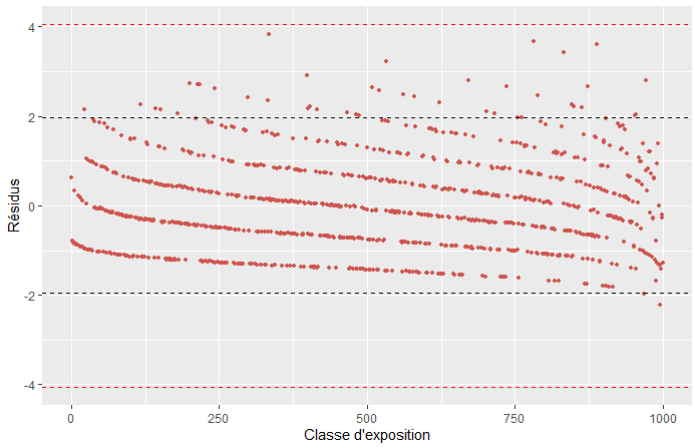


FIGURE 4.6 – Résidus en fonction des classes d'exposition

dasticité et la linéarité de $g(\mathbb{E}[Y|X])$.

Par ailleurs, nous remarquons aussi que la plupart des résidus a des valeurs entre -2 et +2, avec quasiment aucune valeur au-delà de +4 ou en dessous de -4. Ceci voudrait dire que les résidus pourraient aussi suivre une loi normale centrée réduite. Nous le confirmons en regardant l'adéquation de la fonction de répartition des résidus à celle d'une loi normale standard comme le montre la figure 4.8 suivante.

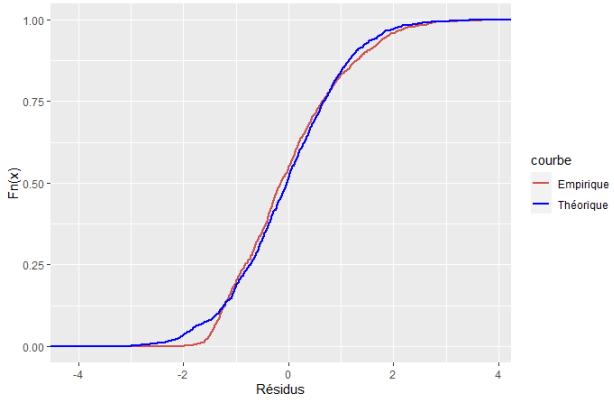


FIGURE 4.7 – Fonction de répartition des résidus : adéquation à la loi normale

L'approximation de la distribution des résidus groupés par une loi normale semble correcte. Nous validons ainsi l'analyse des résidus du modèle de fréquence avant de passer à l'analyse spatiale en prélude à la construction du zonier.

4.5 Analyse géographique des résidus

Les résidus géographiques sont définis par :

Résidus géographiques Fréquence = Fréquence observée - Fréquence prédite non géographique où la fréquence observée est définie comme à la page 37. Il s'agit donc de valeurs définies sur $\mathbb{R}$ et centrée sur 0.

Il faut à ce niveau choisir quel type de résidus utiliser. Nous disposons des résidus bruts sortis tels quels de la modélisation et des résidus standardisés. Parmi les résidus standardisés, nous avons les résidus de Pearson, les résidus de déviance et les résidus d'Anscombe.

En notant μ_i, l'espérance de la fréquence et $\hat{\mu}_i$ l'espérance prédite, les résidus de Pearson sont définis par

$$\hat{\epsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathbb{V}ar\left(\hat{\mu}_i\right)}}$$

Par définition ils suivent la loi normale (théorème central limite).

En reprenant les mêmes notations, les résidus d'Ansombe sont définis par :

$$\hat{\epsilon}_i^A = \frac{A\left(y_i\right) - A\left(\hat{\mu}_i\right)}{A'\left(y_i\right)\sqrt{\mathbb{V}ar\left(\hat{\mu}_i\right)}}$$

où A est une transformation pour que les résidus s'approche le plus possible d'une loi normale centrée réduite.

Les résidus de déviance sont définis par :

$$\hat{\epsilon}_i^D = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{d_i}$$

où $d_i = \frac{2y_i(\theta(\hat{\mu}_i) - \theta(y_i)) - b(\theta(\hat{\mu}_i)) + b(\theta(y_i))}{a(\phi_i)}$ est le ième terme de la déviance, θ, b, a et ϕ définis dans le GLM (page 14).

Dans le soucis de conserver l'intégralité de l'information apportée par les résidus, nous utilisons les résidus bruts ; l'objectif étant en effet de construire le zonier à partir de l'information non modélisée par le GLM. Par ailleurs l'utilisation des résidus bruts permet de simplifier l'interprétation des résultats dans la partie du lissage des résidus. Le tableau suivant présente les principales statistiques sur les résidus.

Ils présentent parfois de grandes valeurs positives puisque les fréquences observées peuvent exploser par rapport aux fréquences prédites. Par ailleurs, ces résidus ayant été obtenus sans prendre en compte l'aspect géographique du risque, il est normal d'avoir des valeurs élevées.

TABLE 4.2 – Caractéristiques des résidus géographiques

Min.	q25	Méd	Moy.	q75	Max.
-0.12427	-0.04551	-0.03560	-0.02168	-0.02729	3.92944

Dans la suite, nous agrégeons les résultats à la maille commune par la moyenne. Le tableau 4.4 présente les caractéristiques des résidus moyens par commune. Les résidus sont ensuite lissés par krigeage. L'objectif du krigeage est de réduire la variabilité entre les communes voisines avant de procéder à la modélisation des résidus par les variables extérieurs géographiques qui par leur structure intègrent déjà un lissage naturel. Le krigeage permet aussi de faire l'interpolation des résidus aux communes de France qui n'étaient pas représentées dans les données.

TABLE 4.3 – Caractéristiques des résidus géographiques moyens par commune

Min.	q25	Méd	Moy.	q75	Max.
-0.10358	-0.03982	-0.03380	-0.02120	-0.02402	1.92633

La carte de la figure 4.8 représente les résidus sur toute la France. Les zones vertes sont les communes moins risquées et les zones rouges sont les communes à risque.

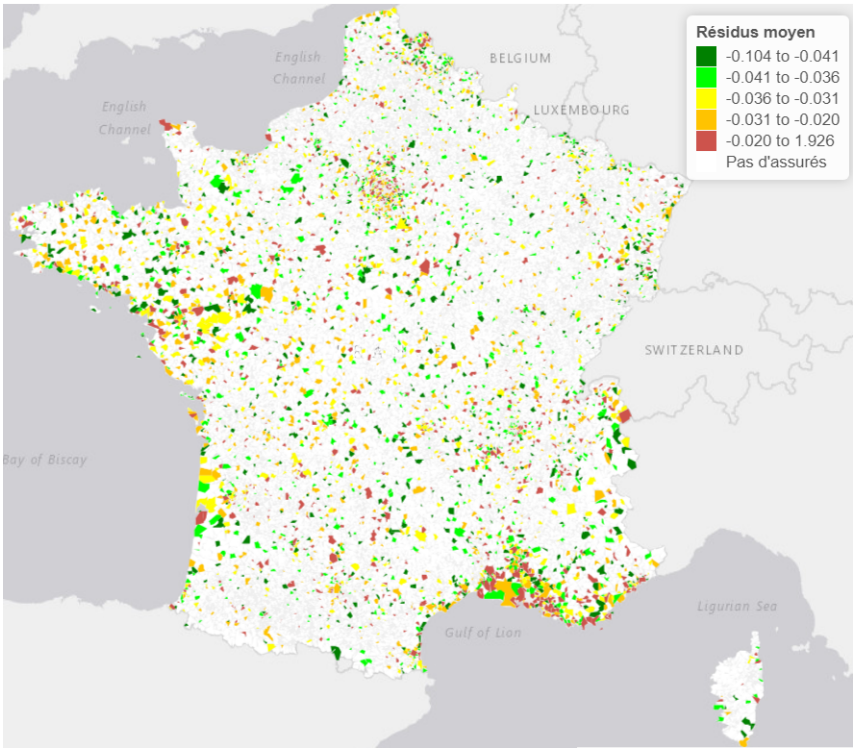


FIGURE 4.8 – Carte des résidus géographiques non lissés

De grandes tendances sont observables sur la carte et suivent les mêmes répartitions du

risque de sinistralité qui ont été mises en évidence plutôt dans ce travail. La carte fait aussi ressortir le caractère peu uniforme des résidus de zones voisines. Cela justifie la nécessité de faire un lissage avant toute modélisation de ces résidus. Le lissage permettra non seulement d'interpoler des résidus aux zones blanches de la carte, mais également de mieux faire ressortir la structure spatiale qui pourrait se dégager des résidus ; structure spatiale que nous cherchons à capter pour construire notre zonier.

4.5.1 Autocorrélation spatiale et lissage par krigeage

La modélisation des résidus géographiques par les méthodes classiques ne peut se faire si la structure d'autocorrélations spatiales n'est pas prise en compte. L'autocorrélation mesure la corrélation d'une variable avec elle-même, lorsque les observations sont considérées avec un décalage dans le temps (autocorrélation temporelle) ou dans l'espace (autocorrélation spatiale). On définit alors l'autocorrélation spatiale comme la corrélation, positive ou négative, d'une variable avec elle-même du fait de la localisation spatiale des observations.

Dans le contexte de la spécification de modèles économétriques, la mesure de l'autocorrélation spatiale peut être envisagée comme un outil de diagnostic et de détection. En présence d'autocorrélation spatiale, on observe que la valeur d'une variable pour une observation est liée aux valeurs de cette même variable pour les observations voisines. L'analyse de l'autocorrélation spatiale des résidus nous permet de faire une analyse quantifiée de la structure spatiale du risque de sinistralité non captée par les variables tarifaires utilisés dans le GLM. Le semi-variogramme est un outil qui permet alors de décrire la corrélation spatiale entre observations spatiales. Il est définit par :

$$\gamma(h) = \frac{1}{2\#N(h)} \sum_{(i,j)\in N(h)} (z_i - z_j)^2$$

Avec $N(h)$ l'ensemble des couples de communes séparés d'une distance $h, \#N(h)$ son cardinal, et les z_i sont les valeurs des résidus des communes.

Lorsque chaque paire d'emplacements a été tracée (on parle alors de semi-variogramme empirique), un modèle est ajusté à travers ces derniers (c'est le modèle de variogramme théorique). En fonction de la forme du semi-variogramme empirique, plusieurs modèles théoriques de semi-variogramme sont proposés dans la littérature : le modèle exponentiel, le modèle sphérique, le modèle gaussien, le modèle puissance, etc. (cf. J-M Floch, 2012). Les

caractérisques utilisées pour décrire ces modèles de variogramme sont la portée, le palier (aussi appélée seuil) et l'effet de pépite.(cf. figure 4.9 pour leur illustration)

La portée : En observant un modèle de semi-variogramme, l'on peut remarquer qu'à une certaine distance le modèle se stabilise. La distance à laquelle le modèle commence à s'aplanir est appelée la portée. Les emplacements d'échantillons séparés par une distance inférieure à la portée sont auto-corrélés spatialement alors que les emplacements séparés par une distance supérieure à la portée ne le sont pas.

Le palier : La valeur que le semi-variogramme atteint à la portée (la valeur de l'axe des ordonnées) est appelée le palier.

L'effet de pépite : Théoriquement, à une distance de séparation nulle, la valeur du semi-variogramme est 0. Toutefois, à une distance de séparation infiniment petite, le semi-variogramme présente souvent un effet de pépite, qui représente une valeur supérieure à 0. Par exemple, si le modèle de semi-variogramme intercepte l'axe des y à la valeur 1, la pépite prend alors la valeur 1. L'effet de pépite peut être attribué à des erreurs de mesure ou à des sources de variation spatiale à des distances inférieures à l'intervalle d'échantillonnage, ou aux deux. Dans notre contexte, l'effet de pépite s'il y en a correspondrait à des différences de résidus entre communes non expliquées par la structure géographique du risque.

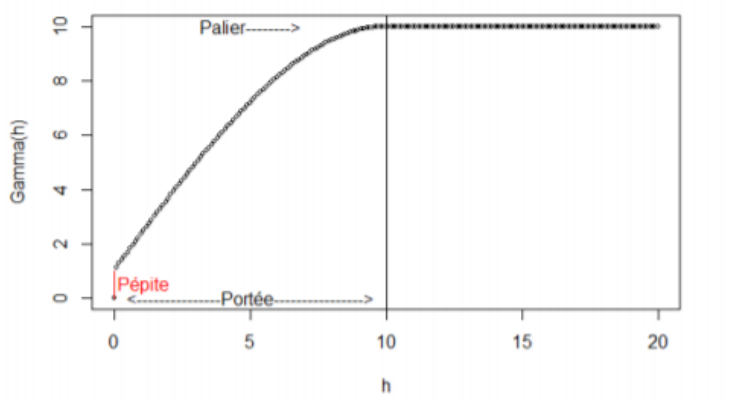


FIGURE 4.9 – Caractéristiques d'un modèle de variogramme

Dans notre cas, le modèle de variogramme théorique exponentiel a été utilisé car il semblait être mieux adapté à la forme du semi-variogramme empirique. Les résultats de l'analyse de l'autocorrélation spatiale des résidus sont présentés dans la figure 4.10

Nous pouvons voir dans le semi-variogramme qu'il y a un palier dont la portée est entre 5 et 10km. L'aspect de croissance entre 0 et 10km nous permet de confirmer la présence d'auto corrélation spatiale dans les résidus géographiques. On note également que le semi-

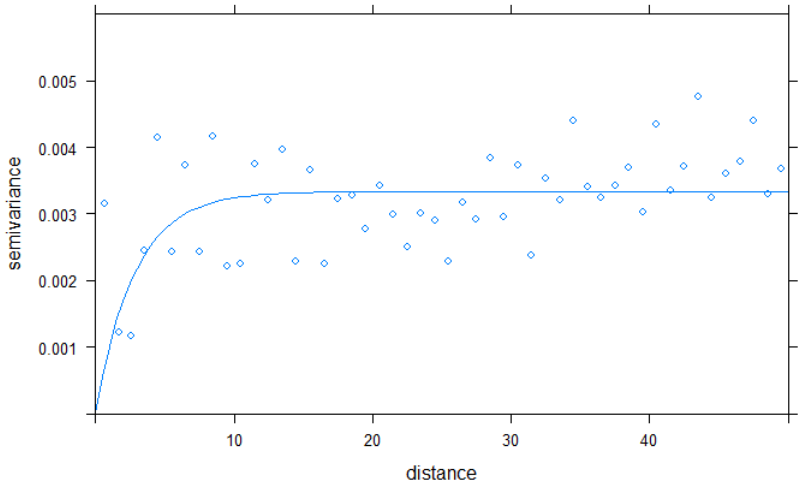


FIGURE 4.10 – Semi-variogramme théorique et empirique des résidus

variogramme théorique s'ajuste assez bien au variogramme empirique. Le modèle théorique exponentiel choisit est donc adéquat et il est bien utile pour le lissage par la méthode de krigeage.

Il n'existe pas un unique modèle de krigeage, bien qu'ils considèrent tous que la variable modélisée (ici les résidus) peut être séparée en deux composantes principales : une tendance déterministe et une erreur auto-corrélée. Si l'on note $z(s)$ le résidu observé à une position géographique s, alors on peut écrire $z(s) = m + e(s)$, où m est une tendance déterministe, et $e(s)$ un terme d'erreur auto-corrélée. On distingue notamment deux modèles de krigeage : le grigeage simple et le krigeage ordinaire.

Le krigeage simple : c'est le modèle de krigeage le plus global étant donné qu'il ne prend pas en compte de variations locales de la tendance déterministe : les variations restent constantes sur l'ensemble de la France.

Le krigeage ordinaire : Il considère que la tendance est constante mais seulement par morceau. Autrement dit, la tendance est constante au niveau d'un voisinage et non plus sur l'ensemble de la France. Nous utilisons le krigeage ordinaire pour le lissage. La carte de la figure 4.11 représente les résidus lissés par krigrage ordinaire.

Nous pouvons remarquer une répartition désormais plus homogène des résidus. Nous retrouvons aussi la répartition de la sinistralité telle que nous la décrivions plus tôt dans cette étude.

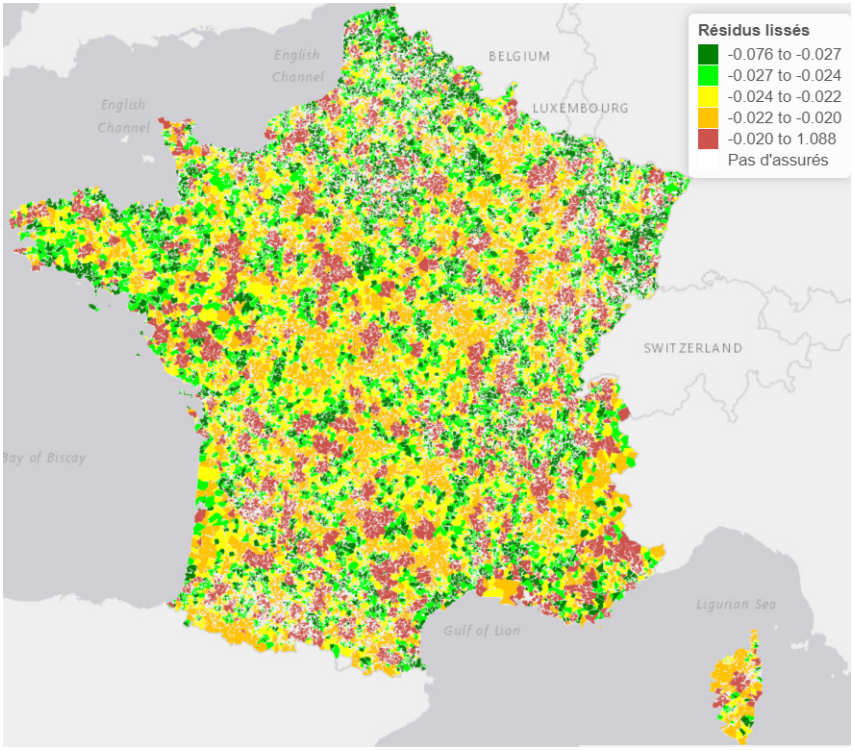


FIGURE 4.11 – Carte des résidus géographiques lissés

4.6 Modélisation des résidus géographiques par les variables externes

Les résidus étant déjà lissés, ici nous les modélisons par les caractéristiques géographiques des communes correspondants. Étant donnée la structure des données, les modèles de machine learning sont plus adaptés pour ce type de modélisation. C'est pourquoi les modèles d'agrégation d'arbre de décision sont utilisés ici. Le random forest (forêt aléatoire) agrège plusieurs arbres construits simultanément alors que le xgboost construit itérativement les arbres de régression dans une logique d'optimisation à chaque nouvel arbre ajouté.

Nous choisissons en premiers les hypers paramètres fondamentaux sur lesquels il faut jouer pour optimiser les résultats de ces deux modèles. Présentons dans la figure suivante l'évolution de l'erreur out-of-bag du random forest en fonction du nombre d'arbre du modèle. Il ressort que pour la forêt aléatoire, le nombre optimal d'arbre à conserver est de 300 arbres.

En utilisant cela comme entrée de modèle, les autres paramètres ont été choisies par une approche de validation croisée ainsi que ceux du XGBoost. En comparant les RMSE et le temps de calcul nécessaire, le meilleur modèle parmi les deux pour modéliser les résidus s'est révélé être le random forest avec 300 arbres de profondeur 9 dont la taille minimale

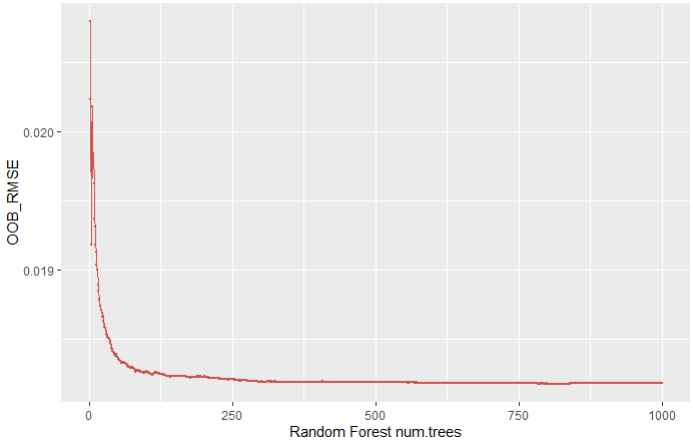


FIGURE 4.12 – Choix du nombre d'arbres de la forêt aléatoire

des feuilles est 5. Le Random forest a en effet un RMSE de 0.016 et tourne assez vite alors que le XGBoost a un RMSE de 0.017 pour un temps de calcul élevé. Les figures 4.13 et 4.14 suivantes présentent l'importance des variables pour le Random Forest et le XGBoost.

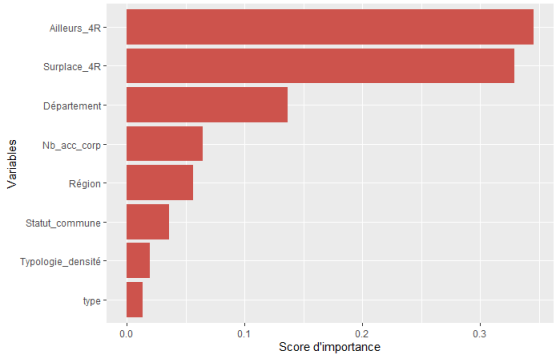


FIGURE 4.13 – Importance des variables du Random Forest

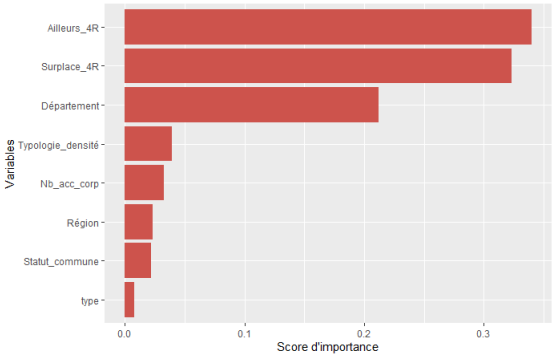


FIGURE 4.14 – Importance des variables du XGBoost

Les résultats montrent que la mobilité des habitants d'une commune détermine fortement le niveau de risque géographique de sinistralité de la commune. L'on peut en effet voir que les deux variables les plus importantes dans les deux modèles sont le nombre de personnes de la commune allant au travail en véhicule quatre roues, qu'elles travaillent sur place ou dans une autre commune. Le nombre de personnes allant travailler dans une autre commune est positivement corrélé au risque d'accidents de la commune. La part importante de l'information non captées par les variables tarifaires classiques proviendraient donc de la mobilité des habitants. L'on pouvait s'attendre à ce résultat car la mobilité accru en véhicule motorisé dans une commune définit directement la densité de circulation dans la commune, ce qui augmente le risque d'accident dans la zone.

L'une des variables qui apparaît comme variable importante pour modéliser le risque géographique des assurés est le nombre d'accident corporels enregistrés dans la commune de 2017 à 2018. Cette variable apporte des informations sur l'historicité globale de l'exposition d'une commune aux accidents de la route. Le fait qu'elle se révèle importante ne nous surprend donc pas, mais nous conforte davantage sur le fait que la prise en compte des antécédents d'assurance même au niveau individuel des assurés est une bonne idée.

Les deux autres variables qui ressortent en terme d'importance pour la modélisation du risque géographique sont le département et la région dans le cas du Random Forest. Leur présence ici montre qu'elles permettent de capter les informations globales à l'échelle régionale ou départementale non rapportées par les variables construites à l'échelle de la commune. Ces deux variables complétant les deux précédentes qui étaient à l'échelle de la commune nous permettent dès lors de nous rassurer sur le caractère robuste du zonier qui sortira de cette modélisation. Les éventuelles corrélations qui existeraient si ces variables étaient introduites en l'état dans le modèle de tarification seront désormais moins fortement remarquées par l'introduction plutôt du zonier. Ce dernier intégrera en effet une information plus robuste car composite de plusieurs informations brutes.

Globalement nous pouvons dire que les deux modèles (Random Forest et XGBoost) ressortent les mêmes tendances et que choisir l'un ou l'autre pour modéliser les résidus géographiques ne ferait pas trop varier le résultat. Nous sommes donc confortés dans le choix du Random Forest qui est le plus rapide des deux modèles en termes de temps de calcul.

4.7 Découpage des résidus en classe : le zonier construit

Deux méthodes de découpages sont mis en concurrence ici. La méthode des Kmeans et la méthodes des quantiles. Le choix du nombre optimal de classes est fait en regardant l'évolution de deux métriques selon le nombre de classes : la métrique silhouette et la méthode du gap statistic.

Notons que ces deux métriques peuvent conduire à un nombre de classes différent. Ici nous choisissons 6 classes en se basant sur la méthode du gap statistic. La statistique atteint son plus bas niveau pour un découpage en 6 classes.(cf. figure 4.15).

Cependant, en observant la répartition des 6 classes découpées par k-means, nous voyons

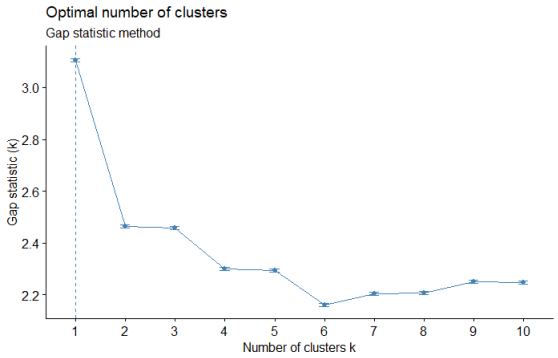


FIGURE 4.15 – Choix du nombre de classes par la méthode de la gap statistic

que les effectifs présentent un déséquilibre important. C'est d'ailleurs une limite du Kmeans, la méthode étant adapté pour le cas où on voudrait constituer des cluster sur la base de plusieurs variables quantitatives. C'est ainsi que pour constituer le zonier, nous avons mis en contribution cette méthode avec la méthode des quantiles. Le zonier est donc construit par un découpage en 6 classe des résidus prédits par le random forest, découpage fait par la méthode des quantiles.

4.8 Analyse de la pertinence du zonier

On peut de manière assez simple apprécier la qualité d'un zonier par son pouvoir discriminant sur la fréquence des sinistres. Et pour ce point, la figure 4.16 nous montre bien la segmentation qui se dégage entre les assurés selon les zones où ils sont affectés.

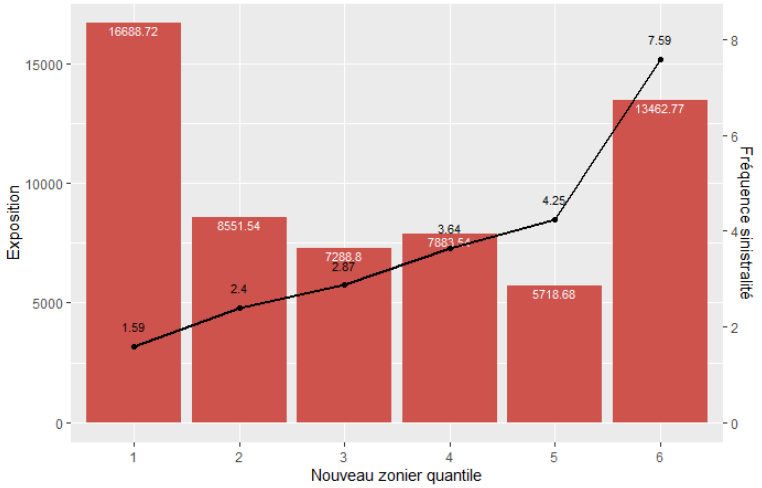


FIGURE 4.16 – Sinistralité selon le nouveau zonier

Nous allons plus loin en introduisant maintenant les deux zoniers à tour de rôle dans le modèle initial de fréquence. On voit directement au niveau des coefficients estimés de

l'ancien zonier qu'il n'y a que la zone 6 qui est significative. Concernant le nouveau zonier, tous ses coefficients sont significatifs. Le tableau suivant résume les performances des trois modèles. Le modèle avec le nouveau zonier est incontestablement le meilleur modèle.

TABLE 4.4 – Résumé des modèles avec et sans zonier

Modèle	AIC	BIC	log-vraisemblance	deviance
Bin neg sans zonier	21586.01	21843.36	-10768	15444
Bin neg avec ancien zonier	21556.1	21852.74	-10749.05	15463.79
Bin neg avec nouveau zonier	20867.29	21173.74	-10403.64	15483.15

4.9 Apport du zonier dans la tarification

Pour l'utilisation du zonier dans la tarification au sein de la compagnie, il sera pris en compte dans le modèle de fréquence des sinistres des partenaires. Chaque partenaire a en effet son modèle de tarification, qui par l'approche fréquence-coût comporte un modèle de fréquence et un modèle de coût. Le zonier étant un zonier fréquence, il sera utilisé dans le modèle de fréquence du partenaire de sorte à avoir un tarif plus optimisé.

Nous illustrons l'utilisation du zonier construit en estimant un modèle de prime pure. Pour cela nous faisons une modélisation rapide des coûts moyen et obtenons ainsi la prime pure à travers la formule de l'approche « fréquence × coût» :

$$\text{Prime pure} = \mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[B]$$

où N représente la fréquence et B le coût. Nous parlons d'une modélisation rapide des coûts parce que contrairement au modèle de fréquence qui a fait l'objet principal de ce mémoire et qui, pour cela, a été calibré et optimisé avec soin, le modèle de coût a été estimé par un GLM de loi Gamma, avec sélection automatique de variables tarifaires construites initialement pour modéliser la fréquence. Nous avons fait un écrêtement simple des coûts en considérant le seuil supérieur équivalent au quantile d'ordre 99.5% (19145.23€). Le graphe de la figure 4.17 représente la prime pure selon les zones du nouveau zonier.

Il ressort de la figure que le zonier intégré au modèle de fréquence a bien permis d'introduire une bonne segmentation de la prime pure des assurés. La prime pure est bien positivement corrélée au niveau de risque de sinistralité attribué à la zone. L'on peut également remarquer que le tarif prenant en compte également d'autres caractéristiques, les limites

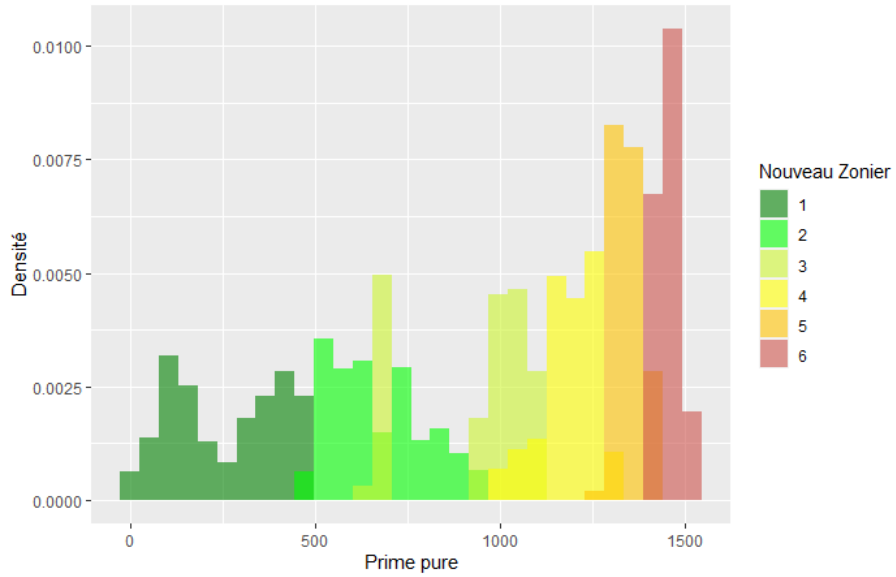


FIGURE 4.17 – Histogramme de la prime pure selon le nouveau zonier

entre les zones ne sont pas exclusives. Des assurés de zones plus risqués peuvent avoir des tarifs plus bas que des assurés de zones moins risqués. Ceci peut s'expliquer par le fait que ces personnes dans les zones risqués ont d'autres caractéristiques moins risquées, par exemple ils sont majoritairement des conducteurs expérimentés (âge du permis élevé).

On pourrait pousser plus loin cette optimisation en construisant un zonier coût moyen et l'intégrant aussi au modèle de coût. Ainsi, le tarif qui en sortira sera optimisé à la fois pour les fréquences et les coûts moyens.

Le nouveau zonier est présenté dans la carte de la figure 4.18 à la page suivante. Nous pouvons noter qu'il reprend bien l'essentiel des analyses faites sur la répartition géographique de la sinistralité des assurés.

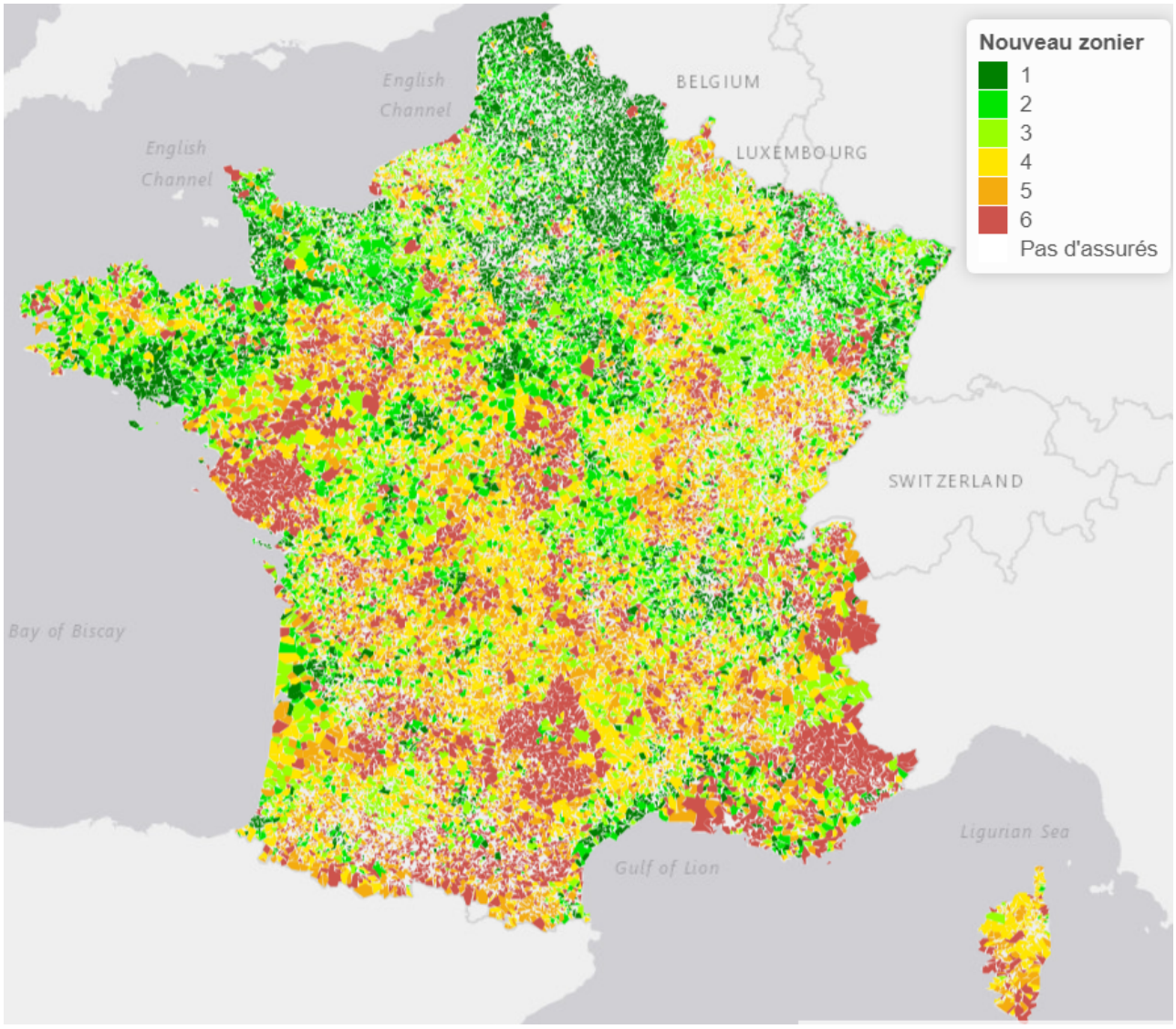


FIGURE 4.18 – Cartographie du territoire de la France avec le nouveau zonier

CONCLUSION

Par son importance au sein de la branche non-vie des compagnies d'assurance et le contexte concurrentiel de son marché, l'assurance automobile fait l'objet d'une attention particulière des acteurs du marché. La croissance de la masse de données disponibles et le développement de nouvelles techniques de tarification sont parmi les enjeux clés de ce marché. C'est ainsi que l'intégration des données géographiques externes pour sophistiquer les tarifs proposés devient une solution de choix pour les assureurs, dont la direction de partenariat de Generali où j'ai effectué mon stage.

L'objectif de ce stage était de construire un zonier pour optimiser la tarification de la garantie DTA au sein du portefeuille de l'Équité. La méthode utilisée est l'approche résiduelle partant d'une modélisation complète de la fréquence des sinistres. Cette étude a abouti à des résultats satisfaisants et a permis de répondre aux besoins de sophistication des produits distribués par l'Équité. Les analyses exploratoires des risques assurés ont été très instructives, et ont servi de point de départ aux choix des variables qui ont été utilisées par la suite.

Au terme de la modélisation de la fréquence, les facteurs discriminants de la sinistralité des assurés se répartissent en trois familles. Dans la première famille, nous avons les caractéristiques de l'assuré : l'âge de son permis, son coefficient bonus-malus, sa profession et le nombre d'accidents qu'il a eu au cours des 36 derniers mois. Dans la deuxième famille, il y a les caractéristiques du véhicule qui sont l'âge du véhicule, et la classe SRA. Le dernier groupe est constitué des caractéristiques du contrat. Il s'agit de l'usage du véhicule et de la présence ou non d'un second conducteur.

Le zonier qui a ensuite été construit en utilisant les résidus géographiques issus du modèle précédent a permis de gagner en pouvoir explicatif du modèle. Il a en effet permis de mieux capter le risque de sinistralité du portefeuille et pourra servir à rendre les primes DTA qui

en découleront plus concurrentielles.

Cette étude n'est toutefois pas exhaustive sur le sujet traité et présente des limites. L'absence de certaines caractéristiques géographiques clés peut être considérée comme une limite majeure du zonier construit. Il serait par exemple intéressant d'avoir sur les communes des informations comme la couverture en route, les caractéristiques de son réseau routier, le nombre de parcs de stationnement et leurs caractéristiques, etc. Des méthodes de construction de zonier auto fondés non plus seulement sur la commune du lieu de garage mais également sur le déplacement du véhicule peuvent constituer des pistes d'amélioration. Le manque de puissance de calcul constitue aussi une limite importante à cette étude car elle nous aurait permis de paramétrer plus finement nos modèles, surtout pour ce qui concerne le lissage spatial par krigeage. Par ailleurs, en raison de la richesse de la composante géographique du risque d'accident et notamment des interactions entre les facteurs dont il dépend, des études peuvent être menées afin de mieux comprendre sa répartition, et sa sensibilité à l'évolution des populations en général et des assurés en particulier.

BIBLIOGRAPHIE

[2017] S. KRANZLIN ; Modélisation du risque géographique en assurance automobile ; mémoire IA.

[2016] B. LE BOUCHER ; Tarification IARD et Open Data ; mémoire IA.

[2015] S. SAID ; Refonte des modèles de tarification de l assurance automobile et création de zoniers tarifaires ; mémoire IA.

[2005] A. CHARPENTIER, M. DENUIT Michel ; Mathématiques de l'assurance non vie : Tome 2, tarification et provisionnement. Economica.

[1994] L. BREIMAN ; , Bagging predictors.

[2001] L. BREIMAN ; Random forests.

[2001] R. TIBSHIRANI, G. WALTHER, T. Hastie ;Estimating the numer of cluster in a dta set via the gap statistic [1984] L. BREIMAN ; J. FRIEDMAN, R. OLSHEN, and C. STONE ; Classification and regression trees.

[1999] J. FRIEDMAN ; Stochastic gradient boosting.

[2008] T. HASTIE, R. TIBSHIRANI, and J. FRIEDMAN ; Elements of statistical learning data mining, inference, and prediction second edition.

[2016] G. LANSLEY and J. CHESHIRE, An Introduction to Spatial Data Analysis and Visualisation in R

[2002] K. P. BURHNAM and D. R. ANDERSON ; Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach, Springer-Verlag.

[2005] Y. YANG ;Can the strengths of AIC and BIC be shared ? ; Biometrika, vol. 92,p. 937–950.

[2005] S. BAILLARGEON ; Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations ; Mémoire M.Sc. ; Université Laval.

[2012] J-M.FLOCH ; Géostatistique ; Insee.

[2016] T. CHEN and C. GUESTRIN ; XGBoost : A Scalable Tree Boosting System ;University of Washington.

[2016] T. CHEN and C. GUESTRIN ; XGBoost : Reliable Large-scale Tree Boosting System ;University of Washington.

LISTE DES FIGURES

LISTE DES TABLEAUX

ANNEXES

Annexe A : Démonstrations mathématiques

$\pi = \mathbb{E}[S]$

Puisque la prime pure correspond à la tarification du risque, il ne faut pas que la valeur de la prime pure soit surévaluée ou sous-évaluée. Ainsi, l'assureur doit chercher à minimiser l'erreur quadratique $\mathbb{E}\left[(S-\pi)^2\right]$.

Par définition, $\mathbb{E}\left[(S-\pi)^2\right] = \mathbb{V}[S-\pi] + (\mathbb{E}[S-\pi])^2$.

Or, π est déterministe donc $\mathbb{V}[S-\pi] = \mathbb{V}[S]$.

Par conséquent, minimiser $\mathbb{E}\left[(S-\pi)^2\right]$ revient à minimiser $(\mathbb{E}[S-\pi])^2 > 0$, c'est-à-dire à minimiser $\mathbb{E}[S-\pi]$.

Finalement $\mathbb{E}[S-\pi] = 0 \Leftrightarrow \mathbb{E}[S] - \mathbb{E}[\pi] = 0 \Leftrightarrow \mathbb{E}[S] = \pi$ D'où $\pi = \mathbb{E}[S]$.

$\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[B]$

On l'obtient à l'aide de l'espérance conditionnelle

$$E[S] = E\left[\sum_{k=1}^{N} B_k\right] = E\left[E\left[\sum_{k=1}^{N} B_k \mid N\right]\right] = E\left[\sum_{k=1}^{N} E\left[B_k \mid N\right]\right]$$
$$= E\left[\sum_{k=1}^{N} E\left[B_k\right]\right] = E\left[\sum_{k=1}^{N} E[B]\right] = E[N]E[B]$$

$\mathbb{V}ar[S] = \mathbb{E}[N]\mathbb{V}ar[B] + \mathbb{V}ar[N](\mathbb{E}[B])^2$
Par la formule de la variance totale, on a :

$$\mathbb{V}ar[S] = \mathbb{E}\left[\mathbb{V}ar\left[S \mid N\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[S \mid N\right]\right]$$

avec d'une part $\mathbb{V}ar[S \mid N] = \mathbb{V}ar\left[\sum_{k=1}^{N} B_k \mid N\right] = \sum_{k=1}^{N} \mathbb{V}ar\left[B_k \mid N\right] = N\mathbb{V}ar[B]$
$\Rightarrow \mathbb{E}[\mathbb{V}ar[S \mid N]] = \mathbb{E}[N\mathbb{V}ar[B]] = E[N]\mathbb{V}ar[B]$
et d'autre part $\mathbb{E}[S \mid N] = \mathbb{E}\left[\sum_{k=1}^{N} B_k \mid N\right] = N\mathbb{E}[B] \Rightarrow \mathbb{V}ar[\mathbb{E}[S \mid N]] = \mathbb{V}ar[NE[B]] = \mathbb{V}ar[N](\mathbb{E}[B])^2$.

La famille exponentielle

Dans le cadre le plus général, la famille exponentielle regroupe les lois de probabilité dont la fonction de masse de probabilité ou la densité est donnée (pour une variable aléatoire X) par

$$f(x;\theta) = \exp\left(\sum_{j=1}^{k} \eta_j(\theta)T_j(x) - B(\theta)\right) h(x), x \in \mathcal{X}$$

où $\eta(.), T(.) : \mathbb{R}^k \mapsto \mathbb{R}^k, h(.) : \mathbb{R}^d \mapsto \mathbb{R}^k$ la fonction de base, $B : \mathbb{R}^k \mapsto \mathbb{R}$ et $\theta \in \Theta \subset \mathbb{R}^k$ est le vecteur de paramètres naturelles (k est le nombre de paramètre de la loi). d est la dimension de la loi. Cette formulation est appelée forme naturelle avec transformation.

Dans le cadre des GLM, on considère une classe plus simple à 1 paramètre où $k = 1$ et $T(x) = x$. Les fonctions B et h sont aussi notées différemment. Dans ce cadre, la fonction de masse de probabilité ou la densité est donnée par

$$\ln f_X(x) = \frac{\theta x - b(\theta)}{a(\phi)} + c(x, \phi), \quad x \in \mathbb{X} \subset \mathbb{R}$$

où $\mathbb{X}$ est le support de la variable aléatoire, typiquement $\mathbb{N}$ ou $\mathbb{R}_+, a, b, c$ trois fonctions différentiables propre à la loi, θ le paramètre d'intérêt et ϕ le paramètre de dispersion. On notera $X \sim \mathcal{F}_{\exp}(\theta, \phi, a, b, c)$.
On montre que les moments se calculent par

$$\mathbb{E}[X] = \mu = b'(\theta), \mathbb{V}ar[X] = a(\phi)V(\mu) = a(\phi)b''(\theta),$$

où V est la fonction de variance. $V(\mu) = b''\left((b')^{-1}(\mu)\right)$

Construction de l'arbre de décision

La première étape de la construction de l'arbre saturé T consiste à diviser la racine, c'est-à-dire l'ensemble des données, en deux sous-ensembles les plus homogènes possibles. Puis, ces deux sous-ensembles sont eux-mêmes divisés en deux classes homogènes, et ainsi de suite. L'enjeu consiste donc à déterminer la règle de découpage, c'est-à-dire la variable explicative i et les modalités $\mathcal{D}$ prises par cette variable permettant une séparation la plus homogène possible. Pour cela, il est nécessaire d'introduire les fonctions d'hétérogénéité. Ce sont des fonctions permettant de mesurer l'hétérogénéité au sein d'un noeud. L'objectif de l'algorithme CART consiste alors à réduire cette erreur lors de chaque division binaire. En notant :

- t : noeud ;
- $D(t) = \{(x_i^1; \ldots; x_i^p; y_i) \in t\}$: données appartenant au noeud t ;
- $N(t) = \text{card}(D(t))$: nombre de valeurs appartenant au noeud t ;
- t_G et t_D : noeuds droit et noeud gauche, descendants du noeud t ;
- $\bar{y}(t) = \frac{1}{N(t)} \sum_{D(t)} y_i$: moyenne empirique des valeurs de la variable à prédire, pour le nœud t,

la fonction d'hétérogénéité correspond à la variance intra-nœud. Elle est définie par :

$$r(t) = \frac{1}{N(t)} \left(\sum_{D(t)} (y_i - \bar{y}(t))^2 - \left(\sum_{D(t_G)} (y_i - \bar{y}(t_G))^2 + \sum_{D(t_D)} (y_i - \bar{y}(t_D))^2 \right) \right)$$

L'optimalité est obtenue lorsque $r(t)$ est maximale. La maximisation de la fonction d'hétérogénéité se fait par calcul de celle-ci pour l'ensemble des partitions possibles. Le critère de séparation de la variable à chaque noeud dépend de la nature qualitative ou quantitative continue/-discrète de la variable.

L'algorithme CART est un algorithme récursif : l'étape précédente permet d'obtenir deux branches (droite et gauche). La maximisation de la fonction d'hétérogénéité est ensuite effectuée sur chacune de ces deux branches, et ainsi de suite jusqu'à ce que plus aucune segmentation ne soit possible, c'est-à-dire lorsque tous les noeuds sont terminaux. Un noeud est terminal lorsque toutes les données contenues dans ce noeud sont telles que $y_i = \bar{y}(t)$ ou

lorsqu'il n'existe plus de variable explicative à tester. Chaque noeud terminal est une feuille et est caractérisé par un unique chemin partant de la racine. Ce chemin correspond à une règle. L'ensemble des règles pour toutes les feuilles constitue le modèle.

Après avoir obtenu l'arbre maximal, celui-ci est élagué car possédant une très grande variance et un biais faible (à l'inverse d'un arbre constitué uniquement de la racine qui engendre alors un prédicteur constant et donc a une très petite variance mais un biais élevé). Ainsi, on aimerait pouvoir diminuer la variance tout en conservant un biais faible. D'où l'élagage. Cette procédure consiste à créer, étant donné un noeud t non terminal de l'arbre, un sous arbre $\hat{T}$ de T : plus précisément, un arbre qui correspond à T privé des descendants de t. A la fin de la phase d'élagage, l'on dispose de plusieurs sous-arbres et donc de plusieurs estimateurs. Ainsi, la phase finale consiste à sélectionner celui qui a la plus faible erreur estimée via deux méthodes : en utilisant un échantillon test ou par validation croisée.

Annexe B : Résultats d'analyses exploratoires

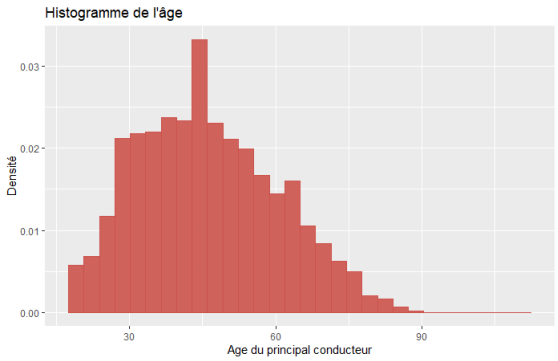


FIGURE 4.19 – Histogramme de l'âge du conducteur principal

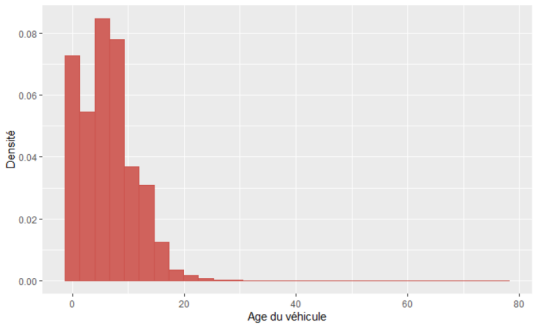


FIGURE 4.20 – Histogramme de l'âge du véhicule

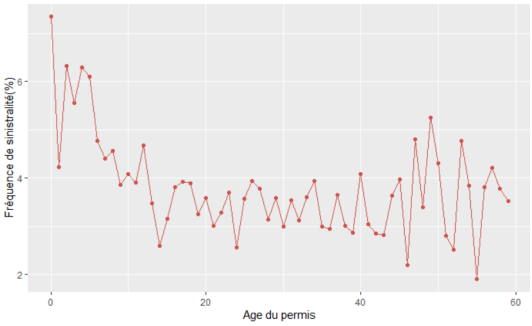


FIGURE 4.21 – Sinistralité par âge du permis

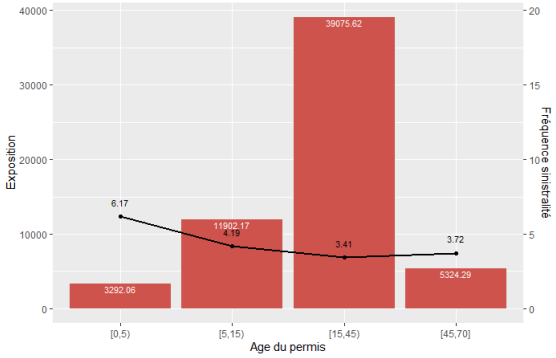


FIGURE 4.22 – Découpage en classe

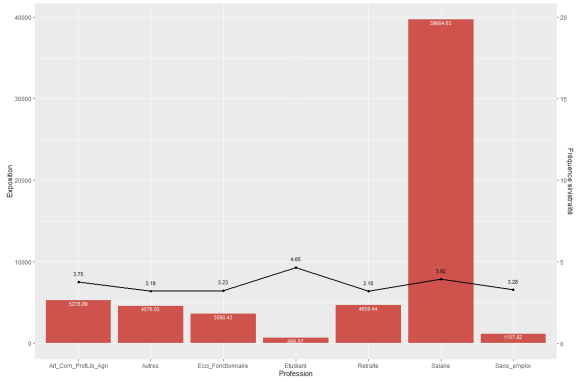


FIGURE 4.23 – Sinistralité par profession

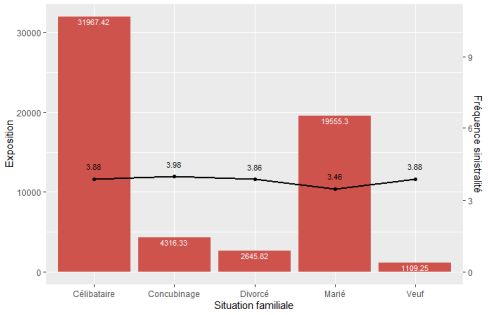


FIGURE 4.24 – Sinistralité par situation familiale

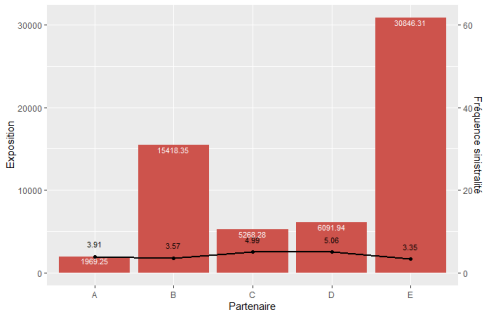


FIGURE 4.25 – Sinistralité par partenaire

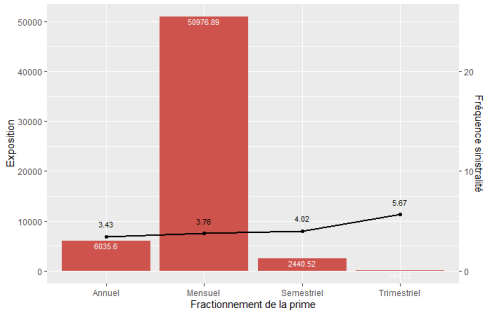


FIGURE 4.26 – Sinistralité selon le fractionnement de la prime

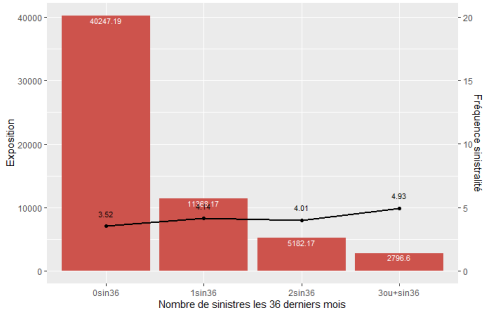


FIGURE 4.27 – Sinistralité actuelle et sinistralité passée

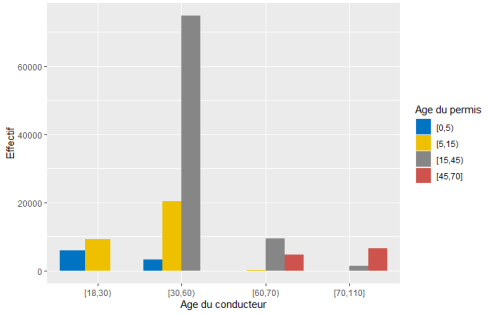


FIGURE 4.28 – Lien entre âge du conducteur et âge du permis

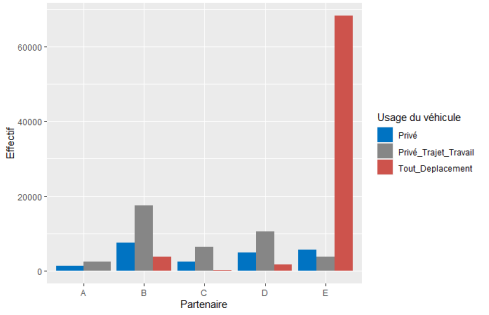


FIGURE 4.29 – Lien entre partenaire et usage du véhicule

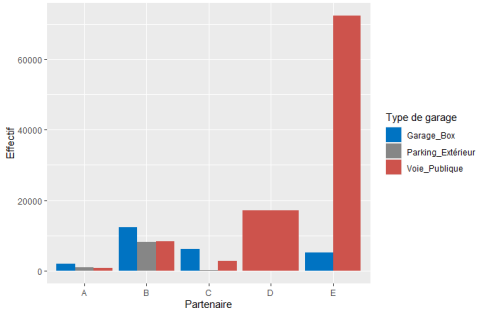


FIGURE 4.30 – lien entre partenaire et type de garage

Annexe C : Résultats modélisation

TABLE 4.5 – Adéquation à la loi de Poisson ($\lambda = 0.01636971$)

Nb. sinistres	Eff. obs	Proba obs	Proba théo	Eff. théo
0	133928	0.98400	0.98376	133895
1	2126	0.01562	0.01610	2192
2	51	0.00037	0.00013	18
Test d'adéquation :		*p-value=2.184e-14*		

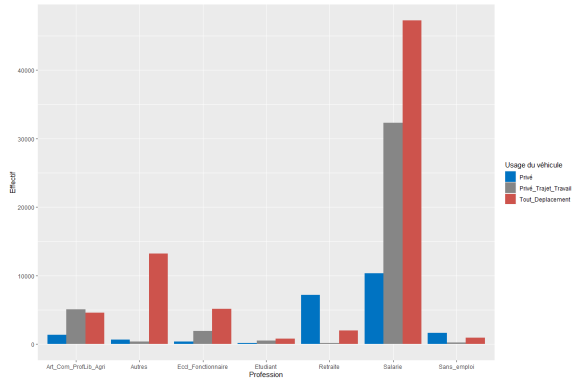


FIGURE 4.31 – Lien entre profession et usage du véhicule

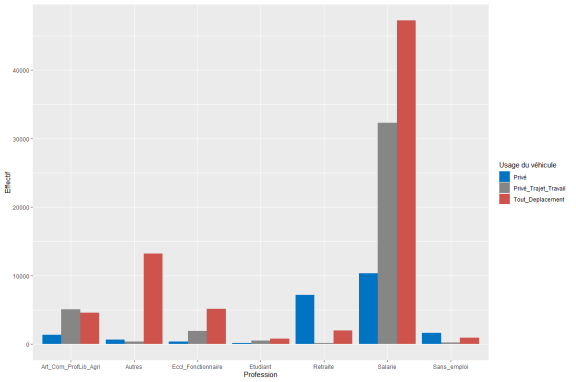


FIGURE 4.32 – Lien entre profession et usage du véhicule

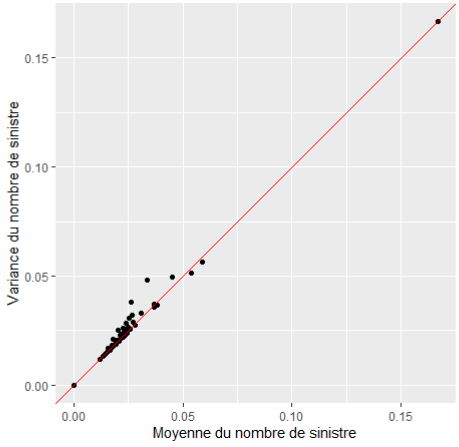


FIGURE 4.33 – Rélation moyenne-variance par âge

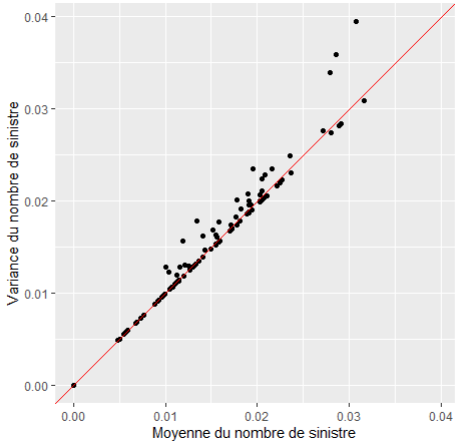


FIGURE 4.34 – Relation moyenne-variance par département

TABLE 4.6 – Adéquation à la loi de Binomiale négative ($r = 0.55644$,$q = 0.97142$)

Nb. sinistres	Eff. obs	Proba obs	Proba théo	Eff. théo
0	133928	0.98400	0.98399	133927
1	2126	0.01562	0.01565	2130
2	51	0.00037	0.00035	47
Test d'adéquation :		*p-value=0.867*		