ENSAE – Spécialisation Actuariat

Mémoire présenté devant l'ENSAE

pour l'obtention du diplôme d'Actuaire ENSAE

et l'admission à l'Institut des Actuaires

le 2 novembre 2015

Par : Antoine GUILLOT

Titre: Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?

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

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

Membre présent du jury de l'Institut des Actuaires : *Signature :*

Florence PICARD

Jérôme VIGNANCOUR

Antoine MATTEI

Alexandre YOU

Membres présents du jury de l'ENSAE :

Nicolas BARADEL

Caroline HILLAIRET

Entreprise : Sia Partners

Nom : Pierre-Antoine MERLE

Signature :

Directeur de mémoire en entreprise :

Nom : Khalid JEBBARI

Signature :

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

Signature du responsable entreprise :

Secrétariat : Pierre BERTIAUX

Bibliothèque : Cécile BAZILLOU

Signature du candidat :

ENSAE ParisTech, 3 Avenue Pierre Larousse, 92245 Malakoff cedex

Résumé

Dans un contexte d'accentuation perpétuelle de la segmentation tarifaire en assurance non-vie, la modélisation des effets non linéaires, des interactions et la sélection des facteurs de risque les plus pertinents sont des enjeux clés pour ce secteur. Afin de répondre à la problématique de la non-linéarité, ce mémoire compare le modèle économétrique classique, à savoir le modèle linéaire généralisé (GLM), avec trois méthodes issues de l'apprentissage statistique : une extension simple mais naturelle du GLM standard, un modèle additif généralisé (GAM), et une forêt aléatoire (*random forest*). Ces différentes techniques ont été comparées sous l'angle de la parcimonie afin de privilégier l'interprétabilité et l'applicabilité opérationnelle des modèles élaborés. En complément de cette analyse comparative, ce mémoire propose des solutions pour traiter deux difficultés pratiques importantes : l'utilisation d'une méthode de régularisation, le Lasso, pour résoudre les problèmes computationnels liés à la sélection de variables ; l'exploitation de la forêt aléatoire pour la prise en compte automatique des interactions.

Abstract

In a situation of constant emphasis on pricing segmentation in P&C insurance, non linear effects and interactions modelling, and features selection are key issues for this sector. In order to answer the non-linearity question, this memoir compares the classical econometric model, the generalized linear model (GLM), with three statistical learning methods: a simple but natural extension of the GLM, a generalized additive model (GAM), and a random forest. These various techniques have been compared under the light of parsimony, in order to promote interpretability and operational applicability. As a complement of this comparative analysis, this memoir offers solutions to deal with two major practical difficulties: use of a regularisation method, the Lasso, to solve computational problems related to variable selection; exploitation of random forest to automatically take into account interactions.

Note de synthèse

Les évolutions réglementaires et technologiques en faveur des assurés ont rendu le marché de l'assurance non-vie particulièrement concurrentiel. Dans ce contexte, la segmentation tarifaire constitue une objectif déterminant dans la lutte contre la sélection adverse et le risque de rachat. Tout d'abord, il est classique de distinguer la modélisation de la fréquence des sinistres par assuré, de celle de leur sévérité. Cette méthodologie apparaît cohérente, et nous l'adoptons donc pour l'ensemble de nos modèles. Ensuite, l'approche de marché consiste à recourir aux modèles linéaires généralisés (GLM) afin de prédire chacune de ces deux variables d'intérêt. Pourtant, ces méthodes traditionnelles ne parviennent pas à modéliser convenablement le phénomène de sinistralité dans toute sa complexité. En particulier, la prise en compte des effets non-linéaires, la détection des interactions, et la sélection des variables les plus pertinentes sont des sujets cruciaux, peu traités par la pratique de marché, et qui méritent une analyse approfondie. Afin de répondre à ces trois enjeux clés, ce mémoire examine divers outils issus de l'apprentissage statistique complexifiant l'économétrie traditionnelle pour mieux intégrer ces aspects.

La méthode de marché, que nous dénommons par l'expression « GLM standard », consiste à discrétiser préalablement les variables numériques de manière arbitraire, afin de produire des classes de risque homogène. Cette approche parvient certes à modéliser sommairement les phénomènes non monotones, avec des effets de convexité, d'extrêmes, ou de seuils. Toutefois, elle est peu précise et demeure particulièrement sensible à la subjectivité des nœuds de discrétisation choisis. Afin d'améliorer les performances prédictives de cette méthode, nous proposons une démarche alternative qui s'efforce de calibrer soigneusement les régresseurs du GLM à l'aide de transformations polynomiales par morceaux – les *splines*. Le principe de cette approche est d'identifier, à l'aide de la représentation graphique des coefficients de régression du GLM standard, des points de rupture notables de la relation de dépendance de la sinistralité en fonction des facteurs de risque. La figure 1 illustre cette méthode graphique par l'exemple de l'âge de l'assuré, dans le cadre du modèle de sévérité. Ces points sont ensuite spécifiés au sein du modèle en tant que nœuds de *splines*. L'extension simple qui consiste à employer des *splines* linéaires, que nous désignons par « GLM amélioré », est d'abord retenue dans un souci de parcimonie. Cette méthode conduit à la formation d'effets linéaires par morceaux (figure 2).

FIGURE 1 – Coefficients associés à l'`Âge de l'assuré` (sévérité)

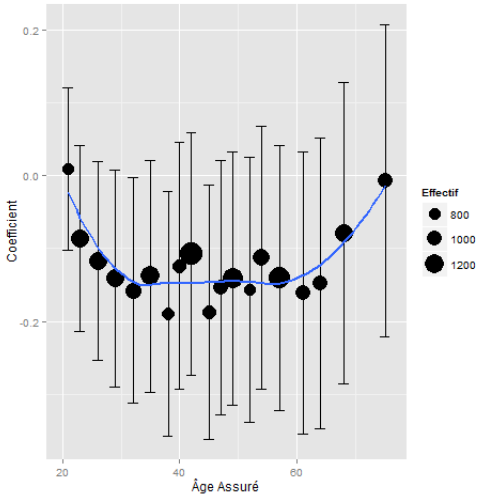
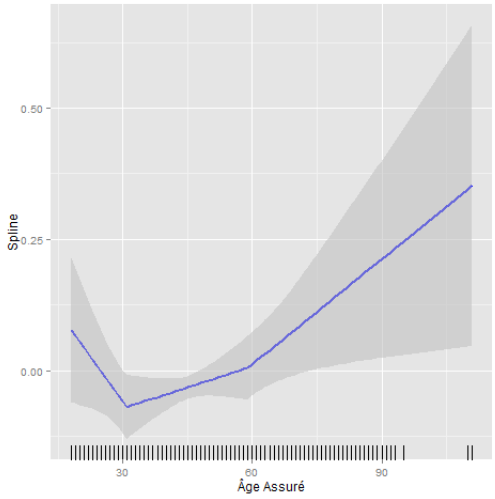


FIGURE 2 – *Spline* linéaire associée à l'`Âge de l'assuré` (sévérité)



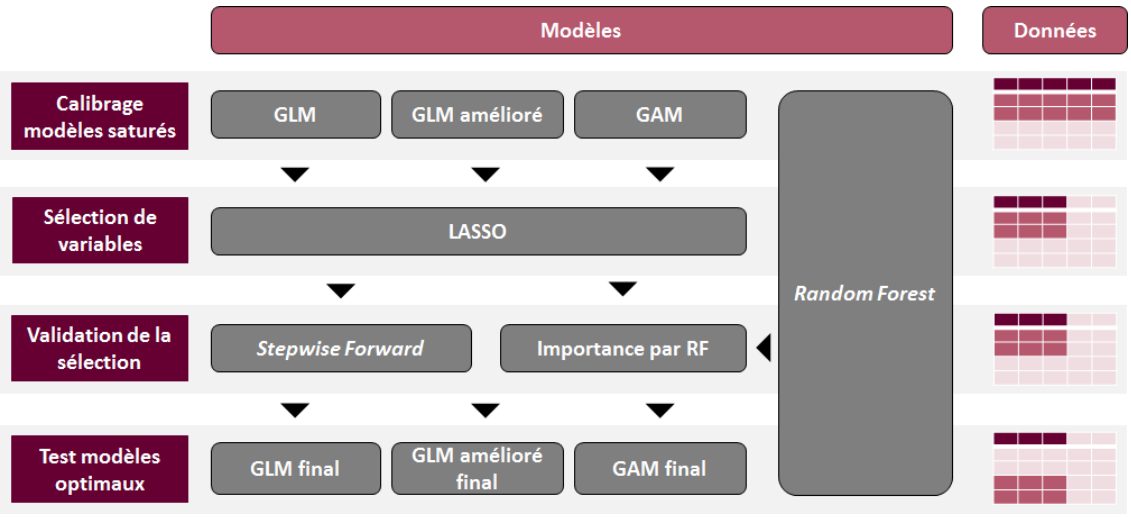
Dans un second temps, nous étudions les perspectives offertes par des *splines* cubiques : il s'agit alors d'un modèle additif généralisé (GAM). Finalement, nous comparons les résultats produits par trois modèles économétriques de complexités croissantes. Comme ces généralisations successives sont déjà qualifiables de techniques de *machine learning*, il nous parait alors intéressant d'inclure dans cette analyse comparative l'utilisation d'une méthode purement non statistique, mais qui peut se concevoir comme une extension tortueuse du GAM, la forêt aléatoire (*random forest*). Cette dernière est une technique dite ensembliste, qui agrège plusieurs arbres décisionnels, pour former un modèle non linéaire robuste.

La méthodologie de calibrage des *splines* développée dans cette étude présente déjà des avantages dans un contexte opérationnel. Elle permet à l'actuaire, à travers l'analyse de graphiques de tendances, de mieux appréhender les données disponibles, et de se forger un *a priori* vis-à-vis de la sinistralité étudiée. De plus, les nœuds de *splines* identifiés par cette approche offrent une information précieuse pour mieux définir les frontières des classes tarifaires commercialisées *in fine*.

Suite au calibrage des différents modèles économétriques, se pose la question essentielle de la sélection de variables. En effet, les difficultés statistiques et informatiques générées par un trop grand nombre de régresseurs imposent de recourir à une technique de réduction de la dimension. La procédure *stepwise* usuelle est marquée par un coût computationnel important et un biais statistique inhérent à son caractère itératif. Face à ces défauts, nous préférons une méthode de régularisation plus efficiente, le Lasso (*Least Absolute Shrinkage and Selection Operator*), qui contraint des coefficients à zéro par l'ajout d'une pénalité au sein de l'ajustement du modèle. La collection de variables résultant de cette technique est validée par sa comparaison avec les sélections issues d'un *stepwise*, et d'une forêt aléatoire. Cette dernière permet effectivement de hiérarchiser les facteurs de risque grâce à une mesure d'importance propre à sa structure. Comme les différentes sélections obtenues divergent peu, le choix du Lasso apparaît pertinent. Les modèles finaux sont enfin élaborés à partir de la collection de variables alors retenue. La figure 3 synthétise la démarche générale suivie par ce mémoire. L'ensemble des phases de calibrage sont réalisées à

partir de données d'apprentissage, représentant la moitié de la base totale. La seconde moitié est conservée pour la production de résultats indépendants.

FIGURE 3 – Démarche générale du mémoire



Les techniques de sélection de variables employées dans cette étude répondent dans une certaine mesure à des contraintes opérationnelles imposées par le marché. Le Lasso a été notamment choisi pour sont efficience en termes de temps de calcul, atout indispensable pour permettre un pilotage régulier du portefeuille. Mais il offre aussi une plus grande robustesse qui est très recherchée dans le cadre d'analyses de l'évolution de la sinistralité par les directions techniques.

Pour la comparaison des différents modèles élaborés, deux mesures de performance sont calculées sur base de test : l'erreur quadratique (tableau 1) et la rentabilité technique du portefeuille. Ces indicateurs semblent favoriser les deux modèles économétriques les plus complexes : le GLM amélioré et le GAM. Néanmoins, il apparaît ardu de formuler des conclusions générales à partir de ces indicateurs macroscopiques qui demeurent par ailleurs difficilement interprétables opérationnellement.

TABLE 1 – Erreur résiduelle quadratique (RMSE)

Variable	GLM Standard	GLM Amélioré	GAM	*Random Forest*
Sévérité	1 752,26	1 749,88	1 751,89	1 766,30
Fréquence	0,199 026	0,196 383	0,196 365	0,197 653
Charge totale	430,81	429,10	428,97	429,91

En complément de ces résultats chiffrés, des graphiques de la sinistralité prédite en fonction de l'évolution d'un facteur de risque donné permettent d'apprécier les impacts univariés modélisés (figure 4). Bien que le GLM standard parvient péniblement à capter les effets non linéaires grâce à la discrétisation de ses variables numériques, il modélise néanmoins ces impacts de manière particulièrement rigide, par une courbe de prédiction « en plateaux ». *A contrario*, le GAM présente un comportement beaucoup plus lisse, mais qui s'éloigne parfois fortement de la sinistralité réelle. Par ailleurs, à travers ces figures, l'ensemble des modèles économétriques manifestent des lacunes

en présence de variables fortement corrélées, phénomène qui n'impacte pas particulièrement la forêt aléatoire. Néanmoins, cette dernière présente un comportement souvent volatil et inégal, ce qui se distingue de la régularité naturelle des autres méthodes.

FIGURE 4 – Sévérité prédite en fonction de l'`Âge de l'Assuré`

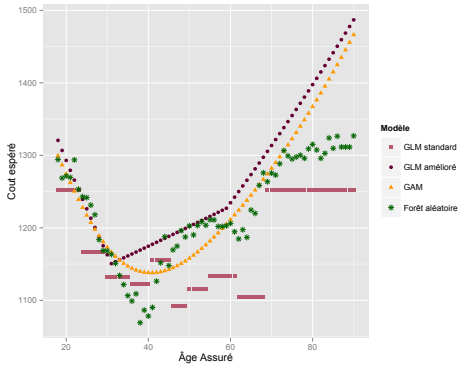
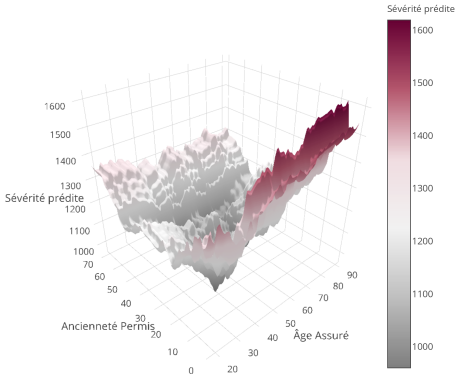


FIGURE 5 – Sévérité prédite selon l'interaction `Âge` x `Ancienneté Permis` par le *Random Forest*



Par ailleurs, en examinant des graphes tridimensionnels représentant des effets croisés entre deux variables explicatives (figure 5), les interactions inhérentes au phénomène de sinistralité apparaissent mieux modélisées par la méthode de *machine learning*, ce qui constitue un argument supplémentaire en sa faveur. Cependant, malgré l'intérêt en matière d'interprétation, cet atout n'est pas suffisant pour propulser les prédictions de la forêt en tête du classement. Finalement, au regard des divers indicateurs disponibles, sans se démarquer profondément des autres modèles, le GLM amélioré forme selon nous un bon compromis entre la souplesse des *splines* et la robustesse de la linéarité.

En définitive, les divers modèles élaborés offrent des points de vue différents sur le même phénomène de sinistralité, ce qui constitue un *benchmark* utile pour mieux identifier certaines défauts d'un modèle tarifaire existant. En particulier, la résistance de la forêt aléatoire face au problème de colinéarité permet de corriger une éventuelle dérive de modélisation d'un GLM classique. De plus, la forêt peut aussi s'exploiter en amont du processus tarifaire pour détecter d'éventuelles interactions importantes à spécifier ensuite au sein du modèle principal de tarification.

Executive summary

Regulatory and technologic evolutions in favor of insureds have made the P&C insurance market particularly competitive. In this situation, pricing segmentation form a leading objective in the fight against adverse selection and surrender risk. First of all, it is usual to distinguish the claims frequency modelling from the claims severity modelling. This methodology appears consistent, so we adopt it in all our models. Then, the market approach is to use generalized linear models (GLM) in order to predict these two variables under study. And yet, these traditional methods fail to properly model loss thoroughly. In particular, taking into account non linear effects, spotting interactions, and selecting the most relevant variables are critical topics, that the market practice rarely deals with and that deserve an in-depth analysis. In order to answer these key issues, this memoir investigates various statistical learning tools which complexify traditional econometry to better integrate these aspects.

The market method, that we named "standard GLM", is discretising beforehand numerical variables in an arbitrary fashion, in order to produce homogeneous risk classes. This approach certainly manages to barely model non monotonous phenomena, with convexity, extremes and thresholds effects. But it is little precise and remain particularly sensitive to the chosen discretisation knots' subjectivity. In order to improve this method's predictive performance, we offer another technique which strives to carefully calibrate the GLM regressors using piecewise polynomial transformations – splines. This approach principle is to identify distinct breaking points within the loss dependent function, thanks to the regression coefficients' figures from the standard GLM. Figure 1 illustrates this graphical method with the insured's age example, for the severity model. These points are then specified as spline knots, within the model. Firstly, in the interest of parcimony, we choose to use linear splines, which form a simple extension we name "improved GLM". This method leads to piecewise non linear effects (figure 7).

Figure 6 – Coefficients related to `Insureds' Age` (severity)

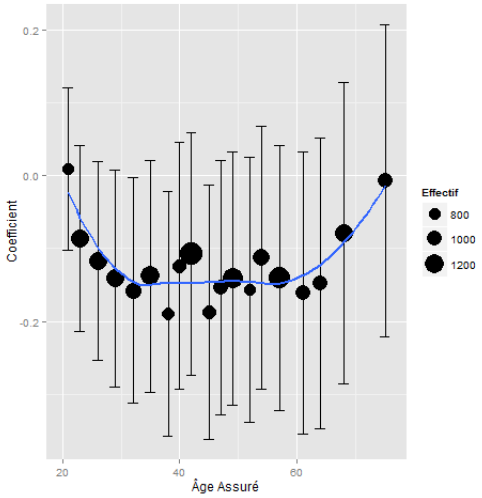
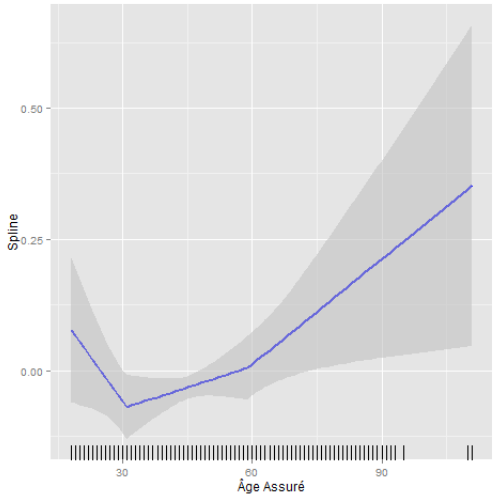


Figure 7 – Linear *spline* related to `Insureds' Age` (severity)



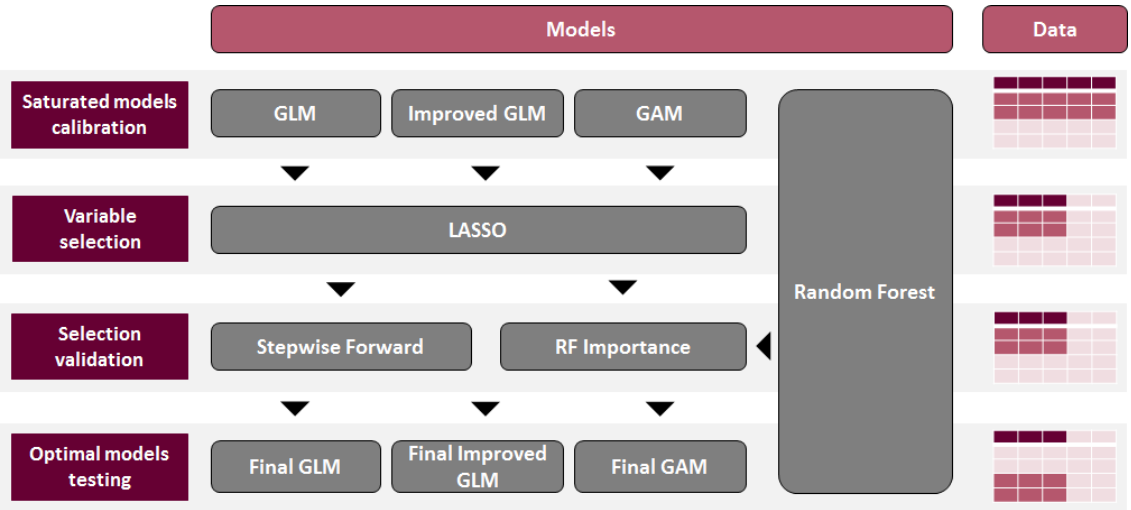
Secondly, we study prospects offered by cubic splines: it is then a generalized additive model (GAM). Finally, we compare three econometric models' results with increasing complexity. Since these nested generalizations are already part of machine learning techniques, it seems interesting

to include a purely non statistical method into our comparative analysis. We pick the random forest, which can be conceived as a tortuous extension of GAM. It is an ensemble technique which aggregates multiple decision trees to form a robust non linear model.

The *splines*' calibration methodology which is developed by this study already presents operational advantages. It allows the actuary, through the analysis of trends graphs, to better apprehend available data, and to build his own *a priori* towards the loss he studies. Besides, the *splines* knots identified by this approach offer a precious piece of information for better defining the frontiers of the pricing classes *in fine*.

After the econometric models calibration, the essential question of feature selection is raised. Indeed, statistical and computing difficulties generated by too many regressors make the use of a dimension reduction technique necessary. The usual stepwise procedure is labeled as a very expensive method and it has a statistical biais which is inherent in its iterative nature. To face these flaws, we prefer a more efficient regularization method, the Lasso (Least Absolute Shrinkage and Selection Operator), that shrinks coefficients to zero with the addition of a penalty within the model adjustement. The variable collection resulting from this technique is validated by its comparison with a stepwise and a random forest output selections. The forest allows indeed to rank risk factors thanks to its own importance measure. As various obtained selections differ only slightly, the Lasso pick seems relevant. Final models are at last built from the chosen variable collection. Figure 8 summarizes the overall methodology followed by this memoir. All calibration stages are performed from training data, which stand for one half of the total database. The other half is withheld for providing independent results.

Figure 8 – The memoir's overall methodology



The variable selection techniques used in this study somehow answer operational constraints imposed by the insurance market. The Lasso has been especially chosen for its efficiency in terms of computational time, which is an essential asset to allow a frequent monitoring of the portfolio. But it also offers a greater robustness which is very valuable to actuaries who analyse the loss evolution.

Concerning the models comparison, two performance measures are computed on a test set: the portofolio square error (table 2) and technical profitability. These indicators tend to favour the

two most evolved econometric models: improved GLM and GAM. However, it seems hard to state any general conclusion from these macroscopic indicators which remain uneasy to be operationally interpreted.

Table 2 – Root Mean Square Error

Variable	Standard GLM	Improved GLM	GAM	Random Forest
Severity	1 752,26	1 749,88	1 751,89	1 766,30
Frequency	0,199 026	0,196 383	0,196 365	0,197 653
Total loss	430,81	429,10	428,97	429,91

In addition to these figures, graphs of predicted loss depending on a given risk factor evolution allow to assess univariate modeled impacts (figure 9). Although standard GLM barely manages to capture non linear effects thanks to its numerical variables discretization, it models these impacts in a particularly stiff fashion, with a step prediction curve. To the contrary, GAM offers a much smoother behavior, but sometimes it diverges strongly from the real loss function. Besides, through these graphs, all econometric models show massive gaps in the presence of strongly correlated variables, a phenomenon that does not particularly affect random forest. However, the forest holds a frequently volatile and uneven behavior, which differs from other methods' regularity.

Figure 9 – Predicted severity depending on `Insureds' Age`

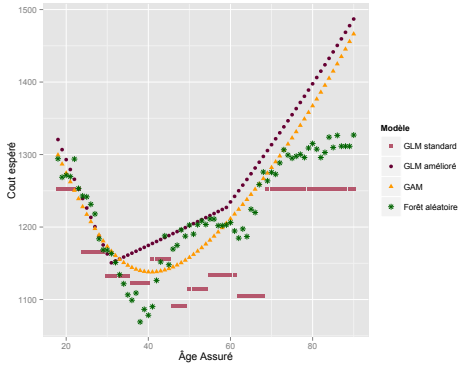
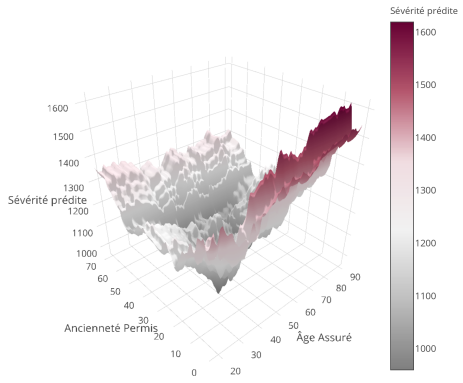


Figure 10 – Predicted severity depending on `Age` x `Licence's Seniority` interaction by Random Forest



Moreover, when looking at tridimensional graphs showing cross effects between two explanatory variables (figure 10), interactions inherent in the claims phenomenon seem better modeled by the machine learning method. which is an additional argument for it. However, despite the interesting interpretations it provides, this asset is not enough to boost the forest's predictions to the top of the ranking. Finally, in view of diverse available indicators, improved GLM forms in our opinion a nice compromise between splines' flexibility and linearity's robustness, although it does not stand out heavily.

Ultimately, the various models built here offer different points of view on the same loss phenomenon, which is a useful *benchmark* to better identify somme flaws of an existing pricing model. In particular, the random forest resistance to collinearity allows to correct any modelling drift of a classic GLM. Moreover, the forest can be also exploited before the pricing process to spot potential

major interaction that may be then specified into the main pricing model.

Remerciements

Je tiens à remercier mes tuteurs de mémoire, Khalid Jebbari et Pierre-Antoine Merle, pour leur encadrement régulier et avisé, qui m'ont guidé avec intérêt tout au long de ces travaux, et ont largement contribué à l'achèvement de ce projet.

Je souhaite remercier Michael Donio, Partner du pôle Actuariat chez Sia Partners, pour son dynamisme et sa propension à insuffler une véritable énergie chez ses consultants.

Je remercie également Hamza Ghrib, Yolan Honoré-Rouge et l'ensemble du pôle Actuariat qui forment, avec leur bonne humeur, un environnement chaleureux et agréable.

Enfin, je remercie Antoine Ly, camarade de promotion et partenaire de projet, ancien consultant et aujourd'hui doctorant, qui partage ma passion pour les enjeux digitaux et avec lequel j'ai pu échangé à tout moment sur mes questions techniques.

Non sunt multiplicanda entia sine necessitate.

« Les entités ne doivent pas être multipliées sans nécessité. »

— John Punch (1603–1661) [1]

[1] Commentaire sur *Opus Oxoniense* de John Duns Scotus, livre III, dist. 34, q. 1. dans *Opera Omnia*, vol.15, Ed. Luke Wadding, Louvain (1639), réédité à Paris : Vives, (1894) p.483a. Cette citation est souvent attribuée à tort au philosophe franciscain Guillaume d'Ockham, qui a inspiré le rasoir d'Ockham (1287–1347), désignant le principe de parcimonie (*lex parsimoniae*).

Table des matières

Introduction

La **tarification en assurance non-vie** n'a cessé de représenter un défi de taille pour les actuaires. Celle-ci repose sur la modélisation de la loi de sinistralité en fonction de divers facteurs de risque. Mais la complexification des méthodes économétriques d'une part, et l'apparition de nouvelles sources de données d'autre part, bouleversent de façon récurrente les modèles tarifaires employés. Aujourd'hui, les algorithmes d'**apprentissage statistique** (*statistical learning*), ensemble de techniques auto-apprenantes issues de l'informatique à fondements statistiques, concurrencent les modèles statistiques traditionnels pour offrir des prédictions toujours plus performantes. À ce titre, ces nouveaux outils figurent au cœur des problématiques assurantielles actuelles, et font donc l'objet de nombreux développements dans la littérature actuarielle. Toutefois, ces méthodes demeurent encore très **hermétiques** pour les néophytes, et les études les explorant offrent trop peu de moyens d'interpréter leurs résultats dans un contexte de tarification, et se focalisent davantage sur leurs aspects théoriques. L'engouement général pour ces techniques innovantes, également regroupées sous le terme générique de *data science*, s'accompagne donc de conclusions mitigées vis-à-vis de leur **applicabilité opérationnelle**. En effet, le caractère « boîte noire » des algorithmes d'apprentissage, allié avec des résultats pratiques limités, entravent la transition digitale escomptée. C'est dans ce contexte que nous menons cette étude comparative de tarification, appliquée à un portefeuille automobile. Nous souhaitons proposer, à travers ce mémoire, une approche tarifaire alternative, exploitant les atouts offerts par l'apprentissage statistique, sans perdre l'intelligibilité de l'économétrie. Et nous nous attacherons à examiner, au terme de cette investigation, les implications opérationnelles de nos différentes approches.

Comment améliorer les prédictions produites par un modèle tarifaire tout en conservant son interprétabilité ?

Les modèles économétriques traditionnels présentent diverses lacunes que les algorithmes d'apprentissage sont susceptibles de combler. En particulier, nous identifions trois enjeux capitaux en tarification automobile auxquels de tels outils peuvent répondre.

Le principal enjeu que nous cernons est la difficulté des modèles classiques à capturer les **effets non linéaires** naturellement présents au sein de la structure de dépendance de la sinistralité. Ces effets non linéaires peuvent présenter des formes variées : des paliers (plancher ou plafond) bornant les valeurs prédites, des seuils marquant une rupture du phénomène, des extrêmes indiquant un caractère non monotone, de la convexité, etc.

Pour répondre à ce problème, nous nous proposons de comparer le modèle économétrique classique, à savoir le modèle linéaire généralisé (GLM), avec trois méthodes issues de la vaste famille que représente l'apprentissage statistique :

- un modèle linéaire calibré avec finesse, comportant des extensions modestes, et qui constitue une généralisation naturelle du modèle standard ;
- un modèle dit « additif », qui s'inscrit dans la continuité de généralisation du modèle précédent ;
- un modèle purement non statistique[2], extrêmement populaire, et qui représente pour nombre d'experts la vitrine de l'apprentissage : la forêt aléatoire (*random forest*).

[2] Il s'agit alors plus d'apprentissage automatique (*machine learning*).

Afin de comparer les performances opérationnelles de ces différentes approches, nous examinerons les divergences en termes de **rentabilité technique** du portefeuille automobile. Nous nous attacherons également à déterminer dans quelle mesure les résultats issus de ces modèles sont interprétables par des non-spécialistes.

À ce premier enjeu, s'ajoutent selon nous deux sujets secondaires, mais qui nous paraissent toutefois importants. En premier lieu, au cours du calibrage d'un modèle économétrique, une **sélection de variables** est d'abord réalisée. Cependant, la procédure communément employée est particulièrement coûteuse en temps de calcul et peut également présenter des défauts de sélection. Pour répondre à cette difficulté, nous nous proposons donc de recourir à une technique d'apprentissage statistique, le Lasso (*Least Absolute Shrinkage and Selection Operator*, Tibshirani, 1994 [33]), présentant une complexité informatique moindre et permettant d'aboutir également à un écrémage des variables initiales. Cette approche consiste à introduire une pénalité au sein de l'estimation du modèle, afin de contrôler l'amplitude des coefficients estimés, et même de contraindre certains paramètres à zéro. Par ailleurs, la forêt aléatoire (*Random Forest*, Breiman, 2001 [7]), utilisée généralement en tant que modèle de prédiction, peut aussi être employée à des fins de sélection préliminaire. Celle-ci agrège des arbres décisionnels afin de produire un modèle non-linéaire robuste. Aussi, nous comparerons, au cours de notre étude générale, les dissimilitudes produites par ces trois méthodes.

En second lieu, un défaut supplémentaire des modèles traditionnels est la non prise en compte automatique des **interactions**. En effet, la contrainte additive imposée à l'économétrie traditionnelle interdit toute introduction d'effets croisés entre deux régresseurs ou plus, à moins de les spécifier manuellement. Parmi les trois modèles supplémentaires considérés, seule la forêt aléatoire permet de prendre en compte *de facto* les interactions de manière automatique. Nous souhaitons donc profiter de cette étude pour examiner si des interactions significatives apparaissent au cours de la modélisation par forêt aléatoire.

La première partie de ce mémoire expose en détails les problématiques principales de la tarification non-vie, et en particulier automobile, qui émergent de la vaste littérature sur le sujet, et précise comment les outils alors énumérés permettent de répondre aux divers enjeux cernés. La seconde partie développe la théorie mathématique sous-tendant les différents modèles employés par la suite, en s'attachant à mettre en évidence le cheminement théorique qui permet de généraliser naturellement les méthodes traditionnelles en des algorithmes d'apprentissage performants. Puis la troisième partie développe les divers outils de sélection de variables proposés. La quatrième partie introduit les données disponibles, les retraitements statistiques nécessaires à leur nettoyage et à leur préparation préalable aux modélisations ultérieures, ainsi que de premières analyses descriptives traduisant la structure interne des données. La cinquième partie met ensuite en œuvre les méthodes économétriques sélectionnées et examine les enjeux de paramétrage y afférant. La sixième partie retrace la démarche de calibrage associée à l'élagage des modèle saturés précédents en des modèles optimaux et étudie la nature des variables sélectionnées par les approches retenues. Enfin, la dernière partie analyse les résultats produits par les différents modèles élaborés et compare leurs performances respectives.

1 État de l'art

La revue de littérature menée dans cette partie permet de préciser la problématique du mémoire et d'introduire les outils d'apprentissage statistique qui seront employés par la suite pour accompagner le processus tarifaire : l'usage de *splines* pour modéliser les effets non linéaires, l'emploi d'une régression pénalisée en tant qu'alternative de sélection de variable, et le recours à la forêt aléatoire pour mieux détecter les interactions.

Les dispositifs de tarification au sein des compagnies d'assurance non-vie exploitent habituellement des modèles économétriques, qui visent à déterminer une relation de dépendance entre le phénomène de sinistralité et divers facteurs de risque. Ces derniers sont liés aux caractéristiques de l'assuré, du contrat et du véhicule. La méthodologie de tarification dominante sur le marché suit le principe de la **prime pure** chargée. Cette approche consiste à déterminer l'espérance de la distribution des sinistres, qui est ensuite majorée par un chargement de sécurité.

Mais l'estimation précise de ce premier moment constitue déjà une opération complexe. Dugas (2003 [12]) énumère par exemple les nombreuses difficultés afférentes à cette tâche : multiplicité des facteurs de risque, distribution à queue épaisse, relation de dépendance non linéaire, influence temporelle, puissance de calcul requise, données manquantes, enjeux commerciaux, etc. Dans le cadre de l'étude du processus de tarification, nous nous proposons d'examiner dans un premier temps les différents écueils soulevés par la littérature actuarielle. Nous expliciterons ensuite les enjeux qui nous paraissent clés et que nous souhaiterions traiter dans le cadre de cette étude. L'énoncé de ces différentes problématiques s'accompagnera également des solutions envisagées par ce mémoire.

1.1 La tarification non vie dans la littérature actuarielle

Le processus de modélisation de la prime pure se décompose en deux grands chantiers :

- la préparation des données, qui garantit que le format et la qualité de celles-ci soient en adéquation avec les exigences des travaux de modélisation ultérieurs ;
- l'élaboration des modèles, qui requiert également un exercice de calibrage spécifique aux méthodes employées.

Au cours de ce processus, de nombreux obstacles mettent en défaut les hypothèses de modélisation retenues et nuisent ainsi à la précision des résultats produits. À partir de la liste des enjeux formulée par Dugas, et d'autres sources bibliographiques, examinons la nature et les caractéristiques de ces diverses difficultés.

1.1.1 Préparation des données

La prise en main des données préalable à toute modélisation statistique présente généralement les plus grands maux auxquels sont confrontés les actuaires. Cette phase comprend le nettoyage et les éventuels retraitements des données afin qu'elles puissent être facilement exploitables par les modèles statistiques.

Il faut noter dans un premier temps que la performance des modèles de tarification est en grande partie prédéterminée par la qualité des données sur lesquelles ils reposent. En particulier,

le traitement des **valeurs manquantes** demeure une véritable problématique et suscite un débat intact au sein de la profession statistique. Les méthodologies d'imputation des données manquantes sont largement abordées dans la littérature statistique générale et celles-ci ne dépendent pas des spécificités de la tarification en assurance. Les implémentations proposées dans un chapitre de Gelman & Hill (2006 [16]) sont de bons exemples d'approches adaptées à cet enjeu. Toutefois, il ne parait pas évident de choisir parmi ces différentes méthodologies sans réaliser une véritable étude comparative. En conséquence, ce mémoire se contentera d'omettre les valeurs manquantes.

Outre ces considérations générales, les données assurantielles présentent des particularités propres au domaine qu'il convient de prendre en compte dans une modélisation tarifaire. En premier lieu, le phénomène de sinistralité non vie suit parfois une **distribution de probabilité à queue épaisse**. Par exemple, les sinistres de la garantie Responsabilité Civile Corporelle peuvent s'élever à plusieurs millions d'euros dans des cas extrêmes particuliers et ceux-ci impactent très fortement la forme de la distribution sous-jacente. Ces sinistres, dits « graves », nécessitent donc d'être distingués des sinistres plus standards, dits « attritionnels », dans le modèle de tarification. Il est donc souhaitable d'écrêter les sinistres extrêmes. Pour ce faire, la **théorie des valeurs extrêmes** offre des outils permettant d'identifier le seuil d'écrêtement idéal. Cette méthodologie est systématiquement employée dans les études de tarification. Le mémoire d'actuariat de Bouche (2014 [5]) ou le très bon article de Benlagha, Grun-Réhomme et Vasechko (2009 [3]) exploitent plusieurs méthodologies particulièrement efficaces dont nos travaux s'inspirent largement.

Enfin, les enjeux commerciaux liés à la construction d'un tarif d'assurance requièrent également un retraitement particulier des données de portefeuille. En effet les Directions de la Souscription imposent généralement la production d'une **segmentation tarifaire** cohérente et commercialement acceptable. Plus précisément, la pratique de marché consiste à discrétiser les variables continues afin d'identifier des classes de risque homogène, comme l'indique Ohlsson (2010 [27]). La façon dont ces classes sont formées est généralement arbitraire, mais nous nous proposons d'affiner cette segmentation dans le cadre de ce mémoire, selon une approche similaire à celle développée par Pouna Siewe (2010 [30]). Non seulement cette méthodologie permet-elle de produire une grille tarifaire exploitable par les équipes opérationnelles, mais elle présente également l'avantage de prendre en compte la structure de dépendance non linéaire de la sinistralité, comme cela sera mis en évidence dans la partie suivante. Ce paramétrage nous parait constituer un enjeu clé au regard de la qualité des modèles élaborés par la suite et il sera, à ce titre, développé largement au cours de cette étude.

1.1.2 Élaboration des modèles

Après ces premiers traitements statistiques, il convient de s'intéresser ensuite à la modélisation de la **prime pure**. La méthodologie classique consiste à recourir au **modèle collectif**, qui considère que la charge de sinistralité totale est une somme aléatoire de variables aléatoires indépendantes et identiquement distribuées représentant chacune le coût d'un sinistre. Autrement dit, ce modèle suppose que la prime pure suit une **distribution composée fréquence-sévérité**. En pratique, cela se traduit par la distinction de la modélisation de la fréquence des sinistres par assuré, de celle du coût unitaire de chacun de ces sinistres. Chacune de ces deux tâches s'inscrit dans le cadre des problèmes de **régression** pour lesquels la variable d'intérêt – le nombre ou la charge de sinistres – est quantitative, par opposition aux problèmes de classification, pour lesquels la variable réponse

est qualitative. Les modèles traditionnels répondant à ce type de question sont historiquement les **modèles linéaires généralisés** (GLM) (Ohlsson, [27]), qui étendent le cadre très limité de la régression linéaire multiple. Ces modèles s'appuient sur une hypothèse de loi de probabilité pour estimer l'espérance de la sinistralité. Ce cadre paramétrique est particulièrement populaire au sein de la profession actuarielle, puisqu'il offre des résultats facilement interprétables et qu'il permet de quantifier l'impact de chaque variable explicative sur la variable réponse. L'existence de distributions particulièrement adaptées à la nature des données modélisées encourage l'usage de ces modèles, et justifie également la distinction fréquence/coût : des données de comptage pour la fréquence, des données continues pour le coût.

Le principal avantage des GLM par rapport aux régressions linéaires classiques est la prise en compte partielle d'effets non linéaires à travers la **fonction de lien**, qui transforme la structure de dépendance initialement linéaire entre la variable réponse et les régresseurs. Celle-ci correspond généralement à la fonction logarithmique en tarification non vie, comme nous le verrons dans la partie de modélisation. Cette fonction est toutefois régie par des contraintes de régularité fortes qui restreignent le champ des effets modélisables. Elle présente notamment la spécificité d'être strictement monotone, ce qui implique que l'effet modélisé d'un prédicteur donné sur la variable de sortie est systématiquement de signe constant. Cette particularité ne permet donc pas de modéliser fidèlement, par exemple, les phénomènes suivants :

- des impacts de signe contraire aux deux extrémités du domaine de définition d'un certain prédicteur (effets d'extrêmes)
- des ruptures de la relation de dépendance en des points de discontinuité précis (effets de seuils)
- des bornes de la variable de sinistralité aux extrêmes (plancher et plafond)
- un comportement convexe (la fonction de lien logarithmique étant naturellement concave)

Aussi, un modèle GLM offre en réalité une modélisation **pseudo-linéaire**, ce qui limite fortement le pouvoir explicatif du modèle.

Afin de s'affranchir de ce carcan pseudo-linéaire contraint par les GLM, de nombreuses alternatives ont été explorées. Pouna Siewe (2010 [30]) et Huther (2014 [21]) développent l'utilisation de modèles additifs généralisés (GAM). Pour aller plus loin, Paglia (2011 [29]) compare les GLM avec plusieurs modèles basés sur des arbres de décision (CART, *Bagging*, *Random Forest*, *Gradient Boosting*). Dans ce sens, Dugas (2003 [12]), et Christmann (2004 [9]) expérimentent diverses méthodes d'apprentissage statistique comme des réseaux de neurones (NN) ou des machines à vecteurs de support (SVM). Ces dernières techniques présentent néanmoins un certain hermétisme pour les néophytes et elles demeurent, à ce titre, difficilement interprétables, ce qui limite fortement leur applicabilité opérationnelle. Dans ce mémoire, en vue de considérer les **implications opérationnelles** de nos résultats, nous nous interdirons de traiter les modèles les plus complexes, à savoir réseaux neuronaux et SVM. Néanmoins, nous expérimenterons diverses alternatives à l'économétrie à des fins de comparaison. Notons par ailleurs que ces différents auteurs comparent davantage la performance prédictive de ces méthodes que leur pouvoir explicatif. Autrement dit, ils se focalisent essentiellement sur les résultats quantitatifs des modèles élaborés, mais omettent souvent d'étudier leur disposition à capturer des tendances particulières, des motifs récurrents au sein des données. Dans ce sens, ce mémoire tentera d'apporter, outre une comparaison des erreurs de prédiction, des conclusions sur la capacité d'un modèle à décrire la vraie nature sous-jacente

des données. En particulier, nous nous attacherons à analyser la façon dont un modèle, même simple, peut être ajusté finement afin d'estimer les effets de non-monotonie, d'extrêmes et de seuils mentionnés plus haut.

1.1.3 Difficultés complémentaires

D'autres sujets mineurs méritent également d'être évoqués au regard de la tarification non vie. Premièrement, l'**exposition** de chaque observation au sein du portefeuille considéré impacte la probabilité de survenance d'un sinistre, et celle-ci doit donc être intégrée au sein de la modélisation en conséquence. Peu d'articles développent concrètement la gestion de l'exposition, mais cette notion est pourtant cruciale lors de l'analyse d'un portefeuille de polices sur plusieurs années consécutives. Contrairement aux premières intuitions, il ne s'agit pas de pondérer simplement les observations en fonction de leur exposition, ce qui introduirait un biais en faveur des contrats les plus exposés (cf partie 2.7), mais bien d'inclure cette variable au sein même de l'écriture de l'estimateur pour les GLM, ou de la fonction d'hétérogénéité pour les arbres de décision. Des détails sur ces techniques se retrouvent toutefois dans certaines sources, comme chez Paglia (2010 [28]) ou Charpentier (2013 [8]), précisions que nous exploiterons pour le calibrage de nos modèles.

Deuxièmement, la modélisation de la fréquence souffre aussi d'une autre difficulté liée aux **classes fortement déséquilibrées**. En effet, les polices ne présentant aucun sinistre sur l'ensemble de la période d'observation sont largement majoritaires au sein du portefeuille. Aussi, la qualité de la prédiction de la fréquence de sinistres par police se mesure à la capacité du modèle employé à détecter correctement une minorité d'observations, à savoir les polices sinistrées. Or les modèles traditionnels sont peu robustes en présence de classes fortement minoritaires puisque la vraisemblance est une quantité moyennée, et sa maximisation ne permet donc pas de prendre compte correctement l'information fournie par ces observations. Cet enjeu de détection de classes déséquilibrées a été identifié comme l'un des défis majeurs du *machine learning*. Et Bouche (2014 [5]) est l'un des premiers actuaires à s'y être intéressés dans le cadre de la détection des sinistres graves. Alors que cette problématique est clé dans une optique de prédiction, elle parait moins prioritaire dans une logique d'estimation des impacts des facteurs de risque sur l'espérance. De plus, cette difficulté n'est un véritable sujet que lorsque l'effectif de la classe minoritaire est de moins de 1% environ. Dans le cadre d'un portefeuille automobile, qui constituera les données d'étude de ce mémoire, il est tout à fait plausible de rencontrer cette difficulté avec les sinistres des garanties Responsabilité Civile[3]. Il convient donc de garder à l'esprit ce phénomène récurrent lors de l'analyse des résultats.

Troisièmement, la volumétrie des bases de données traitées impose de recourir à des outils toujours plus performants afin de produire des résultats en un temps raisonnable. D'une manière générale, les **problématiques computationnelles** prennent un intérêt grandissant au sein de la communauté statistique et méritent donc d'être examinées afin de s'assurer de la pertinence opérationnelle de nos conclusions. Cette contrainte doit donc être considérée avec attention lors du choix des algorithmes employés par la suite. En particulier, les procédures de sélection de variables classiques, qui parcourent souvent de manière quasi-exhaustive l'ensemble des combinai-

[3]Comme nous le verrons par la suite, la fréquence des sinistres de la garantie Responsabilité Civile Corporelle s'évalue avec les données disponibles entre 1‰ et 2‰, alors que ceux de la garantie Responsabilité Civile Matérielle représentent entre 4% et 5%.

sons possibles, sont ainsi consommatrices de temps de calcul considérables. En réponse à cette problématique, l'apprentissage statistique offre des alternatives efficaces, notamment avec les régressions pénalisées. Ces techniques alternatives seront donc envisagées au cours de l'élaboration des modèles.

Quatrièmement, les coefficients estimés par des modèles linéaires se rapportent à une seule variable : il s'agit d'évaluer l'impact de chaque facteur de risque sur la sinistralité et ce, de manière indépendante. Aussi, non seulement les modèles économétriques classiques ne peuvent modéliser les phénomènes non monotones, mais ils sont également incapables de capter automatiquement des effets croisés entre plusieurs variables, c'est-à-dire de prendre en compte le comportement de la sinistralité lorsque deux prédicteurs évoluent simultanément (Paglia, 2011 [29]). Ces **interactions** entre facteurs de risque constituent, après la non-monotonie stricte, un autre aspect de la non-linéarité qui caractérise le phénomène de sinistralité. Traiter cette nouvelle difficulté reviendra à s'affranchir du caractère additif des modèles économétriques. Cette amélioration est envisageable à travers, par exemple, l'utilisation d'arbres décisionnels, dont nous examinerons la pertinence dans le cadre de ce mémoire.

Enfin, l'étude d'un portefeuille sur un historique de plusieurs années requiert de considérer les **impacts temporels** sur l'estimation des paramètres, en particulier via l'effet inflationniste sur la sévérité des sinistres. Ce point sera pris en compte à travers l'utilisation de la variable de l'année calendaire, comme nous le verrons par la suite.

1.2 Enjeux clés

Après cette première revue de littérature, nous sommes en mesure de cerner une collection d'enjeux qui nous paraissent primordiaux au regard de la problématique définie : l'amélioration des prédictions produites par un modèle tarifaire non vie. Le principal enjeu ciblé est celui de la modélisation non linéaire, dont les implications sont jugées cruciales sur les résultats obtenus. À ce premier sujet, s'ajoutent également deux enjeux secondaires, mais de portée non négligeable sur nos travaux : les problématiques computationnelles afférentes à la sélection de variables, et la prise en compte des interactions au sein du modèle tarifaire. Nous présentons dans cette sous-partie la nature de ces trois enjeux, ainsi que les différentes solutions envisagées dans ce mémoire pour y répondre.

1.2.1 Modélisation non linéaire

Au cours de l'examen précédent des défis liés à la tarification non-vie, les implications liées au comportement non linéaire de la sinistralité se sont révélées récurrentes. Il est donc naturel de souhaiter généraliser les méthodologies classiques de tarification afin de prendre en compte ces effets non linéaires, et en particulier non monotones. Comme indiqué précédemment, la discrétisation des variables continues préalable à l'élaboration des modèles constitue déjà une première approche susceptible de modéliser ces impacts de seuils, d'extrêmes, et de convexité. Elle représente une pratique courante sur le marché de l'assurance non-vie puisqu'elle permet, de surcroît, de produire une segmentation tarifaire cohérente. Toutefois, la démarche actuelle consiste à former des classes soit d'effectifs homogènes, soit selon des intervalles réguliers. Ce choix semble arbitraire et ne bénéficie pas d'une caution statistique particulière. Dans l'optique d'améliorer la performance prédictive

des modèles bâtis, nous souhaitons donc affiner la détermination de ces points de discontinuité. Pour ce faire, nous proposons une approche de calibrage méthodique en plusieurs étapes qui exploite les résultats de modèles économétriques intermédiaires ainsi que des outils graphiques, afin d'identifier la localisation et le nombre optimaux des seuils de discrétisation. Nous désignerons le modèle issu de cette démarche sous l'expression « GLM amélioré ».

Cette approche manuelle consiste à recourir à l'utilisation d'une transformation linéaire par morceaux de la variable continue considérée. Il se trouve que cette simple extension constitue un premier pas vers une généralisation répandue des GLM que sont les modèles additifs généralisés (GAM). En effet, les transformations élémentaires impliquées dans notre GLM amélioré sont des objets mathématiques particuliers de l'ensemble plus général des *splines*. Ce terme désigne des polynômes par morceaux spéciaux, avec des applications en optimisation polynomiale, et dont nous expliciterons les tenants et aboutissants dans une partie théorique détaillée. Ils représentent également les composés principaux des modèles additifs et justifient, à ce titre, la généralisation naturelle du GLM amélioré en GAM. C'est pour ces raisons théoriques qu'il nous paraît raisonnable d'expérimenter, outre l'extension simple du GLM d'abord proposée, les outils plus complexes offerts par les GAM. Mais alors que Pouna Siewe (2010 [30]) exploite d'abord les résultats d'un GAM pour élaborer un GLM amélioré similaire à notre proposition, nous nous attacherons à respecter dans notre étude le cheminement naturel de la théorie sous-tendant ces méthodes. Nous analyserons d'abord les modèles les plus simples afin de paramétrer convenablement les discrétisations requises au GLM amélioré, puis nous exploiterons ces mêmes résultats dans le calibrage du GAM. Précisons que ces deux extensions successives s'inscrivent à la jonction entre l'économétrie traditionnelle, à fondements statistiques, et l'apprentissage automatique (*machine learning*), à fondements informatiques. Cet ensemble de méthodes hybrides, aux frontières indistinctes, peut se désigner sous le terme d'apprentissage statistique (*statistical learning*). En transgressant les hypothèses théoriques des GLM, et en altérant ainsi le cadre très restrictif de ces modèles, nous parcourons effectivement déjà le domaine empirique de l'apprentissage.

Sans plonger entièrement dans le monde du *machine learning*, les diverses méthodes exposées précédemment effleurent déjà le caractère innovant de ces techniques heuristiques issues de l'informatique. Toutefois, afin de parfaire notre étude comparative, il convient d'examiner un modèle purement non statistique, qui complexifie encore davantage les modèles à fondements statistiques. Pour ce faire, recensons – de manière très schématique, et non exhaustive – les trois grandes familles de méthodes[4] de *machine learning* les plus populaires, que le lecteur pourra découvrir à travers l'excellent ouvrage de Hastie, Tibshirani et Friedman (2009 [20]), celui de Bishop (2007 [4]), ou encore de Izenman (2008 [23]) :

- Les arbres de décision
- Les réseaux de neurones artificiels (NN)
- Les machines à vecteurs de support (SVM)

Dans notre choix d'une méthode d'apprentissage parmi ces grandes familles, insistons encore une fois sur notre préférence pour les **modèles parcimonieux**, qui privilégient des fonctions de prédiction simples. D'une part, l'applicabilité opérationnelle d'un modèle tarifaire est largement conditionnée par sa propension à être **facilement interprétable**, alors que les modèles les plus

[4]Rappelons que nous sommes exclusivement confrontés, dans le cadre de cette étude de tarification non vie, à des problèmes de régression supervisée, et non de classification ou de *clustering*. C'est la raison pour laquelle nous n'énumérons ici que les méthodes susceptibles d'expliquer ou de prédire une variable quantitative.

complexes se démarquent généralement par leur caractère « boîte noire ». Il est donc recommandable de délaisser les deux dernières familles : réseaux neuronaux et SVM. D'autre part, la performance prédictive d'un modèle tarifaire dépend de deux quantités d'évolutions contraires : le **biais**, qui mesure l'écart moyen aux données réelles ; la **variance**, qui mesure la volatilité moyenne par rapport à des données de test indépendantes. Mais alors que le biais décroît généralement avec la complexification du modèle, sa variance augmente systématiquement : c'est le phénomène de **sur-apprentissage**. Aussi, afin de limiter l'erreur de prédiction, somme de ces deux quantités, nous confirmons notre premier choix pour des arbres décisionnels. Toutefois, bien que ces techniques présentent une complexité moindre que les deux autres familles, elles sont également réputées pour conserver une volatilité assez importante. De sorte qu'il est courant de recourir à des méthodes dites ensemblistes qui agrègent plusieurs arbres décorrélés pour former des modèles plus robustes en présence de données de test. La méthode ensembliste la plus célèbre au sein de la communauté *data science* est sans doute la forêt aléatoire (*random forest*). Celle-ci est d'ailleurs prônée empiriquement comme l'un des modèles prédictifs les plus performants et ce, de manière récurrente. Alors qu'il est classique de commencer par bâtir un arbre décisionnel simplement élagué, puis de comparer celui-ci avec différentes méthodes ensemblistes (*bagging*, *random forest*, *boosting*), nous nous contenterons ici de comparer nos divers modèles économétriques (GLM, GLM amélioré, GAM), qui forment le cœur de ce mémoire, avec ce qui semble représenter la vitrine du *machine learning*, une technique de référence pour les puristes, la forêt aléatoire.

1.2.2 Sélection de variables

Après ce premier sujet qui sera longtemps discuté au cours de ce mémoire, se pose une seconde difficulté, certes mineure au regard du choix des modèles employés, mais avec des conséquences substantielles sur les prédictions obtenues. Suite à l'élaboration initiale des modèles économétriques, ceux-ci peuvent présenter une grande variabilité des prédictions, conséquence d'un nombre élevé de paramètres – en l'occurrence, de variables expliquées. En effet, les modèles linéaires sont particulièrement sensibles au risque de sur-apprentissage en grande dimension[5]. Pour limiter ce risque, il convient donc de procéder à une sélection de variables, afin de réduire le nombre de paramètres du modèle ainsi construit. De plus, la préférence pour des modèles parcimonieux contribue à produire des segmentations tarifaires simples et opérationnellement applicables.

Les procédures classiques de sélection de variable, les *stepwise*, basées sur le parcours quasi-exhaustif de l'ensemble des combinaisons possibles, sont couramment utilisées, mais celles-ci présentent plusieurs défauts qui les rendent moins pertinentes, en particulier dans le cas de données volumineuses. Premièrement, le coût informatique élevée qui caractérisent ces techniques compromet sérieusement leur applicabilité opérationnelle. Deuxièmement, la grande volatilité de ces méthodes vis-à-vis de la base d'apprentissage employée dégrade leur performance prédictive. Pour pallier ces deux défauts, il est possible de faire appel à des alternatives issues de l'apprentissage statistique, moins gourmandes en temps de calcul, et plus robustes en présence de données indépendantes. Parmi celles-ci, les régressions pénalisées, aussi appelées méthodes de régularisation, ou encore *shrinkage estimators*, permettent de réduire la valeur des coefficients estimés à l'aide

[5] Ce trait peut s'intuiter en visualisation une régression linéaire sur n points en dimension n : la droite obtenue passe exactement par tous ces points (une droite en dimension n est entièrement caractérisée par n paramètres), ce qui correspond à un apprentissage parfait, mais une très faible propension à se généraliser à des données indépendantes.

d'un terme de pénalité ajouté à la fonction objectif du modèle. Ces techniques constituent des outils particulièrement utiles à des fins de réduction de la variance du modèle. Mais pour l'une d'entre elles, le Lasso, développé par Tibshirani (1994 [33]), cela permet également de contraindre certains coefficients à zéro, sélectionnant par la même occasion les variables les plus pertinentes, et réduisant alors la dimension du modèle.

Notons toutefois que cette alternative n'a aucune raison de résulter en la même sélection que les *stepwise*, et qu'elle n'a pas, à notre connaissance, de caution statistique tangible. En outre, elle produit des **estimations biaisées**, qui ne sont pas directement comparables avec ceux des GLM classiques. Une grande prudence doit donc être prise lors de l'interprétation des résultats des différents modèles. Il demeure cependant important d'examiner les différences de sélection découlant du *stepwise* d'une part, et du Lasso d'autre part, afin de valider les sélections résultantes. C'est pourquoi nous accordons dans ce mémoire une analyse spécifique, dédiée à la comparaison des techniques de sélection de variable.

Dans une autre mesure, la forêt aléatoire, qui comporte déjà un ajustement visant à contrer le risque de sur-apprentissage, ne souffre pas particulièrement de la multiplicité des facteurs de risque. La présélection des variables explicatives n'est donc pas réellement nécessaire. Vu sous un autre angle, ce modèle peut même être considéré comme une technique de sélection de variables en soi. En effet, à l'issue de l'élaboration de la forêt, il est possible de calculer une mesure d'importance des variables, basée sur leur contribution respective à la construction des arbres, et qui les ordonnent ainsi selon une hiérarchie chiffrée. En définitive, comme pour l'enjeu principal de ce mémoire, l'extension des modèles classiques, guidé par une logique comparative, nous souhaitons comparer les implications des différentes techniques en matière de sélection de variables, pour l'économétrie. À la procédure classique, le *stepwise*, s'oppose l'outil d'apprentissage, le Lasso. Et comme notre étude sollicite l'utilisation d'une forêt aléatoire, il est opportun d'en profiter pour examiner le classement des variables produit par le critère d'importance décrit ci-dessus. Finalement, trois outils alimenteront notre analyse.

1.2.3 Interactions

Enfin, un dernier enjeu important émerge de la revue de littérature. Celui-ci a trait à la prise en compte des **interactions** au sein des modèles employés. En effet, les modèles économétriques ne permettent pas, par défaut, l'intégration d'effets croisés, à moins de les spécifier manuellement dans la formule de régression. Quoi qu'il en soit, il n'existe pas de méthode automatique efficace aboutissant à l'estimation de coefficients croisés pertinents. Les procédures du type *stepwise* peuvent certes être adaptées, mais cela impliquerait une multiplication exponentielle du nombre d'itérations, ce qui est absolument inapproprié. La gestion des interactions nécessite donc le recours à de nouvelles extensions des modèles classiques. Notons que les phénomènes d'interactions constituent, avec la non-monotonie, la convexité et les bornes, une autre manifestation du caractère non-linéaire d'un modèle. Il est vrai que la non-monotonie, qui est propre à une seule variable, représente sans doute la source la plus courante des défauts de modélisation des GLM. C'est pourquoi ce premier aspect, désigné avec abus sous le terme « non-linéarité », est traité avec davantage d'attention dans le cœur de ce mémoire. Toutefois, les interactions méritent aussi, de notre point de vue, une analyse dédiée. Parmi l'ensemble des modèles testés, seule la forêt aléatoire permet de modéliser les interactions. Cette distinction vis-à-vis des modèles économétriques provient d'une

caractéristique des arbres décisionnels la composant. Aussi, nous développerons en profondeur, au sein de la partie théorique, les étapes successives de généralisation conduisant à cette extension. Ce cheminement théorique prolongera d'abord les modèles GAM en un algorithme de sélection d'interactions, le MARS, puis en un algorithme de construction d'un arbre de décision binaire, le CART. Ensuite, nous étudierons les interactions les plus pertinentes exposées par la forêt aléatoire dans une partie pratique spécifique (partie 7.2).

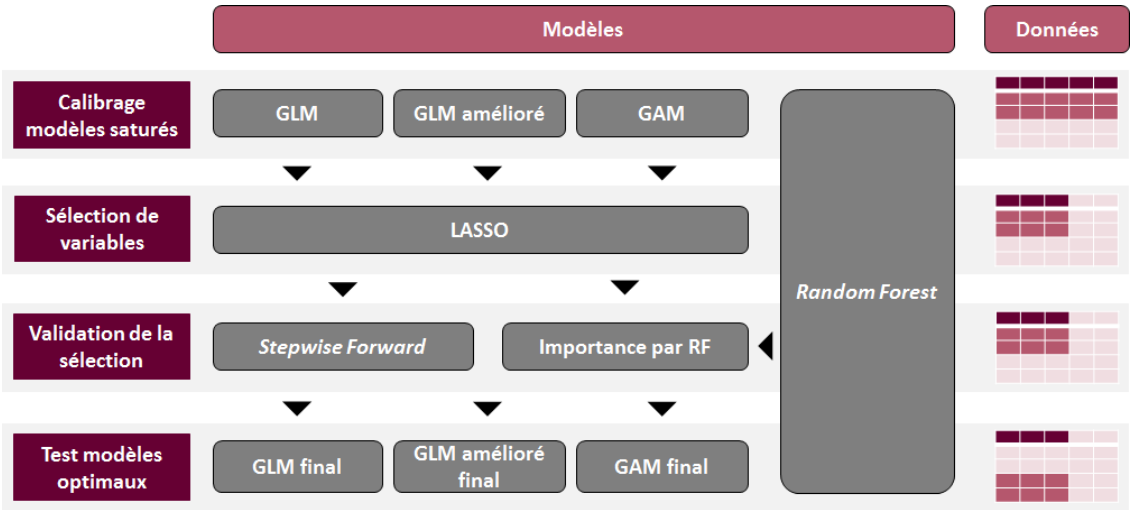
1.3 Démarche

Afin de traiter convenablement les trois enjeux développés précédemment, nous proposons de suivre une démarche organisée selon quatre étapes clés :

1. Le calibrage des régresseurs menant à l'élaboration des modèles économétriques saturés, c'est-à-dire sans sélection de variable, d'une part, et la construction de la forêt aléatoire, d'autre part.
2. La sélection des variables à l'aide de l'opérateur Lasso, appliqué en sur-couche des modèles économétriques, et menant ainsi aux modèles optimaux définitifs.
3. La validation de cette sélection à l'aide d'outils complémentaires : la procédure *stepwise forward*, et le critère d'importance issu de la forêt aléatoire de l'étape 1.
4. L'évaluation des performances de l'ensemble de ces modèles finaux.

Les trois premières étapes correspondent donc à des phases de **calibrage** successives, qui sont réalisées à partir d'une base dite d'apprentissage, qui représente 50% des données disponibles. La dernière étape, qui consiste à comparer de manière impartiale les différents modèles élaborés, est réalisée sur une base indépendante, dite de test, qui correspond à la seconde moitié des données initiales. Le schéma 11 ci-dessous synthétise cette démarche.

FIGURE 11 – Démarche générale du mémoire



2 Modélisation de la prime pure

Les modèles d'apprentissage statistique développés dans cette partie peuvent être exploités en pratique dans le cadre d'une révision tarifaire. Sans remettre réellement en question le modèle en place, ils sont susceptibles d'être utilisés à des fins de *benchmark*. Plus précisément, le modèle additif GAM conduit, à travers l'usage de *splines*, à détecter automatiquement des seuils délimitant des classes tarifaires homogènes. Le modèle classique produit parfois des prédictions grossières et peut donc dévier fortement de la réalité sur certains segments d'assurés. Il est alors possible d'identifier ces segments clés à l'aide du modèle additif, et de corriger en conséquence les écarts observés au sein même du modèle existant. De même, la forêt aléatoire apporte une vision complémentaire de la sinistralité, notamment avec sa capacité à modéliser automatiquement les interactions. En examinant les prédictions issues de cette méthode alternative sur les segments principaux, les actuaires tarificateurs sont ainsi en mesure d'apprécier la qualité du modèle en vigueur, et de refondre le tarif de ces catégories le cas échéant.

La dérégulation du marché de l'assurance non vie a rendu nécessaire la modélisation statistique de la sinistralité attendue. En effet, la concurrence a incité les assureurs à employer des techniques de segmentation de plus en plus fines afin d'offrir des tarifs compétitifs sur l'ensemble des segments cibles. L'objectif est bien de limiter *in fine* les phénomènes de sélection adverse qui conduisent à faire fuir les bons risques pour ne conserver que les mauvais risques au sein du portefeuille, et impacter ainsi négativement la rentabilité technique résultante.

Il s'agit donc de modéliser la prime pure d'un assuré en fonction de différentes caractéristiques relatives à :

- l'assuré (socio-démographiques)
- l'objet assuré (techniques)
- la police (contractuelles)

Ces modèles prédictifs sont couramment utilisés par les directions techniques du marché de l'assurance non-vie. Toutefois, les méthodes de modélisation les plus répandues sont très élémentaires, et ne permettent pas de capturer fidèlement les tendances non-linéaires propres au phénomène de sinistralité automobile. Avant d'exposer divers modèles d'intérêt, une première sous-partie introduit le principe économique du modèle collectif qui propose de distinguer fréquence et sévérité de la sinistralité. Ensuite, deux grands types de modèles économétriques ayant des comportements variés vis-à-vis des effets non-linéaires sont exposés : le modèle linéaire généralisé (GLM), reflétant la pratique de marché ; le modèle additif généralisé (GAM), extension naturelle de ce premier. Après ces premières propositions de modélisation statistique, des outils davantage issus de l'informatique sont envisagés : la régression multivariée par *spline* adaptative (MARS), procédure de spécification des interactions ; l'arbre de décision, méthode de classification descendante hiérarchique produisant un modèle similaire au MARS ; la forêt aléatoire, agrégeant plusieurs arbres décisionnels à des fins de robustesse.

2.1 Modèle collectif

La prime pure des risques non-vie est traditionnellement modélisée dans le cadre du modèle collectif, selon lequel la charge totale de sinistres S_i d'un assuré i a la forme suivante :

$$S_i = \sum_{k=1}^{N_i} Y_{i,k} \tag{1}$$

où N_i est le nombre aléatoire de sinistres de l'assuré i et $Y_{i,k}$ est le coût de son $k^{\text{ème}}$ sinistre, avec les hypothèses suivantes :

- les $Y_{i,k}$ sont indépendants et identiquement distribués, pour tout i et pour tout k
- $(Y_{i,k})_{k\geq 1}$ est indépendante de N_i pour tout i

Ce modèle distingue donc la modélisation de la fréquence des sinistres par assuré (modèle de comptage) et celle de la sévérité des sinistres quel que soit l'assuré (modèle de montant). Les arguments en faveur de cette distinction sont multiples. Tout d'abord, il n'y a pas de raison particulière de penser que les déterminants de la fréquence et du coût des sinistres soient les mêmes. Une modélisation spécifique à chaque quantité parait donc *a priori* plus pertinente. Ensuite, la fréquence est connue pour présenter un comportement beaucoup plus stable que le coût, et le lien entre les facteurs de risque et la fréquence est généralement davantage matériel. Il est donc recommandable de valoriser cette robustesse naturelle du modèle de fréquence, et de considérer la problématique de coût indépendamment. Dans le cadre de ce mémoire, nous emploierons systématiquement cette distinction lors de l'élaboration des différents modèles, par souci de cohérence.

2.2 Modèle Linéaire Généralisé

Les modèles tarifaires classiques s'appuient sur le cadre très codifié du modèle linéaire généralisé (GLM) qui constitue une extension de la régression linéaire multiple. Sans bousculer brutalement les hypothèses propres à cette dernière, cette généralisation apporte néanmoins davantage de flexibilité dans les structures modélisées. Après un rappel du principe général de cette famille de modèles, nous introduisons ici les lois de probabilités sollicitées par notre étude comparative.

2.2.1 Principe général

Les modèles linéaires généralisés (GLM) sont une extension de la régression linéaire ordinaire dans laquelle la variable réponse Y_i peut être vue comme une réalisation d'une distribution particulière de la famille exponentielle. En régression linéaire classique, une variation constante d'un prédicteur entraîne une variation constante de la variable expliquée, peu importe le point de l'ensemble de définition considéré. La relation de dépendance entre la sortie et les régresseurs est donc exclusivement linéaire. *A contrario*, les GLM permettent de modéliser une dépendance non linéaire dans le sens où c'est l'**image** de la variable réponse par une fonction arbitraire g (appelée fonction de **lien**) qui dépend linéairement des variables explicatives :

$$g\left(\mathbb{E}[Y_i|X_i]\right) = X_i^t\beta \tag{2}$$

où X_i^t désigne la transposée du vecteur d'observations X.

Un modèle GLM est donc composé de 3 éléments clés :

- une distribution de probabilité appartenant à la famille exponentielle
- un prédicteur linéaire $X^t\beta$
- une fonction de lien g monotone et différentiable

Tout d'abord, la distribution de probabilité est choisie en accord avec la forme des données à modéliser. Pour les données de comptage, il convient de choisir une distribution discrète : la loi de Poisson. Pour les données de montant, il est préférable d'employer une distribution continue positive, comme la loi gamma ou log-normale. Ces lois doivent faire parties de la famille exponentielle qui correspond aux densités de la forme :

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} \tag{3}$$

Il convient de remarquer que si le paramètre θ_i peut dépendre de l'individu i, le paramètre de dispersion $\phi > 0$ est identique pour tout i.

Ensuite, le prédicteur linéaire est l'héritage principal de la régression classique. Il implique deux contraintes fondamentales sur la fonction de prédiction produite par le modèle :

- elle est pseudo-linéaire, modulo la transformation de la réponse par la fonction de lien ;
- elle est additive en les facteurs de risque.

Enfin, la fonction de lien est un élément important du modèle GLM puisqu'elle spécifie la forme de la dépendance de l'espérance de la réponse en fonction de la structure linéaire. Alors qu'il est commun de recourir à un lien **logit** $g : t \mapsto \ln\left(\frac{t}{1-t}\right)$ pour les modèles binomiaux, car cette fonction prend ses valeurs dans l'intervalle $[0, 1]$, il est préférable d'employer un lien log pour les modèles de tarification. En effet, cette fonction permet d'obtenir des modèles dits multiplicatifs, ce qui présente l'avantage de prendre en compte les effets des facteurs de risque de façon proportionnelle. Plus précisément, une variation selon l'une des variables explicatives produit un impact sur la sortie en proportion des valeurs des autre variables, et non pas en valeur absolue comme c'est le cas pour les modèles additifs[6] :

$$\begin{aligned}\mathbb{E}[Y_i|X_i] &= \mathrm{e}^{X_i^t\beta} \\ \mathbb{E}[Y_i|X_i] &= \prod_{j=1}^{p} \mathrm{e}^{\beta_{ij}X_{ij}}\end{aligned} \tag{4}$$

Il convient de préciser que l'expression « modèles additifs » est souvent utilisée avec abus dès lors que les variables explicatives interagissent **additivement** au sein du modèle. Aussi, avec cette définition, un GLM est intrinsèquement additif, puisque modulo la fonction de lien, les variables n'ont pas d'interactions particulières entre elles, à moins de les spécifier manuellement. Dans ce sens, nous verrons que la famille des GLM peut être étendue aux GAM (modèles additifs généralisés), qui s'affranchissent de l'hypothèse de linéarité mais conservent néanmoins cette propriété d'additivité, malgré l'utilisation de fonctions de lien non identitaires.

[6] Les modèles additifs correspondent à la fonction de lien **identité**.

2.2.2 Modèle de comptage : Poisson

Les distributions de probabilité usuelles adaptés à la modélisation d'une variable discrète positive $Y_i \in \mathbb{N}^+$ sont la loi Poisson et la loi Binomiale Négative. Sous certaines conditions, il est possible de montrer (cf. Beard, 1985 [18]) que le processus du nombre de sinistres en fonction du temps suit un processus de Poisson. C'est ce qui motive donc le choix de la distribution de Poisson pour modéliser le nombre de sinistres pour toute période temps. La fonction de densité de la loi de Poisson s'écrit :

$$f_{Y_i}(y_i, \lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} \mathrm{e}^{-\lambda_i} \tag{5}$$

La fonction de lien canonique g est la fonction log. Pour cette distribution particulière, le modèle multiplicatif est donc canonique.

À partir de ces éléments, il est possible de déterminer l'estimateur des moments du paramètre de la distribution de Poisson par :

$$\begin{aligned} \hat{\lambda_i} &= \mathbb{E}[Y_i | X_i] \\ \hat{\lambda_i} &= \mathrm{e}^{X_i^t \beta} \end{aligned} \tag{6}$$

Les paramètres de la régression sont ensuite estimés en déterminant le maximum de la vraisemblance associée au modèle de l'ensemble des observations i :

$$\begin{aligned} \mathcal{L}(\beta, (x_i, y_i)_i) &= \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i) \\ &= \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \mathrm{e}^{-\lambda_i} \\ \mathcal{L}(\beta, (x_i, y_i)_i) &= \prod_{i=1}^{n} \frac{\left(\mathrm{e}^{x_i^t \beta}\right)^{y_i}}{y_i!} \mathrm{e}^{-\mathrm{e}^{x_i^t \beta}} \end{aligned} \tag{7}$$

Ou, plus simplement, de la log-vraisemblance :

$$\ln\left(\mathcal{L}(\beta, (x_i, y_i)_i)\right) = \sum_{i=1}^{n} y_i x_i^t \beta - \mathrm{e}^{x_i^t \beta} - \ln(y_i!) \tag{8}$$

Cela revient à annuler la différentielle de la log-vraisemblance, soit selon chaque direction β_j :

$$\left.\frac{\partial \ln(\mathcal{L}(\beta))}{\partial \beta_j}\right|_{\hat{\beta}_k} = \sum_{i=1}^{n} x_{ij} \left(y_i - \mathrm{e}^{x_i^t \hat{\beta}}\right) = 0 \tag{9}$$

où x_{ij} correspond à la $j^{\text{ème}}$ coordonnée de l'observation i.

Comme cette équation n'admet pas de solutions par formule fermée, la résolution numérique se fait par l'algorithme de Newton-Raphson, procédure itérative dont chaque mise à jour a la forme :

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} - \left(\left.\frac{\partial^2 \ln(\mathcal{L}(\beta))}{\partial \beta^2}\right|_{\hat{\beta}^{(t)}}\right)^{-1} \cdot \left(\left.\frac{\partial \ln(\mathcal{L}(\beta))}{\partial \beta}\right|_{\hat{\beta}^{(t)}}\right) \tag{10}$$

Le modèle de Poisson est-il réaliste ? Il est possible de constater en pratique une forte varia-

bilité de la fréquence pour un même profil d'individus, c'est-à-dire qu'un ensemble d'individus présentant les mêmes caractéristiques tarifaires peuvent se différencier significativement en termes de fréquence de sinistralité. Or le modèle décrit ici estime la même valeur du paramètre λ_i, donc de l'espérance de fréquence, pour l'ensemble de ces observations (équation 6). Ce phénomène d'hétérogénéité de la variable expliquée au sein d'une même classe tarifaire se désigne sous le terme de **surdispersion**[7] : la variance des observations au sein d'une classe tarifaire est plus grande que la variance d'une distribution de Poisson. Indiquons qu'il est toutefois possible d'ajuster le modèle pour contrer ce phénomène en considérant le paramètre λ_i comme étant la réalisation d'une variable aléatoire. Considérons ainsi une suite $(\Lambda_1, \dots, \Lambda_n)$ de variables aléatoires indépendantes et identiquement distribuées à valeurs dans $[0, \infty[$ et spécifions la contrainte suivante :

$$\mathbb{E}[Y_i|\Lambda_i] = \Lambda_i \tag{11}$$

Cette hypothèse induit la réécriture suivante du modèle 6 :

$$\mathbb{E}[Y_i|X_i] = \mathbb{E}\left[\mathbb{E}(Y_i|\Lambda_i)|X_i\right] = \mathbb{E}[\Lambda_i|X_i] \tag{12}$$

La distribution des Y_i est alors déterminée par celle des Λ_i, conditionnellement aux facteurs de risque X_i. Il est courant de spécifier une distribution gamma pour les $\Lambda_i|X_i$, ce qui présente l'avantage d'induire une distribution binomiale négative pour les $Y_i|X_i$ et de produire ainsi une forme analytiquement exprimable. En définitive, cette extension du modèle de Poisson classique, communément appelée **modèle de mélange** Poissonnien, est réputée pour mieux s'ajuster aux données assurantielles, mais elle requiert toutefois des développements théoriques qui dépassent le cadre de ce mémoire et complexifient largement les modèles pseudo-linéaires qui forment le cœur de notre étude. Aussi, tout en gardant à l'esprit l'impact potentiel du phénomène de surdispersion, nous ne le traiterons pas de manière formelle.

2.2.3 Modèle de montant : Gamma

Contrairement au modèle de fréquence, il n'existe pas de distribution évidente pour le modèle de coût. Toutefois, la loi Gamma est une distribution peu restrictive qui permet de modéliser des données continues positives à queue épaisse, et est à ce titre classique dans les analyses par GLM, comme l'illustrent de nombreux articles (cf. Murphy, 2000 [26] par exemple). En particulier, l'utilisation de la loi Gamma suppose qu'il existe une relation de proportionnalité entre l'erreur standard et l'espérance de la distribution des données, ce qui est plausible pour la sévérité des sinistres. La densité de distribution correspondante peut s'écrire sous la forme suivante :

$$f_{Y_i}(y_i) = \frac{\beta_i^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \mathrm{e}^{-\beta y_i} \tag{13}$$

où $\alpha > 0$ est le paramètre de forme et $\beta_i > 0$ est le paramètre d'intensité. Il convient de remarquer que seul le paramètre d'intensité dépend de l'assuré i. Cette particularité est vérifiable en écrivant la distribution Gamma sous forme exponentielle et en identifiant alors les expressions des

[7] Du terme anglais *overdispersion*, dont il est fait par exemple mention chez [27].

paramètres associés :

$$\theta_i = -\frac{\beta_i}{\alpha} \quad \text{et} \quad \phi = \frac{1}{\alpha}$$

Par un raisonnement similaire à l'estimation des paramètre du modèle Poissonnien, il est possible de déterminer les équations du maximum de vraisemblance du modèle de coût :

$$\left.\frac{\partial \ln(\mathcal{L}(\beta))}{\partial \beta_j}\right|_{\hat{\beta}_k} = \sum_{i=1}^{n} x_{ij} \left(y_i \mathrm{e}^{-x_i^t \hat{\beta}} - 1 \right) = 0 \tag{14}$$

Il convient de remarquer à ce stade que ces équations sont très proches de celles du modèle de fréquence. Cela provient de la certaine proximité entre la distribution de Poisson et la Gamma au sein de la famille exponentielle. Plus précisément, toutes deux font partie d'une même sous-famille de distributions qui présente un intérêt tout particulier en tarification non-vie : les **Tweedies**. Celles-ci englobent également l'ensemble des distributions composées Poisson–Gamma et sont donc parfois utilisées pour modéliser directement la charge totale de sinistres.

2.3 Modèle Additif Généralisé

Dans le cadre de la modélisation de la fréquence ou du coût de la sinistralité automobile, il est classique d'être confronté à l'apparition d'effets non linéaires typiques selon les facteurs de risque. L'augmentation du risque aux extrêmes est, par exemple, un phénomène caractéristique de la variable de l'âge. De manière plus générale, il est courant d'observer une dépendance non monotone du risque en certains prédicteurs, c'est-à-dire que le coefficient d'impact relatif à ces variables peut être positif ou négatif selon le point d'observation de leur domaine de définition. Il est désormais possible de distinguer les premières limites des modèles linéaires. Il est vrai que les GLM offrent, à travers la fonction de lien, un cadre de modélisation plus vaste que celui de la régression linéaire classique. Mais le comportement de la fonction prédite est largement conditionnée par la nature de cette fonction de lien. Si cette dernière est la fonction logarithmique, elle permet en effet de convertir l'additivité du modèle en impacts multiplicatifs. Mais ce cadre ne permet pas de capter les effets non monotones, car la structure linéaire sous-jacente subsiste en réalité. Aussi, nous nous contenterons d'avancer que la dépendance modélisée est tout au plus **pseudo-linéaire**.

Dans ce contexte, nous souhaitons briser le carcan linéaire pour un espace de fonctions de prédiction candidates plus vaste encore. Une généralisation naturelle des GLM est le recours aux modèles additifs généralisés (GAM). Ceux-ci s'affranchissent d'une hypothèse fondamentale des GLM – la linéarité – pour prendre en compte les effets d'extrêmes, de seuils et de convexité mentionnés précédemment. Ils conservent en revanche la seconde contrainte fondamentale – l'additivité –, afin de préserver en partie la certaine robustesse liée aux GLM. Plus formellement, les GAM les plus simples peuvent s'écrire de la manière suivante :

$$g\left(\mathbb{E}[Y_i|X_i]\right) = \sum_{j=1}^{p} f_j\left(X_{ij}\right) \tag{15}$$

où la fonction de lien g et la distribution de probabilité sous-jacente des observations $(Y_i)_i$ sont les deux caractéristiques inchangées de l'équation 2. Seule la structure linéaire est modifiée pour remplacer les coefficients β_j par des fonctions non linéaires f_j spécifiques à chaque prédicteur,

ce qui rend les GAM semi-paramétriques. Sous condition que les fonctions f_j aient une forme paramétrique spécifiée[8], comme ce sera le cas dans cette étude, la méthode de maximisation de la vraisemblance convient toujours dans cette situation.

Néanmoins, le choix des fonctions f_j, qui déterminent le caractère non linéaire du modèle, constitue une véritable problématique. Ces fonctions peuvent être sélectionnées parmi un ensemble de transformations classiques :

- les fonctions polynomiales
- les fonctions en escaliers
- les *splines*

Nous exposons par la suite ces différentes approches. Gardons toutefois à l'esprit que malgré l'amélioration que constitue ce nouveau cadre de modélisation, il est impossible de s'affranchir totalement de l'*a priori* indispensable à tout économètre lors du calibrage d'un GAM. Ces modèles, bien que plus performants dans certaines conditions, ne présentent guère un caractère d'automatisation, et demeurent à ce titre particulièrement sensibles à l'avis de l'expert qui les bâtit.

2.3.1 Régression polynomiale

Les régressions polynomiales constituent l'extension historique des régressions linéaires classiques en introduisant des transformation polynomiales des prédicteurs au sein du modèle. Typiquement, des polynômes de degré 2 ou 3 suffisent à capter le comportement caractéristique du risque automobile. Au-delà, les modèles construits s'exposent au sur-apprentissage. L'utilisation du carré de l'âge est un exemple de pratique récurrente qui illustre la pertinence de ces régressions :

$$g\left(\mathbb{E}[Y_i|X_i]\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Age}^2 \tag{16}$$

2.3.2 Fonction en escalier

Dans une autre mesure, les fonctions en escalier apportent également une amélioration notable aux modèles linéaires. En effet, en partitionnant le domaine de définition de certaines variables numériques en plusieurs intervalles, puis en attribuant à chacun de ces intervalles une modalité constante, celles-ci permettent de capter des effets non linéaires selon ces facteurs de risque. Cette idée astucieuse est, entre autres, inspirée par le théorème d'approximation uniforme suivant :

Théorème. *Toute fonction continue f sur un segment $[a, b]$ et à valeurs dans un espace métrique est réglée, c'est-à-dire limite uniforme de fonctions en escalier.*

Cette discrétisation des variables quantitatives en variables qualitatives ordonnées ne constitue pas une technique particulièrement novatrice puisqu'elle est souvent employée, comme nous le verrons par la suite (partie 5.2), dans le cadre simple des GLM et ce, sans recourir spécialement à la généralisation qu'offrent les GAM. Cette approche s'implémente en pratique par le choix de

[8] C'est-à-dire que ces fonctions sont entièrement définies par des coefficients explicites, analogues aux coefficients β_j du GLM.

K nœuds de discrétisation $s_1, \ldots, s_K$ afin de former $K+1$ indicatrices :

$$\begin{aligned} S_1(X_j) &= \mathbb{1}\{X_j < s_1\} \\ S_2(X_j) &= \mathbb{1}\{s_1 < X_j < s_2\} \\ S_3(X_j) &= \mathbb{1}\{s_2 < X_j < s_3\} \\ &\vdots \\ S_K(X_j) &= \mathbb{1}\{s_{K-1} < X_j < s_K\} \\ S_{K+1}(X_j) &= \mathbb{1}\{s_K < X_j\} \end{aligned} \tag{17}$$

qui constitueront les nouveaux régresseurs du modèle :

$$g\left(\mathbb{E}[Y_i|X_i]\right) = \beta_0 + \beta_1 \cdot S_1(X_{ij}) + \cdots + \beta_{K+1} \cdot S_{K+1}(X_{ij}) \tag{18}$$

Il en résulte de ce fait l'estimation d'un coefficient par intervalle, ce qui permet de prendre en compte des impacts locaux. Plus généralement, tout comportement non linéaire peut être approché par des fonctions en escalier, ce qui est facilement implémentable avec cette méthode. Toutefois, le choix des points de discrétisation est autrement plus délicat. Souvent, cela est effectué de manière arbitraire, en utilisant des quantiles de la distribution, ou en fixant des intervalles de longueur régulière. Les nœuds frontières peuvent aussi être calibrés par validation croisée, dont le principe est décrit en partie 3.1. Mais cette approche automatique ne nous permet plus d'appréhender la vraie nature sous-jacente des données, et elle semble de surcroît un peu excessive si elle est employée pour l'ensemble des variables explicatives. Afin de conserver le caractère interprétable des discrétisations opérées, nous proposons donc par la suite une démarche de calibrage (partie 5.2), basée sur des résultats graphiques, plus à même de produire un modèle crédible du phénomène de sinistralité.

2.3.3 Courbe *Spline*

En combinant les deux extensions précédentes, il est possible de construire une classe de fonctions particulièrement adaptée à la modélisation de la variété de phénomènes non linéaires pouvant être rencontrés : les **courbes *splines***[9]. Ces dernières ne sont en réalité que des fonctions polynomiales par morceaux, c'est-à-dire des fonctions continues par morceaux qui présentent un caractère polynomial sur chacun de leurs morceaux, et qui sont soumises à des contraintes de régularité supplémentaires. Ainsi, une *spline* de degré[10] d doit présenter une continuité $d-1$, aussi notée C_{d-1}, aux points frontières – les nœuds. Avec ce cadre, il est possible de retrouver les cas particuliers que représentent les deux exemples précédents :

- les fonctions polynomiales de degré d sont des *splines* de degré d sans nœuds ;
- les fonctions en escalier sont des *splines* de degré 0 avec des nœuds.

[9] L'anglicisme est communément employé, malgré l'existence d'une traduction en le terme de **cerce**, désignant également une latte souple en bois qui sert à tracer des courbes harmonieuses passant par un certain nombre de points mais ne pouvant être tracées à l'aide du compas.

[10] Nous emploierons toujours par la suite le degré pour caractériser une *spline*, par analogie avec les polynômes, bien qu'il soit fait plus souvent mention de la classification par **ordre** dans la littérature, ce qui correspond au nombre de degrés majoré de 1. Une *spline* de degré d est donc également d'ordre $d+1$.

Ces fonctions, originellement employées à des fins d'interpolation, se révèlent d'une grande utilité en économétrie. Les *splines* les plus classiques, que nous exploiterons essentiellement par la suite, sont les suivantes :

- les lignes polygonales, ou lignes brisées, qui correspondent à des *splines* de degré 1
- les *splines* cubiques, de degré 3

Cependant, l'utilisation d'une *spline* dans un modèle de régression ne parait pas triviale. Elle se traduit en réalité par l'introduction de plusieurs prédicteurs élémentaires additionnels, chacun correspondant à l'un des degrés de liberté de la *spline* sélectionnée. En effet, il est naturel d'estimer autant de coefficients qu'il n'y a de degrés de liberté. Il faut donc s'attendre à démultiplier le nombre de régresseurs en fonction du nombre de degrés de liberté ajoutés au modèle initial. Mais quelle est donc la forme de ces prédicteurs élémentaires ? Elle peut différer, selon la base de l'espace vectoriel des *splines* qui est choisie pour représenter de façon unique ces polynômes par morceaux. Et tout comme il existe une infinité de bases génératrices des polynômes standards, il en existe tout autant pour décrire les courbes *splines*. L'exhibition d'une telle base est cependant ardue car il est nécessaire de tenir compte des contraintes de régularité caractérisant les *splines*.

Déterminons dans un souci d'illustration les degrés de liberté d'une *spline* cubique avec K nœuds. Une *spline* de degré 3 à K nœuds est constitué de $K+1$ polynômes de degré 3, ce qui représente déjà $4(K+1)$ paramètres. À ce nombre, il faut retrancher les degrés liés aux contraintes de régularité aux nœuds. En effet, pour chacun des K nœuds, la *spline* cubique doit être C_2, soit trois contraintes par nœud[11], ce qui nous fait décompter $3K$ paramètres. Finalement, une première estimation empirique de $K+4$ degrés de liberté peut être émise.

Exhibons maintenant une base génératrice des *splines* cubiques afin de confirmer ce résultat. Commençons par considérer les polynômes simples $1, X_j, X_j^2, X_j^3$ auxquels sont ajoutées les troncatures cubiques aux nœuds s_k suivantes :

$$(X_j - s_k)_+^3 = \begin{cases} (X_j - s_k)^3 & \text{si} \quad X_j > s_k \\ 0 & \text{sinon} \end{cases} \tag{19}$$

Il est alors possible montrer que ces polynômes tronqués présentent une discontinuité aux nœuds qu'en leur dérivée troisième. Toutes ces fonctions sont donc bien des *splines* cubiques. De plus, il est aisément constatable qu'elles ne sont pas linéairement liées puisque chacune d'entre elles est nulle sur un intervalle distinct, à l'exception des polynômes simples qui ne sont jamais nuls mais qui sont évidemment indépendants entre eux et avec n'importe quelle fonction non polynomiale. En définitive, nous avons exhibé une famille libre de $K+4$ *splines* cubiques. La dimension de l'espace vectoriel associé est donc d'au moins ce nombre. Pour conclure, il faudrait décrire une méthode permettant de générer n'importe quelle *spline* cubique à partir de cette famille de vecteurs, ce qui mériterait un raisonnement un peu plus étoffé.

Cet exemple met en évidence l'existence de fonctions exprimables simplement permettant de retranscrire l'utilisation d'une *spline* dans un modèle de régression. Notons par ailleurs que si les vecteurs de base décrits ci-dessus présentent une expression analytique relativement simple, ils ne sont pas particulièrement attractifs d'un point de vue computationnel puisque les puissances

[11] Une *spline* cubique doit être continue, continûment dérivable, et deux fois continûment dérivable en chacun de ses nœuds, ce qui peut s'interpréter empiriquement comme étant trois contraintes élémentaires distinctes.

de grands nombres peuvent conduire à de véritables problèmes de troncatures. En réalité, les procédures informatiques utilisent davantage une représentation moins triviale conceptuellement, mais qui présente un avantage numérique considérable : les **B-*splines*** (ou *basis splines*). Celles-ci ont la particularité de posséder un support compact minimal, ce qui peut être astucieusement exploité pour réduire la complexité informatique des algorithmes utilisés. La description formelle des B-*splines* et de leurs implications computationnelles dépasse largement le cadre de ce mémoire, et le lecteur intéressé pourra se référer au manuel de de Boor (1978 [11]) qui fait figure d'autorité en la matière.

2.3.4 *Spline* naturelle

Les ***splines* naturelles** sont des *splines* respectant des contraintes de régularité additionnelles : ces fonctions doivent présenter un comportement linéaire aux deux frontières extérieures du domaine de définition de la variable sous-jacente. Ces deux contraintes supplémentaires assurent que les *splines* employées produisent des estimateurs plus stables aux frontières, et moins incertains. Cet atout est donc particulièrement important en tarification puisque les tarifs ne peuvent pas diverger de manière brutale sur ces segments extrêmes, dans lesquels la population est d'ailleurs souvent minoritaire. L'utilisation de ces fonctions dans un modèle de régression est similaire à celle des *splines* classiques, puisque qu'elles sont également entièrement caractérisées par des *basis splines*. Ces fonctions plus stables sont donc préférables pour l'élaboration de modèles GAM.

À ce stade, nous avons exposé diverses transformations f_j susceptibles de modéliser des effets non linéaires au sein d'un modèle GAM, comme décrit par l'équation 15. Les régressions polynomiales, les fonctions en escalier, les *splines* de formes variées constituent une liste non exhaustive mais déjà diversifiée d'outils pour composer un modèle additif. De nombreuses autres possibilités existent mais elles demeurent toutefois plus complexes. Parmi celles-ci, Pouna Siewe (2010 [30]) et Huther (2014 [21]) emploient deux techniques additionnelles, que n'avons pas décrites ici. Premièrement, les *smoothing splines* sont des alternatives aux *splines* traditionnelles qui, au lien de produire des transformations directes d'une variable explicative, se traduisent par l'ajout d'un terme de pénalisation à la fonction d'erreur du modèle, à la manière du Lasso, décrit dans une partie ultérieure. Cependant, ce nouveau type de *spline* n'est pas calibré en amont du modèle, comme pour les *splines* traditionnelles. À la place, elles doivent faire l'objet d'une estimation spécifique à chaque prédicteur, *ceteris paribus*, mais au sein même de la construction du modèle. Cela est réalisable à l'aide d'un algorithme d'ajustement rétrograde (*backfitting algorithm*), qui optimise à chaque itération les différentes *smoothing splines* de manière indépendante. Cette méthode est donc beaucoup moins simple à mettre en œuvre, et requiert de surcroît un temps de calcul plus important. Deuxièmement, les **régressions locales** utilisent, pour chaque observation de test dont il faut prédire l'étiquette, le modèle bâti à partir des points de la base d'apprentissage localisés dans un voisinage de celle-ci. Cette seconde technique peut être également intégrée dans un modèle GAM global. Néanmoins, elle est aussi très coûteuse en temps de calcul, et se révèle de surcroît très médiocre en grande dimension, tout comme le sont tous les modèles basés sur le principe du voisinage, tels que la méthode des **plus proches voisins**.

2.4 Régression Multivariée par *Spline* Adaptative

Lorsque dans la littérature statistique, des modèles économétriques traditionnels sont confrontés à des arbres de décision (CART), la façon donc ces deux familles de méthodes s'articulent dans le vaste ensemble que représente l'apprentissage statistique est rarement évoquée. Les modèles MARS (*Multivariate Adaptive Regression Spline*) représentent le lien qui unifie GAM et CART dans une même logique de modélisation. Sans englober les GAM, les MARS s'affranchissent d'une hypothèse fondamentale de ces derniers, l'additivité, tout en introduisant de nouvelles contraintes, l'utilisation de *splines* très simples.

Plus précisément, un MARS est fondamentalement un algorithme combinant des *splines* linéaires de la forme $(X_j - s)^+$ et $(s - X_j)^+$ pour former un modèle de prédiction non linéaire. La contrainte sur les *splines* utilisées limite considérablement l'ensemble des fonctions de prédiction candidates, et contribue par conséquent à lutter contre le sur-apprentissage. En revanche, ce modèle introduit automatiquement des termes d'interaction entre variables, *a contrario* des GAM. C'est cette différence significative qui permet aux MARS de se démarquer sensiblement des modèles traditionnels. D'autre part, remarquons que le recours à une procédure automatisée constitue une autre particularité des MARS, et marque ainsi l'important fossé conceptuel qui divise la statistique et l'apprentissage automatique, l'économétrie et l'algorithmique, et que nous franchissons ainsi avec ces modèles.

L'algorithme MARS se compose de deux étapes distinctes. La première étape, qui s'apparente à une procédure *stepwise forward*, et qui est décrite par l'algorithme 1, permet de construire de manière itérative un modèle qui se révèle souvent trop flexible. La seconde étape consiste à supprimer les termes superflus par une procédure *backward* en utilisant la **validation croisée généralisée**, ce qui revient à minimiser une fonction d'erreur pénalisée par le nombre de paramètres[12].

Il faut noter ici que deux paramètres permettent de limiter le sur-apprentissage du modèle : une limite du nombre de termes additifs composant le modèle et une limite du degré d'interactions de chaque terme.

2.5 Arbre de décision

Les arbres de décision sont des modèles d'apprentissage non linéaires qui s'inspirent largement de l'esprit des MARS, en particulier sur les sujets des interactions et des *splines* élémentaires. Leur principe est de construire une **classification hiérarchique descendante** des observations en des **catégories homogènes** vis-à-vis de la variable réponse. Chacune de ces classes est ensuite **étiquetée** par la valeur moyenne des sorties des observations qu'elle contient.

À titre d'exemple, la figure 12 représente l'arborescence produite par la croissance d'un arbre binaire à partir de deux variables explicatives. La fonction de prédiction issue de cet arbre partitionne le plan cartésien selon l'illustration 13.

[12]Cette approche ne s'apparente donc pas à de la validation croisée, puisqu'il s'agit bien de comparer les erreurs résiduelles sur les données d'apprentissage, avec une fonction de perte légèrement modifiée.

Algorithme 1 : Étape *forward* d'une procédure MARS

Données : Un ensemble de *splines* candidates :

$$C = \left\{ (X_j - s)^+, (s - X_j)^+ \;\middle|\; \begin{array}{l} s \in \{x_{1j}, \ldots, x_{Nj}\} \\ j = 1, 2, \ldots, p \end{array} \right\}$$

`/* Notons que nous avons ici` $|C| = 2pN$ `splines candidates */`

Entrées : Deux paramètres de tolérance :

- T = nombre maximal de termes dans le modèle final
- I = degré maximal d'interaction pour chaque terme

début

Initialisation avec le modèle trivial : $M : Y_i \sim 1 + \varepsilon_i$

tant que $\begin{cases} |M| < T \\ |interactions| < I \end{cases}$ **faire**

pour chaque *paire de splines* $(X_j - s)^+, (s - X_j)^+ \in C$ **faire**

pour chaque *terme* $h_m \in M$ *tel que* $X_j \notin h_m$ **faire**

Construire le modèle augmenté M' :

$M' \leftarrow M + h_m \cdot (X_j - s)^+ + h_m \cdot (s - X_j)^+$

Calculer l'erreur résiduelle du modèle actualisé M' : $||Y_i - \hat{Y}_i||_2$

fin

fin

Mettre à jour le modèle $M \leftarrow M'$ avec l'ajout qui minimise l'erreur résiduelle

fin

fin

Sorties : Retourner le modèle M

FIGURE 12 – Arbre décisionnel bâti avec deux variables explicatives

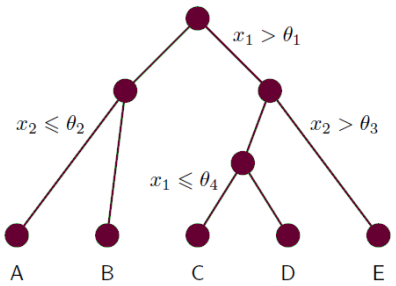
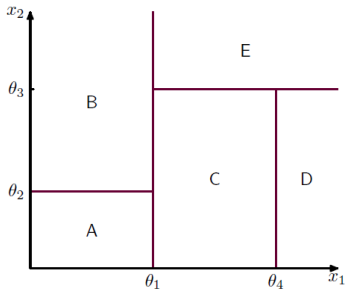


FIGURE 13 – Partition du plan cartésien associée à l'arbre décisionnel



2.5.1 Croissance de l'arbre

La croissance de l'arbre est effectuée par **partitionnement récursif** de la population initiale selon les profils $(X_j)_j$ en minimisant à chaque palier l'erreur de prédiction du modèle. L'algorithme le plus utilisé est celui du **CART** (*Classification And Regression Tree*, Breiman [6]) qui produit des **arbres binaires**, et dont la procédure est détaillée par l'algorithme 2.

À chaque nœud de l'arbre, sont considérées successivement les divisions binaires déterminées par les indicatrices $\mathbb{1}\{X_j < s\}$, pour toutes les variables quantitatives X_j et toutes les valeurs s existantes dans la population considérée. Il y a donc, pour chaque variable, $m_j - 1$ divisions possibles, où m_j est le nombre de valeurs de la variable X_j. Dans le cas de variables qualitatives,

toutes les partitions binaires possibles des modalités sont considérées, soit $2^{m_j-1}-1$ possibilités. Parmi toutes ces divisions possibles, celle qui **minimise** l'erreur de prédiction associée est choisie :

$$(j^*, s^*) = \underset{(j,s)}{\arg\min} \left\{ \sum_{x_i \in \{X|X_j<s\}} \left(y_i - \bar{y}_{\{X|X_j<s\}}\right)^2 + \sum_{x_i \in \{X|X_j \geq s\}} \left(y_i - \bar{y}_{\{X|X_j \geq s\}}\right)^2 \right\} \tag{20}$$

où $\bar{y}_{\{X|X_j<s\}}$ correspond à la moyenne des étiquettes des observations respectant la condition $X_j < s$. La fonction objectif de l'équation 20 est appelée **hétérogénéité**, ou encore **impureté** de l'arbre. Il s'agit donc à chaque étape de maximiser la réduction d'hétérogénéité du CART. Cet algorithme peut être poursuivi jusqu'à l'obtention de nœuds qui ne contiennent que des observations dont les profils sont identiques. À l'issue de cette procédure, aucune division ne permet plus de produire deux nœuds-fils non vides : une **feuille** de l'arbre a donc été atteinte. À ce stade, un arbre dit **saturé** a été construit. Celui-ci peut alors être utilisé pour prédire l'étiquette d'une nouvelle observation. Pour ce faire, il faut insérer cette observation témoin dans la feuille correspondant à son profil : les tests binaires de chaque palier nous permettent de suivre le chemin à travers l'arbre qui nous guide jusqu'à celle-ci. La valeur prédite est alors déterminée en moyennant[13] les étiquettes des observations d'apprentissage se trouvant dans cette feuille.

2.5.2 Élagage de l'arbre

Parce que le nombre de feuilles produites par un arbre saturé peut être très élevé, le modèle précédant est particulièrement exposé au risque de **sur-apprentissage**. Ce risque apparaît lorsque survient un sur-ajustement vis-à-vis de la base d'apprentissage. Par conséquent, le modèle est difficilement généralisable à des données indépendantes. Ce phénomène se traduit par des paramètres trop nombreux en régression standard, excès contre lequel il est possible de lutter grâce à la sélection automatique de variables ou la pénalisation. De même, le sur-apprentissage est présent dans les arbres de décision à travers une arborescence profonde et complexe. C'est la raison pour laquelle il est indispensable d'élaguer ceux-ci, c'est-à-dire de sélectionner un sous-arbre qui produise une **prédiction robuste** vis-à-vis de bases de test indépendantes. Plusieurs méthodes d'élagage existent. La plus simple consiste à comparer les erreurs de prédiction de tous les sous-arbres possibles sur une base de validation indépendante. Mais comme le nombre de sous-arbres possibles peut être très important, il est préférable d'adopter l'alternative moins coûteuse décrite ci-après. Définissons dans un premier temps le **paramètre de complexité** α qui est un réel positif auquel est associé un sous-arbre unique par l'expression :

$$\forall \alpha \in \mathbb{R}^+, T(\alpha) = \underset{T}{\arg\min} \left\{ \sum_{m=1}^{|T|} \sum_{x_i \in K_m} (y_i - \bar{y}_{K_m})^2 + \alpha |T| \right\} \tag{21}$$

où $|T|$ est une pénalité qui correspond au nombre de feuilles du sous-arbre T, et K_m désigne la $m^{\text{ième}}$ feuille de T. Le coefficient α est en réalité un paramètre de pénalisation, qui détermine le poids accordé à la pénalité ajoutée à l'erreur de prédiction. Le cas particulier $\alpha = 0$ correspond à l'arbre saturé. A l'opposé, à partir d'un certain seuil, lorsque le terme pénalisant est trop important, seul l'arbre tronc à une seule feuille subsiste. Avec cette définition, il est possible de montrer que

[13] Dans le cas d'une classification, la règle de prédiction est celle du vote : à la nouvelle observation est attribuée l'étiquette majoritaire de la feuille correspondante.

Algorithme 2 : Croissance d'un CART

Entrées : Un arbre tronc T qui attribue la moyenne globale des Y_i à l'ensemble de la population

tant que *l'erreur de prédiction associée au nouvel arbre T est réduite par la dernière itération* **faire**

- **pour chaque** *nœud terminal C de l'arbre T* **faire**
  - **pour chaque** *prédicteur X_j* **faire**
    - **si** *X_j est quantitatif* **alors**
      - **pour chaque** *valeur s de X_j* **faire**
        - Classer les observations présentes dans le nœud C selon la division binaire associée à ce seuil s entre ses deux nœuds-fils C_1 et C_2
        - Étiqueter chacune de ces observations par la valeur moyenne de Y_i dans le nœud-fils auquel elle appartient : $\bar{y}_{\{X|X_j<s\}}$ ou $\bar{y}_{\{X|X_j\geq s\}}$
        - Calculer la réduction d'hétérogénéité associée à l'ajout de C_1 et C_2 au sein de l'arbre : $\sum_{x_i\in C}(y_i-\bar{y}_C)^2-\left[\sum_{x_i\in C_1}(y_i-\bar{y}_{C_1})^2+\sum_{x_i\in C_2}(y_i-\bar{y}_{C_2})^2\right]$
      - **fin**
    - **si** *X_j est qualitatif* **alors**
      - **pour chaque** *partition binaire de l'ensemble des modalités de X_j* **faire**
        - Classer les observations de C selon cette division en C_1 et C_2
        - Étiqueter ces observations par le valeur moyenne de Y_i dans C_1 ou C_2
        - Calculer la réduction d'hétérogénéité associée à cette division
      - **fin**
    - **fin**
    - Mettre à jour l'arbre T avec la division binaire du nœud C qui maximise la réduction d'hétérogénéité
  - **fin**
- **fin**

fin

Sorties : La fonction de prédiction associée à l'arbre T

l'élagage par examen exhaustif de l'ensemble des sous-arbres possibles est équivalent au calibrage du paramètre α. Cela permet donc de réduire sensiblement le nombre de solutions à parcourir. En calibrant[14] α, il est alors possible de déterminer un sous-arbre robuste.

2.5.3 CART *versus* MARS

Lorsque nous parlions des MARS (partie 2.4), nous évoquions comment ceux-ci formaient un lien théorique important entre les GAM et les CART. Pour bien comprendre les atouts et faiblesses des arbres de décision, il nous parait judicieux d'expliciter le cheminement scientifique qui permet de convertir un modèle économétrique traditionnel en algorithme d'apprentissage statistique.

Il a déjà été vu que les MARS amélioraient la flexibilité des GAM à travers l'introduction d'éventuelles interactions, tout en se restreignant à l'utilisation de *splines* très simples. Ces deux familles de modèles exhibent donc une souche commune mais ne sont pas pour autant incluses l'une dans l'autre. De même les arbres de décision ne sont ni une généralisation, ni un cas particulier des MARS, mais présentent néanmoins une singulière familiarité avec ceux-ci. Plus précisément,

[14] Par validation croisée, par exemple, dont le principe est détaillé en partie 3.1. Le lecteur pourra également se référer à Therneau (1997 [31]) pour toute précision complémentaire.

les CART peuvent être considérés comme des MARS pour lesquels :

- les *splines* linéaires sont remplacées par des fonctions en escalier ;
- un terme du modèle déjà retenu pour former une double interaction ne peut plus être disponible pour des interactions supplémentaires, ce qui fixe ainsi les divisions binaires.

Il conviendra de distinguer alors les spécificités du CART en comparaison des modèles précédents :

- il autorise les interactions, et est à ce titre, tout comme les MARS, plus flexible que les GLM ou même les GAM ;
- il est facilement interprétable de par sa structure d'arborescence, à condition qu'il ne soit pas agrégé par une méthode ensembliste, comme une forêt aléatoire (partie 2.6) ;
- mais il produit des prédictions globalement moins lisses qu'un MARS, puisqu'il peut présenter des discontinuités aux nœuds, ce qui est possible dès lors que des *splines* de degré 0 sont employées ;
- et il ne permet plus d'identifier clairement les structures additives car il est limité par son arborescence binaire.

2.6 Forêt aléatoire

Comme nous l'avons vu dans la partie précédente, l'arbre décisionnel produit par l'algorithme CART apporte une grande flexibilité dans la modélisation d'une variable d'intérêt, à travers notamment la capture d'effets non linéaires et l'introduction d'interactions. Toutefois, ces progrès s'accompagnent également d'un risque augmenté de sur-apprentissage. En effet, l'arbre CART est généralement instable et peut varier grandement en fonction de la base d'apprentissage utilisée.

Pour pallier ce défaut, des techniques d'agrégation de plusieurs arbres, aussi appelées **méthodes ensemblistes**, sont couramment utilisées afin de produire des modèles plus robustes. Ces approches introduisent volontairement diverses formes d'aléa au sein de l'algorithme de croissance pour obtenir ainsi un large éventail d'arbres légèrement dissemblables, et présumés décorrélés. Le modèle consistant à moyenner les prédictions issues de ces arbres offre alors des résultats plus stables, généralisables à des bases de test indépendantes.

Parmi ces méthodes, la **forêt aléatoire** (*random forest*), initialement introduite par Breiman (2001 [7]), est sans aucun doute la plus populaire. Non seulement celle-ci est-elle communément considérée comme le modèle d'arbres décisionnels le plus performant, mais elle est également, pour nombre de spécialistes, la vitrine même du *machine learning*. En effet, tout article de recherche comparant diverses méthodes d'apprentissage statistique emploie systématiquement la forêt aléatoire, et celle-ci s'impose souvent comme un modèle de référence aux yeux de l'auteur. De même, les conclusions pratiques de ces études célèbrent régulièrement la qualité des prédictions produites par cette méthode. Aussi, c'est pour ces raisons empiriques que nous nous proposons d'utiliser cette technique à des fins de comparaison avec nos modèles économétriques. Le principe de cet algorithme, qui fait appel à deux formes d'aléa distinctes, est décrit ci-après.

2.6.1 Ré-échantillonnage aléatoire

La première modification du CART introduite par la forêt aléatoire impacte l'échantillon sur lequel est construit l'arbre de décision. En effet, une technique de ré-échantillonage aléatoire (*bootstrap*

aggregating ou *bagging*) est appliquée à la base originelle en amont de la croissance de chaque nouvel arbre. Il s'agit de sélectionner, à chaque itération, un **échantillon avec remise** de la base d'apprentissage, puis de construire un arbre classique sur cet échantillon. L'agrégation des différents arbres ainsi bâtis constitue la méthode du *bagging*. Notons que la taille de l'échantillon formé par *bootstrap* est généralement la même que celle de la base initiale, mais il est tout à fait envisageable de recourir à du sous-échantillonage, pour des raisons computationnelles. Enfin, précisons que le nombre d'arbres composant le modèle peut être calibré par validation croisée (partie 3.1), bien qu'un grand nombre d'itérations favorise généralement la robustesse des prédictions.

2.6.2 Parcours d'un sous-espace aléatoire

Une seconde évolution de l'algorithme doit être implémentée afin d'obtenir une véritable forêt aléatoire. Il s'agit d'incorporer un nouvel aléa, impactant cette fois-ci non plus les observations, mais les variables utilisées dans la construction de l'arbre. À chaque palier, lors de la génération d'un nouveau nœud, la procédure de recherche de la division binaire optimale est modifiée pour parcourir un **sous-ensemble aléatoire** des variables candidates. Cette nouvelle évolution est censée contrer le risque de corrélation des arbres produits par un simple *bagging*. Ce risque est particulièrement élevé dans la situation dans laquelle certaines variables détiennent un pouvoir explicatif important et apparaissent donc dans la plupart des arbres construits, les rendant ainsi fortement corrélés entre eux. La taille du sous-espace doit être idéalement calibrée par validation croisée, mais des valeurs empiriques sont également proposées par Breiman : $\sqrt{p}$ pour un problème de classification et $\frac{p}{3}$ pour un problème de régression, où p désigne le nombre de variables explicatives initiales. L'algorithme 3 résume la procédure générant une forêt aléatoire.

Algorithme 3 : Génération d'une forêt aléatoire

Entrées : Deux paramètres d'apprentissage :
- Nombre d'itérations : B
- Taille du sous-espace : m

pour $k \leftarrow 1$ **a** B **faire**
  Sélectionner un échantillon avec remise de la base initiale
  Initialiser l'arbre T_k s'appuyant sur cet échantillon avec sa racine (moyenne globale des Y_i)
  pour chaque *nœud terminal* C *de* T_k **faire**
    Sélectionner un sous-espace aléatoire de taille m des variables explicatives
    Déterminer la division optimale parmi l'ensemble des partitions binaires des m variables considérées
  fin
fin

Sorties : Retourner le modèle agrégé : $T = \frac{1}{B}\sum_{k=1}^{B} T_k$

En définitive, la forêt aléatoire s'impose parmi les diverses méthodes de *machine learning* comme un modèle de référence. Non seulement elle introduit au sein des arbres bâtis deux sources d'aléa distinctes afin de limiter la variance des erreurs de prédiction, mais elle apporte aussi des nouveaux outils d'aide à l'interprétation, qui se révéleront particulièrement pertinents lors de

l'étude comparative. Ces outils d'ordre pratique, qui tirent profit des échantillonnages propres à la forêt, sont décrits en partie 3.4.

2.7 Exposition

Comme indiqué au cours de la revue de littérature, l'exposition de chaque observation au sein du portefeuille d'étude impacte fortement la mesure de la fréquence des sinistres et mérite, à ce titre, une considération spécifique. En effet, le nombre de sinistres imputables à chaque contrat correspond à une exposition différente. Cette mesure de durée n'est donc pas constante pour chaque observation. Afin de prendre en compte cette hétérogénéité, il faut plutôt modéliser le taux de sinistres par durée élémentaire et ce, quelque soit le modèle considéré. Examinons en premier lieu le cas économétrique. En reprenant les notations précédentes, si E_i est la variable d'exposition, alors un modèle économétrique plus cohérent est :

$$\mathbb{E}\left[\left.\frac{Y_i}{E_i}\right| X_i\right] = \mathrm{e}^{X_i^t\beta} \tag{22}$$

Comme la variable E_i est déterministe, elle peut sortir de l'espérance, et passer dans l'autre membre du modèle, et même s'intégrer au sein du prédicteur linéaire :

$$\begin{aligned}\mathbb{E}[Y_i|X_i] &= E_i \cdot \mathrm{e}^{X_i^t\beta} \\ \mathbb{E}[Y_i|X_i] &= \mathrm{e}^{\ln(E_i) + X_i^t\beta}\end{aligned} \tag{23}$$

Il convient de remarquer que l'exposition n'est pas une variable explicative du modèle en soi. Il n'apparaît donc pas de coefficient de régression associé à celle-ci au sein de l'estimation. En pratique, cela se traduit par l'ajout d'un *offset* dans la spécification opérationnelle du modèle, c'est-à-dire l'introduction d'une variable additionnelle dont le paramètre associé est fixé à 1. Précisons par ailleurs que ce paramétrage n'est pas équivalent à l'utilisation de régressions pondérées. Ces dernières conviennent lorsqu'une observation représente plusieurs individus, par exemple dans le cas de sondages pour lesquels il est commun d'utiliser des panels représentatifs de la population ciblée. Dans cette situation, il est alors naturel de pondérer chaque observation au sein de la vraisemblance du modèle en fonction de l'importance du groupe d'individus qu'elle représente. Toutefois, cette méthode ne convient pas pour l'intégration de la variable d'exposition puisqu'il ne s'agit pas de donner ici moins de poids aux observations partiellement exposées, mais bien de recalculer leur probabilité de sinistralité **sans biais**.

De même, dans un souci de cohérence, il est nécessaire de corriger l'effet de l'exposition dans les modèles basés sur les arbres de décision. Il s'agit toujours de prédire le taux de sinistres $\frac{Y_i}{E_i}$ au sein de la construction de l'arbre. Cette modification apparaît alors dans l'écriture de la fonction d'hétérogénéité (équation 20) :

$$\sum_{X_i \in K_m} \left(\frac{Y_i}{E_i} - \bar{y}_{K_m}\right)^2 \tag{24}$$

Le lecteur retrouvera ainsi l'approche proposée par Paglia (2010 [28]). Et comme celui-ci l'indique, il est généralement possible dans la plupart des logiciels d'introduire une fonction d'hétérogénéité alternative pour alimenter les algorithmes implémentés par défaut.

3 Sélection de variables

Les techniques de sélection de variables présentées ici permettent de répondre à diverses contraintes opérationnelles du marché de l'assurance. À ce titre, elle sont donc tout à fait envisageables en tant que méthodes alternatives de l'approche existante. Tout d'abord, le Lasso répond à la contrainte du temps de calcul, qui est particulièrement importante en situation opérationnelle, puisque les assureurs modernes pilotent leur portefeuille de manière extrêmement régulière. Certains assureurs directs réalisent même des révisions tarifaires mensuellement. L'évolution des modèles doit donc suivre impérativement cette cadence rapide, pour produire des résultats fréquents. Ensuite, le Lasso répond aussi à la contrainte de précision des tarifs. En effet, la robustesse reconnue de cette technique conduit en pratique à estimer des tarifs plus justes en moyenne sur de gros portefeuilles.

Bien que les modèles linéaires soient réputés pour être globalement robustes, et présenter une certaine résistance au phénomène de sur-apprentissage en comparaison de méthodes plus flexibles, cet atout est néanmoins mis à mal dès lors que le nombre de variables explicatives augmente considérablement. En effet, les modèles économétriques classiques sont particulièrement sensibles au **fléau de la dimension**[15], le célèbre phénomène de détérioration du pouvoir prédictif en grande dimension. Dans ce contexte, il est crucial de recourir à des techniques de réduction de la dimension. Cela peut être effectué par des méthodes factorielles, telles que l'Analyse en Composantes Principales (ACP), qui présentent cependant l'inconvénient de masquer les variables originelles, ou alors par des techniques de sélection de variables, qui conservent l'interprétabilité du modèle élagué. Parmi ces dernières, les procédures *stepwise* sont les plus usitées, mais celles-ci peuvent être coûteuses en temps de calcul, et conduire à des sélections biaisées. Il nous paraît donc judicieux de privilégier une technique alternative, le Lasso, qui pallie en partie ces deux défauts. Afin de s'assurer de la validité de notre choix, nous examinons tout de même les résultats produits par deux méthodes complémentaires : le *stepwise*, qui demeure incontournable malgré ses divers inconvénients, et la forêt aléatoire, qui est également exploitable pour sélectionner les variables les plus pertinentes. Après l'introduction de préliminaires en matière de validation de modèles, nous décrivons ci-après les trois approchés étudiées. Nous examinerons ultérieurement les sélections auxquelles elles conduisent (cf. partie 6).

3.1 Préliminaire : validation croisée

Les techniques de sélection de variables sont généralement des procédures itératives, réalisant des comparaisons successives de variations du modèle initial, afin d'obtenir à terme une combinaison idéale de variables explicatives. Ces comparaisons nécessitent l'utilisation d'une mesure de la fiabilité des modèles considérés. La mesure classique d'erreur d'estimation $||Y_i - \hat{Y}_i||_{L2}$ évalue dans un sens la performance du modèle, mais elle est directement impactée par le phénomène de sur-apprentissage. Elle ne permet donc pas d'apprécier la capacité du modèle à se généraliser à des données indépendantes de la base d'apprentissage. Face à cet enjeu, l'économétrie traditionnelle a

[15] La première mention de cette *curse of dimensionality* provient sans doute de Bellman [2] qui décrit le phénomène d'augmentation exponentielle du volume d'un espace avec sa dimension, ce qui implique que les données disponibles deviennent rapidement clairsemées (*sparse*) et perdent ainsi en significativité statistique.

connu la popularisation croissante de mesures régularisées, pénalisant l'indicateur d'apprentissage du modèle par sa complexité. Plus précisément, de telles métriques adaptées aux GLM comptent notamment parmi elles les critères d'information (AIC, BIC), qui sont des pondérations entre la vraisemblance statistique du modèle, et son nombre de paramètres (i.e. variables). Il existe d'autres indices également très répandus, comme les très nombreuses versions de R^2 ajustés, analogues au R^2 de la régression linéaire classique, mais intégrant de même la vraisemblance du modèle, et une mesure de sa complexité. L'ensemble de ces métriques sont toutefois arbitraires, et peuvent d'ailleurs produire des résultats dissemblables. Le BIC est par exemple réputé pour retenir des solutions davantage pénalisées qu'avec l'AIC. Une réponse théoriquement simple à ce problème, mais souvent délicate à mettre en œuvre, serait de recourir à une base de test indépendante de la base d'apprentissage, afin de mesurer ainsi l'erreur généralisée du modèle. L'ensemble des approches suivant ce principe est désigné sous l'expression de **validation croisée**, dont nous détaillons les caractéristiques ci-après.

La **validation croisée** est une technique empirique d'évaluation de la performance prédictive d'un modèle statistique. Il s'agit en réalité de mesurer sa capacité à se généraliser à une base de données indépendante de la base utilisée pour son estimation. La méthodologie originelle consiste en réalité à partitionner la base initiale en deux ou trois sous-ensembles complémentaires, qui ont alors des fonctions différentes :

- les modèles candidats sont bâtis sur une première sous-base, dite d'**apprentissage**, par exemple 50% ;
- la sélection de modèles est effectuée à partir des mesures de performance calculées sur une seconde sous-base, dite de **validation**, par exemple 30% ;
- la qualité du modèle final est éventuellement évaluée sur une troisième sous-base, dite de **test**, par exemple 20%.

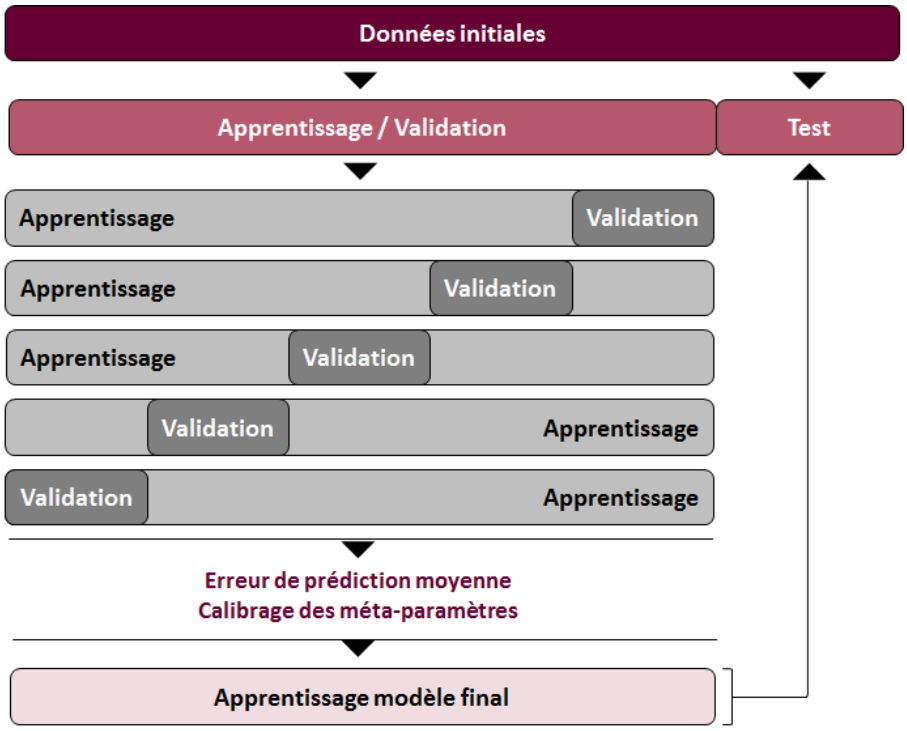
Cette approche a été vivement critiquée pour son caractère heuristique, sans fondement théorique, mais elle s'est révélée particulièrement efficace dans les problèmes de prédiction. Toutefois, elle demeure délicate à mettre en œuvre, car l'utilisation d'une ou deux bases de validation/test suppose de réduire considérablement la base utilisée pour l'apprentissage du modèle, ce qui impacte sensiblement sa qualité. En outre, la manipulation de ces bases requiert une certaine logistique informatique, en particulier lorsque les partitions apprentissage/test sont nombreuses et aléatoires, comme dans les méthodes de validation croisée les plus complexes, que nous présentons maintenant.

Les mesures calculées par validation croisée peuvent présenter une certaine dépendance vis-à-vis de la partition sélectionnée. Afin de s'affranchir de l'arbitrarité de cette démarche, il est parfois recommandé de répéter la procédure précédente en intervertissant les bases d'apprentissage et de validation, c'est-à-dire de bâtir un second modèle sur la base de validation puis d'évaluer sa performance sur la base d'apprentissage. Les mesures moyennées sur ces deux modèles sont alors plus robustes et moins sensibles aux divergences de structure entre les deux sous-bases. Pour aller plus loin, il est même envisageable d'échantillonner aléatoirement la base initiale en plusieurs partitions apprentissage/validation indépendantes. Les mesures ainsi calculées sont alors moyennées sur l'ensemble des partitions utilisées.

La principale application de la validation croisée est donc la sélection de modèles et, *a fortiori*, la sélection de variables. Mais elle est aussi utilisée en tant que méthode de calibrage d'un **meta-**

paramètre[16] au sein même de l'apprentissage du modèle. En particulier, cette technique sera employée pour le calibrage du paramètre de pénalisation λ du Lasso (cf. partie 3.3). La version couramment utilisée est celle du 5-***fold*** qui consiste à diviser la base de données en cinq sous-bases indépendantes de tailles égales. Le modèle est bâti sur l'ensemble formé par les quatre premières sous-bases, puis sa performance est évaluée sur la cinquième sous-base, qui fait alors office de base de validation. Puis cette étape est répétée pour les cinq partitions apprentissage/validation envisageables (cf. figure 14). Enfin, la valeur du méta-paramètre optimale est déterminée en minimisant l'erreur de prédiction moyennée sur ces cinq configurations pour un jeu arbitraire de valeurs candidates. Cette valeur est alors utilisée pour l'élaboration du modèle final sur la totalité de la base de données.

FIGURE 14 – Validation croisée par 5-*fold*



Dans cette étude, seule une unique partition binaire apprentissage/validation *versus* base de test a été opérée, selon un ratio de coupe de 50%, sans échantillonnage ou *bootstrap* supplémentaires. L'ensemble des modèles ont été donc été élaborés sur la première moitié de la base. Les méthodes nécessitant le calibrage de méta-paramètres, comme le Lasso ou la forêt aléatoire, se sont appuyées sur l'approche 5-*fold*. Les résultats produits par les différents modèles ont ensuite été comparés sur la seconde moitié de la base.

3.2 *Stepwise*

Afin d'élaguer le modèle économétrique des variables les moins pertinentes, et pour ainsi limiter le sur-paramétrage, l'approche classique est une procédure itérative bien connue des économètres : la

[16] Il s'agit d'un paramètre additionnel dont le calibrage est nécessaire à l'estimation des paramètres primaires du modèle.

méthode *stepwise*. Celle-ci peut prendre plusieurs formes. La version *forward*, consiste par exemple, en partant du modèle trivial composé uniquement d'une constante, à considérer successivement l'ajout d'une nouvelle variable explicative au modèle. En sélectionnant la variable la plus pertinente à chaque itération, et en arrêtant le processus dès que le pouvoir explicatif du modèle n'augmente plus, cette approche réduit ainsi sensiblement la complexité du modèle, en faveur de son interprétabilité. Mais cette technique peut souvent se révéler inappropriée de par sa complexité informatique, et son importante volatilité. Nous exposons ci-après le principe de cette méthode ainsi que ses différents défauts, qui nous ont menés à privilégier *in fine* une approche alternative.

3.2.1 Procédure automatique

Le principe de la procédure *stepwise forward* est décrit avec plus de précision par l'algorithme 4.

Algorithme 4 : Procédure *stepwise forward*

Données : Ensemble de prédicteurs $P = \{X_j\}$

Entrées : Initialisation avec le modèle trivial à constante : $M : Y_i \sim 1 + \varepsilon_i$

tant que *l'AIC du modèle M est réduit par la dernière itération* **faire**

  pour chaque $X_j \in P$ **faire**

    Construire le modèle augmenté $M' : M' \leftarrow M + X_j$

    Calculer l'erreur résiduelle du modèle actualisé $M' : ||Y_i - \hat{Y}_i||_2$

  fin

  Mettre à jour le modèle $M \leftarrow M'$ avec l'ajout qui minimise l'erreur résiduelle

  Supprimer la variable X_j sélectionnée de l'ensemble P

fin

Sorties : Retourner le modèle M

La mesure de performance utilisée pour la comparaison, à chaque itération de la procédure, est communément le critère d'information **AIC** (*Akaike's Information Criteria*). Ce critère pénalise la vraisemblance du modèle par le nombre de paramètres – en l'occurrence, de variables –, afin de privilégier des modèles parcimonieux et limiter ainsi le sur-apprentissage. Le choix du meilleur modèle se traduit donc par la minimisation de l'AIC, qui est défini par la quantité :

$$\mathrm{AIC} = 2k - 2\ln(L)$$

où k est le nombre de paramètres, et L la vraisemblance du modèle.

Remarquons que de nombreuses procédures alternatives permettent de sélectionner les variables les plus pertinentes : l'approche *backward* qui est le procédé inverse de la version *forward* ; une méthode hybride combinant ces deux techniques ; ou encore la sélection exhaustive par force brute, appelée *best subset selection*. De même, différentes mesures du pouvoir explicatif du modèle peuvent être utilisées au cours de la procédure afin de conserver les termes les plus pertinents : le BIC (*Bayesian Information Critera*) qui a la réputation de produire des modèles plus parcimonieux que l'AIC ; ou la large variété de R^2 ajustés, qui sont des variantes du R^2 impliquant également la vraisemblance du modèle d'une part, et diverses formes de pénalités, basées entre autres sur le nombre de degrés de liberté, d'autre part. Alors que ces nombreuses alternatives peuvent produire des résultats légèrement différents, elles ne se distinguent pas de manière drastique. Finalement,

nous retenons la procédure *forward* qui est la moins coûteuse en temps de calcul puisqu'elle débute par l'examen de modèles comportant peu de variables. Pour le critère de sélection, nous choisissons l'AIC, qui se justifie par ses fondements lié à la théorie de l'information : il s'agit d'une estimation de l'information perdue lors de la simplification de la réalité opérée le modèle.

Quelles sont les différents problèmes induits par la procédure *stepwise* ? Pourquoi préférer une approche alternative ? Tout d'abord, le coût computationnel de ce type de méthode itérative est particulièrement problématique. En particulier, pour le modèle de fréquence, le volume des données d'apprentissage interdit son utilisation en un temps raisonnable. Mais des raisons statistiques vont également à l'encontre de cette procédure. Plus précisément, cette technique invalide systématiquement les hypothèses qui lui sont généralement attribuées et elle est donc, à ce titre, vivement critiquée. Nous discutons maintenant de quelques-unes de ces critiques.

3.2.2 *p-hacking*

À propos de la régression *stepwise*, deux sujets polémiques méritent, à notre sens, d'être discutés : le premier concerne les critiques théoriques à l'égard de cette procédure ; le second est un dilemme opérationnel qui résulte de cette controverse.

Premièrement, les résultats de la régression *stepwise* sont fréquemment exploités de manière impropre, sans prendre en compte les implications de la procédure de sélection. En particulier, les estimateurs et les intervalles de confiance généralement présentés, y compris ceux du tableau 11, sont propres au modèle final, issu de cette procédure, et ne sont donc pas ajustés pour refléter convenablement l'incertitude liée à cette sélection. Cette pratique a suscité de vives critiques (Harrell, 1996 [19]) dont les principales sont les suivantes :

- **les tests de significativité entre deux itérations sont biaisés**, puisqu'ils sont basés sur les mêmes données, et leurs résultats sont donc mécaniquement améliorés ;
- **les hypothèses de distribution sur lesquelles s'appuient ces tests sont invalides**, puisque la sélection de variables conduit à produire des modèles dont l'erreur est davantage corrélée aux régresseurs, en comparaison du modèle initial ;
- **les *p-values* du modèle final n'ont donc plus la même signification** compte tenu des hypothèses violées, et leur correction est un problème difficile ;
- **les intervalles de confiance du modèle final sont faussement étroits**, puisqu'ils sont mécaniquement améliorés par cette procédure qui s'auto-approuve.

Illustrons ce dernier point par une autre pratique courante en économétrie. Après l'estimation d'une régression linéaire, il est souvent de mise d'identifier les variables les plus significatives, puis de ré-exécuter le modèle avec ces seuls régresseurs. Cette technique n'est qu'une autre sélection de variables, similaire à la régression *stepwise*, mais davantage manuelle qu'automatique. Suite à cette opération, il est fréquent d'observer une amélioration des *p-values* et des intervalles de confiance. Cette approche souffre en réalité des mêmes problèmes que la régression *stepwise* : en sélectionnant les variables les plus pertinentes à l'aide de la *p-value*, cela contribue à **améliorer mécaniquement** ce critère dans le modèle élagué. De manière généralle, cette pratique abusive est désignée sous le terme de ***p-hacking***[17].

À ce stade, d'aucuns argumenteraient que la régression *stepwise* n'est affectée par ces problèmes

[17] Aussi appelé *data dredging*, ou triturage de données en français, ce terme désigne l'extrapolation statistique issue de résultats fortuits et d'une exploration exhaustive sur un même échantillon de données.

que lorsque le critère de sélection est basé sur un test statistique, et que l'utilisation de l'AIC, par exemple, en est épargnée. *A contrario*, le test du rapport de vraisemblance, l'AIC et le BIC conduisent en réalité à des sélections similaires ! Considérons en premier lieu le critère de l'AIC. À une itération donnée de la procédure, la différence d'AIC entre les deux modèles successifs s'écrit :

$$[2(k+1) - 2\ln(L_{k+1})] - [2k - 2\ln(L_k)] = 2 - 2\ln\left(\frac{L_{k+1}}{Lk}\right) \tag{25}$$

Le second terme $\frac{L_{k+1}}{Lk}$ est la statistique du rapport du vraisemblance sur laquelle se base la *p-value* du test homonyme. Et le premier terme n'est qu'un seuil arbitraire que la statistique de test doit dépasser pour constituer une amélioration valide au sens de l'AIC. En réalité, la sélection induite par l'AIC revient à calculer la *p-value* modifiée $\mathbb{P}(\chi_1^2 > 2)$, au lieu $\mathbb{P}(\chi_1^2 > 0)$ pour le test du rapport de vraisemblance standard[18]. Par conséquent, ces deux approches sont simplement des conceptions différentes d'une même logique, et elles souffrent donc toutes deux des problèmes évoqués ci-dessus. La conclusion est analogue pour le BIC. Enfin, la mesure de la déviance, elle aussi basée sur la vraisemblance, présente un comportement identique. Rappelons que cette quantité est définie par l'expression :

$$Dev = -2\ln\left(\frac{L_0}{L_{Saturé}}\right) \tag{26}$$

où L_0 correspond à la vraisemblance du modèle considéré, et $L_{Saturé}$ correspond à la vraisemblance du modèle saturé (ou complet), c'est-à-dire du modèle théoriquement parfait qui prédirait correctement toutes les observations. Ce modèle saturé est obtenu en utilisant un paramètre pour chaque observation, afin que les données d'apprentissage soient ajustées parfaitement. Ainsi, la déviance d'un modèle mesure l'écart de celui-ci avec le modèle parfait. Une faible déviance traduirait donc une performance élevée, du moins au sens de la vraisemblance. Avec ces précisions, il est assez aisé de montrer que la statistique du rapport de vraisemblance vaut exactement le rapport des déviances, car dans l'expression 26, la quantité $L_{Saturé}$ s'annule au numérateur et au dénominateur du ratio.

En définitive, il nous parait important de prendre conscience de cette difficulté, et des précautions nécessaires lors de l'analyse des résultats découlant de cette procédure. Avec les aspects computationnels, ces différents écueils statistiques semblent donc constituer une seconde motivation pour l'utilisation d'une technique alternative, issue de l'apprentissage statistique, à savoir le Lasso, qui est présenté en partie 3.3. Toutefois, malgré ces divers obstacles, nous avons conservé l'utilisation de la régression *stepwise* à des fins de validation de la sélection résultant du Lasso, et aussi par souci de cohérence avec les pratiques de marché. Mais cette procédure demeure difficilement applicable en un temps raisonnable au modèle de fréquence, car la base associée correspond à l'ensemble du portefeuille, et non plus aux contrats sinistrés seuls, comme c'est le cas pour la modélisation de la sévérité.

Dans une autre mesure, ce premier sujet a des implications importantes sur une question moins théorique. Précisons d'abord que cette étude s'appuie sur de nombreuses variables multimodales. Celles-ci sont encodées, en amont de l'estimation des modèles économétriques, par la création d'une indicatrice pour chacune des modalités, multipliant ainsi considérablement le nombre de

[18] La loi du χ^2 à 1 degré de liberté est la distribution asymptotique de cette statistique, en raison d'un degré de liberté de différence entre les deux modèles testés.

paramètres du modèle ultérieur. Cette opération est nécessaire pour chacune des variables catégorielles, mais une étape similaire est également suivie pour les variables numériques discrétisées, ou transformées par une *spline*. La prise en compte de ces dernières au sein du modèle se traduit aussi par la multiplication des régresseurs : un par intervalle dans le premier cas, ou par *spline* de base dans le second. Cette multiplication *in fine* des variables explicatives pose le problème de leur prise en compte au sein de la procédure *stepwise*. Par défaut, les fonctions implémentées dans les logiciels statistiques examinent, à chaque itération de la procédure, un facteur de risque dans sa globalité. Plus précisément, elles ne font pas la distinction entre les différentes variables élémentaires (indicatrice uni-modale, ou *spline* de base) composant un facteur de risque donné. Aussi, avec cette approche, un facteur de risque ne peut être ajouté au modèle que de manière unanime, et les différentes modalités ou *splines* d'un même facteur ne peuvent pas être dissociées au cours de la procédure.

Cette démarche est-elle justifiée ? À notre connaissance, le débat sur le sujet n'est pas tranché. D'une part, cette approche permet de favoriser la sélection de variables qui sont pertinentes dans leur globalité, mais pas forcément à travers leurs modalités élémentaires. En effet, il est courant d'observer, au sein d'un modèle économétrique, deux variables peu explicatives individuellement, mais conjointement très significatives. Ce type d'interaction est donc conservé avec une telle démarche. *A contrario*, considérer chaque régresseur élémentaire comme variable indépendante du modèle en tant que telle conduit à délaisser ces variables conjuguées, mais identifie mieux des modalités qui sont significatives lorsqu'elles sont dissociées de leurs sœurs. Ces deux approches ont donc chacune leur propre motivation. Néanmoins, au regard de la discussion précédente, retenir l'une des modalités d'un facteur donné sur la base d'un argument de significativité résulte sans doute en un biais accru. La distinction des *splines* au sein de la régression *stepwise* a-t-elle des conséquences différentes sur les estimateurs de celles des indicatrices uni-modales ? Quoi qu'il en soit, l'importante multiplicité des prédicteurs nous pousse tout de même à privilégier l'approche la plus parcimonieuse qui fait donc le *distinguo* entre les différents régresseurs élémentaires. De la même manière, nous prenons conscience des implications de ce choix sur les résultats obtenus *in fine*.

3.3 Lasso

Les procédures de sélection de variables décrites ci-dessus sont des outils classiques en économétrie, mais sont marquées, comme nous venons de le voir, par un coût computationnel considérable, ainsi qu'une importante volatilité. Dans le cadre de notre étude, le modèle de fréquence souffre profondément de cet aspect puisqu'il s'appuie sur l'ensemble des données disponibles. De plus, la régression *stepwise* présente également des défauts éventuels de pertinence dans son choix de régresseurs, qui sont discutés en partie 3.2.2. Nous avons donc besoin de recourir à des techniques plus efficientes, mais tout aussi efficaces. À cette fin, cette partie présente ci-après les méthodes dites de régularisation, et dont le **Lasso** est l'emblème le plus célèbre.

3.3.1 Méthodes de régularisation

Les **méthodes de régularisation** correspondent à l'ajout d'une pénalité envers la complexité du modèle afin de favoriser la parcimonie. L'objectif principal de ces méthodes est d'améliorer la

robustesse du modèle sous-jacent vis-à-vis de données indépendantes de la base d'apprentissage. Elles rejoignent en ce sens les critères d'information (AIC, BIC) utilisés pour la sélection de variables. La pénalisation se traduit généralement par l'utilisation d'une norme sur les paramètres à estimer, qui est ajoutée à la fonction objectif à minimiser. Il s'agit alors de *shrinkage estimators*. L'utilisation de la norme L_2 correspond à la **régression *ridge*** (régularisation de Tikhonov), celle de la norme L_1 au **Lasso** (*Least Absolute Shrinkage and Selection Operator*). Ce terme supplémentaire va donc contraindre le programme d'optimisation à favoriser le choix de coefficients plus faibles, limitant par la même occasion le caractère particulièrement volatil des coefficients les plus élevés, et contribuant ainsi à la robustesse du modèle. En somme, il s'agit d'augmenter volontairement le biais du modèle, en faveur d'une variance réduite.

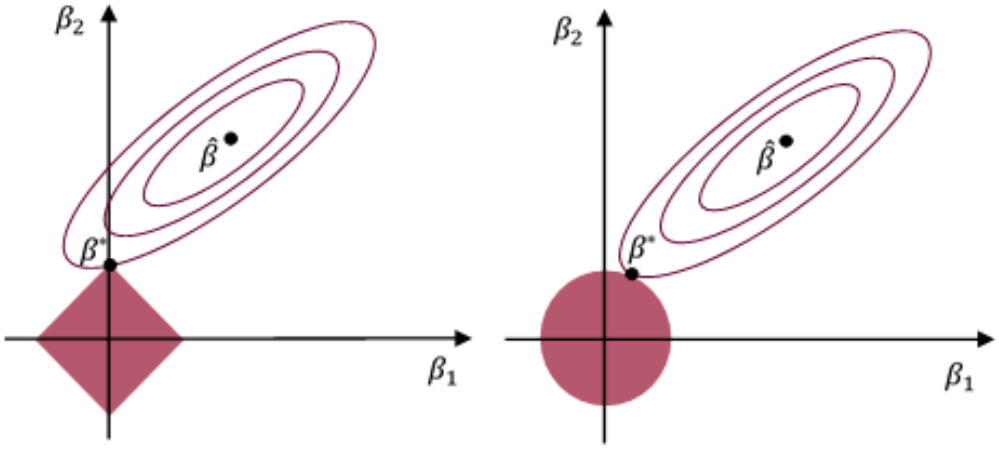
La description du cadre général de ces méthodes requiert de présenter la régularisation dite de l'***elastic net***, combinaison des deux méthodes de pénalisation précédentes. L'estimateur associé dans le cas d'une régression aux moindres carrés est alors défini par la minimisation de la fonction objectif générale R suivante :

$$\hat{\beta} = \arg\min_{\beta} \left[\|Y - X^t\beta\|_{L_2}^2 + \lambda \left((1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2 + \alpha\|\beta\|_{L_1} \right) \right] \tag{27}$$

où α détermine l'équilibre entre les deux pénalités et λ est le coefficient de pénalisation. Les valeurs $\alpha = 0$ et $\alpha = 1$ correspondent respectivement aux cas particuliers de la régression *ridge* et du Lasso. Et le cas $\lambda = 0$ est celui de la régression linéaire classique. Ce coefficient de pénalisation est donc d'une importance cruciale, puisqu'il définit le poids attribué à la pénalité. Plus celui-ci est élevé, plus la pénalisation est forte, et les coefficients estimés sont alors plus proches de zéro. Le calibrage de ce paramètre est par conséquent capital. Cela est généralement effectué à l'aide d'une technique de validation croisée, dont l'application au Lasso est décrite plus loin (partie 3.3.3). De même, l'équilibre α entre les deux types de pénalité est susceptible d'être calibré par une méthode similaire. Toutefois, cette pénalisation complexe se révèle généralement excessive, et il suffit souvent de se restreindre à l'une des deux méthodes élémentaires la composant.

Les deux pénalisations introduites ici présentent des effets relativement opposés vis-à-vis des variables corrélées, en particulier en grande dimension, lorsque de nombreux prédicteurs sont susceptibles d'être fortement corrélés entre eux. Alors que la régression *ridge* est connue pour **réduire de façon simultanée** la valeur des coefficients de variables corrélées, le Lasso est plutôt indifférent aux phénomènes de corrélation et a tendance à **retenir arbitrairement l'un d'entre eux et écarter les autres**. La manifestation de ces deux phénomènes contraires peut s'intuiter à travers la figure 15, qui représente graphiquement l'exemple d'une régression avec deux variables. Les deux problèmes de d'optimisation exposés ici y sont illustrés en dimension 2. S'y retrouvent les courbes d'équi-erreur de la régression non pénalisée, ainsi que les zones de contrainte imposées par les pénalités. L'intersection de ces deux surfaces correspond alors à la solution graphique du problème de régularisation considéré. Le Lasso favorise visiblement des solutions sur des axes de l'espace, contraignant ainsi les coefficients relatifs à ces variables à zéro.

FIGURE 15 – Programmes d'optimisation du Lasso (à gauche) et de la régression *ridge* (à droite) en dimension 2



Les ellipses correspondent aux courbes d'équi-erreur $||Y_i - \hat{Y}_i||$ de la régression non pénalisée, dont le minimum est atteint par l'estimateur standard $\hat{\beta}$. Les formes pleines représentent les zones de contrainte sur les coefficients β_1 et β_2 imposées par les pénalités, définies par $|\beta_1| + |\beta_2| \leq t$ pour le Lasso et $\beta_1^2 + \beta_2^2 \leq t^2$ pour la régression *ridge*. L'intersection de ces deux surfaces constitue la solution graphique β^* de la régression pénalisée. Le Lasso conduit naturellement à favoriser des coefficients nuls, une solution dite *sparse* (clairsemée), car les sommets du cercle unité pour la norme 1 apparaissent sur les axes du plan. En l'occurrence, la solution du Lasso est ici nulle selon sa première coordonnée : $\beta_1^* = 0$.

D'aucuns pourraient alors défendre l'*elastic net*, argumentant qu'il offre ainsi un bon compromis entre ces deux approches, la première favorisant la sélection de nombreux régresseurs potentiellement pertinents, dans une situation sans *a priori* particulier, et la seconde écrémant sensiblement l'ensemble des variables disponibles, ce qui est profitable en grande dimension. Cependant, le grand nombre de variables auxquelles nous sommes confrontés dans cette étude nous contraint d'employer une vraie technique de réduction de la dimension. Et seul le Lasso nous permet d'écarter véritablement une partie des facteurs de risque utilisés. C'est donc cette méthode qui sera exclusivement mise en œuvre dès lors que les procédures *stepwise* se révèlent inadaptées.

3.3.2 Résolution numérique

Discutons désormais brièvement de l'implémentation concrète de ces techniques. Il faut d'abord remarquer que le programme d'optimisation permettant l'estimation des paramètres d'intérêt, décrit par l'équation 27, n'admet pas de solution analytique par formule fermée comme en régression linaire classique, en raison de la présence de la norme L_1, non inversible[19]. La résolution se fait donc par un algorithme de descente de gradient. En effet, le problème d'optimisation se décline selon chaque direction β_j par l'équation impliquant la dérivée partielle de la fonction objectif R :

$$\left.\frac{\partial R}{\partial \beta_j}\right|_{\beta=\tilde{\beta}} = -\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - x_i^t\tilde{\beta}) + \lambda(1-\alpha)\beta_j + \lambda\alpha = 0 \tag{28}$$

où $\tilde{\beta}$ est l'estimateur de l'itération précédente, fixé dans la dérivation précédente pour toutes les coordonnées $k \neq j$.

[19]Notons que la régression *ridge* admet une solution analytique, grâce au caractère quadratique de la norme L_2

Dans le cas d'une régression GLM, le terme de pénalité s'ajoute à la log-vraisemblance du modèle. Pour la modélisation de la fréquence, par exemple, le programme de maximisation de l'équation 7 devient alors :

$$\begin{aligned} &\max_{\beta\in\mathbb{R}^p}\left[\sum_{i=1}^{n}\{y_i\ln(\lambda_i)-\ln(y_i!)-\lambda_i\}-\lambda\left\{(1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2+\alpha\|\beta\|_{l_1}\right\}\right] \\ \Leftrightarrow &\max_{\beta\in\mathbb{R}^p}\left[\sum_{i=1}^{n}\left\{y_i x_i^t\beta-\ln(y_i!)-\mathrm{e}^{x_i^t\beta}\right\}-\lambda\left\{(1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2+\alpha\|\beta\|_{l_1}\right\}\right] \end{aligned} \tag{29}$$

La méthode de résolution numérique correspond alors à une variante de l'algorithme de Newton-Raphson, homologue de la descente de gradient pour les vraisemblances. Le lecteur pourra se référer aux articles de Friedman (2007 [14] et 2010 [15]) pour la description détaillée de cette procédure. À ce stade, il convient de remarquer que les hypothèses habituelles sous-tendant l'estimation par maximum de vraisemblance ne sont plus vérifiées. Aucune analogie n'est véritablement possible entre la vraisemblance pénalisée ici considérée, et la quantité traditionnellement maximisée dans le cadre des GLM. En particulier, les coefficients estimés par le Lasso sont *de facto* **biaisés**, ce qui est d'ailleurs logique puisqu'une source de biais est volontairement introduite au sein de la fonction de perte. De même, les erreurs standards ainsi que les intervalles de confiance usuellement calculés n'ont plus aucun sens dans cette situation, car ils sont conditionnés par une hypothèse de distribution asymptotiquement gaussienne des estimateurs, qui est ici violée. Aussi, la comparaison des résultats entre modèles classiques et modèles pénalisées doit être opérée avec la plus grande précaution, en gardant bien à l'esprit les divergences de signification précédentes.

3.3.3 Calibrage du méta-paramètre

Comme indiqué précédemment, le Lasso comporte un méta-paramètre capital, qui traduit le poids accordé à la pénalisation, et détermine ainsi l'équilibre biais–variance du modèle construit. La valeur optimale de λ est généralement déterminée par validation croisée, dont le principe a été exposé en partie 3.1. Dans cette étude, l'approche suivie correspond à l'examen d'une sélection de 50 valeurs candidates, comparées par la méthode 5-*fold*, à l'aide de la mesure de la déviance. Les 50 valeurs candidates n'ont cependant pas été proposées de manière arbitraire. L'algorithme proposé par Friedman (2010, [15]), et qui est ici employé, exploite une stratégie de mise à jour du méta-paramètre qui est computationnellement efficiente. Celle-ci part d'une valeur initiale $\lambda_{\max}$, qui est la plus petite valeur pour laquelle tous les coefficients estimés sont nuls, détermine ensuite une valeur finale $\lambda_{\min} = \varepsilon \cdot \lambda_{\max}$ (typiquement $\varepsilon = 0,001$), puis construit une suite de valeurs décroissantes entre ces deux bornes sur l'échelle logarithmique : $\lambda_{\max} > \lambda_2 > \lambda_3 > \cdots > \lambda_{49} > \lambda_{\min}$

L'erreur de validation croisée est ensuite calculée pour chacune de ces 50 valeurs candidates. Cette erreur est estimée par la moyenne des déviances sur les 5 partitions du 5-*fold* et, par conséquent, cet estimateur est incertain. L'approche standard consiste à retenir la valeur du méta-paramètre pour laquelle l'erreur est minimale, noté $\lambda_{Dev}^{\min}$. Toutefois, le modèle issu de cette approche apparaît parfois trop complexe et a tendance à sur-apprendre. Face à ce phénomène, une alternative consiste à sélectionner la valeur du méta-paramètre la plus grande, et pour laquelle l'erreur est majorée par l'erreur standard de $\lambda_{Dev}^{\min}$. Cette valeur est notée λ_{Dev}^{1se}. Autrement dit,

cela revient à choisir le modèle le plus simple, mais qui ne peut pas être distingué du meilleur modèle en termes d'erreur, étant donnée l'incertitude de l'estimation de cette erreur. Les expressions mathématiques de ces deux choix sont données ci-après.

$$\lambda_{Dev}^{\min} = \arg\min_{\lambda_j} \frac{1}{k} \sum_{i=1}^{k} Dev_i(\lambda_j) \tag{30}$$

$$\lambda_{Dev}^{1se} = \max_{\lambda_j} \left\{ \lambda_j : \left(\frac{1}{k} \sum_{i=1}^{k} Dev_i(\lambda_j) \right) \leq 1 \cdot se(\lambda_{Dev}^{\min}) \right\} \tag{31}$$

3.4 Forêt aléatoire

Les deux procédures de sélection de variables présentées précédemment peuvent paraître profondément différentes, mais elles ne sont pas, en réalité, extrêmement éloignées. L'estimateur du Lasso est ici déterminé par la méthode de descente de gradient. Toutefois, un autre algorithme populaire permet aussi de résoudre la programme du Lasso : le *least angle regression*[20] (LARS). Celui-ci présente une connexion très forte avec le *stepwise forward*, et produit des solutions très proches de celles du gradient. En somme, le Lasso et le *stepwise* sont tous deux des **procédures itératives**. Face à ces approches plutôt conventionnelles, une troisième méthode, **à caractère récursif**, la forêt aléatoire, peut aussi être employée en tant que technique de sélection de variable en soi. En effet, la structure hiérarchique des arbres permet de classer les variables en jeu dans la croissance de la forêt par ordre d'importance. Et cette mesure d'importance peut tout à fait être exploitée comme critère de sélection des variables les plus pertinentes. Nous présentons ci-après les outils nécessaires au calcul de cette mesure.

3.4.1 Erreur *Out-of-bag*

Introduisons dans un premier lieu les erreurs dites *out-of-bag*, issues des forêts aléatoires. Nous avons vu précédemment que l'erreur de validation croisée est particulièrement efficace pour mesurer de manière robuste la performance prédictive d'un modèle. La pertinence de cette nouvelle mesure d'erreur provient de l'indépendance entre les différentes sous-bases considérées. Or, ce caractère d'indépendance se retrouve également au sein de la forêt aléatoire, entre les différents arbres bâtis, à travers notamment le *bootstrap* de la base initiale. Il est donc naturel de considérer une mesure analogue à celle issue de la validation croisée, mais basée ici sur une partition apprentissage/test de la forêt F elle-même. Pour ce faire, il faut d'abord considérer, pour chaque observation (x_i, y_i) de la base d'apprentissage, la sous-forêt F_i des arbres basés sur un échantillon de la base ne contenant pas cette observation (x_i, y_i). Autrement dit, cette sous-forêt F_i n'exploite pas l'information fournie par l'observation (x_i, y_i). Le modèle *out-of-bag* se définit ensuite par l'association, à chaque observation (x_i, y_i), de la sortie prédite par la sous-forêt F_i, à savoir la moyenne des prédictions des arbres la composant. Chaque (x_i, y_i) représente donc une observation de test pour ce modèle. L'erreur *out-of-bag* correspond alors à l'erreur de prédiction moyenne du modèle *out-of-bag* sur la base initiale. Breiman (2001 [7]) présente sur ce sujet des preuves empiriques qui montrent que l'erreur *out-of-bag* est aussi précise que l'utilisation d'une base de test de même taille que la base d'apprentissage.

[20] Cet alternative à la descente de gradient est toutefois plus coûteuse en temps de calcul (Efron et al., 2004 [13] ; The Lasso Page [32]

3.4.2 Mesure d'importance

Les outils précédents permettent désormais de calculer une mesure d'importance des variables explicatives au sein de la forêt aléatoire. Il s'agit ici de déterminer l'importance de chacun des prédicteurs au sens de leur contribution marginale à la construction de la forêt et donc, *a fortiori*, à la réduction de l'hétérogénéité liée au modèle. La mesure d'importance utilisée dans cette étude est inspirée du test statistique de permutation. Plus précisément, l'importance d'un régresseur correspond à l'augmentation marginale de l'erreur *out-of-bag* due à la permutation des valeurs des observations selon cette variable, moyennée sur l'ensemble des arbres. Autrement dit, si une variable n'est pas importante (hypothèse nulle du test), alors le ré-arrangement des valeurs de cette variable ne dégradera pas la précision des prédictions.

4 Portefeuille d'étude

Cette partie présente les données d'étude sur lesquelles est basée la mise en œuvre pratique de la méthodologie développée précédemment. L'ensemble des traitements statistiques préliminaires, indispensables à toute modélisation ultérieure, sont également décrits ici : discrétisation des variables numériques, écrêtement des valeurs extrêmes, etc.

Afin de mettre en œuvre les démarches proposées précédemment, nous avons à notre disposition deux bases de données de taille satisfaisante, recensant l'ensemble des contrats et des sinistres relatifs à un portefeuille conséquent, sur un large périmètre, et un horizon étendu. Ces données doivent faire l'objet de divers retraitements statistiques nécessaires à leur bonne exploitation au sein des modèles ultérieurs : présélection des facteurs de risque, encodage des variables qualitatives, écrêtement des valeurs extrêmes, etc. Après la présentation du portefeuille d'étude, cette partie décrit les différentes opérations préliminaires réalisées sur les données, ainsi que les résultats de diverses analyses descriptives visant à se familiariser avec leur structure interne.

4.1 Présentation des bases de données

Les bases d'étude à disposition (base police et base sinistre) correspondent à l'historique de 2009 à 2013 du portefeuille automobile d'un assureur de taille moyenne. Ces deux bases fournissent respectivement de nombreuses informations relatives aux clients et aux sinistres afférents à tous les contrats présents dans le portefeuille au cours de la période d'observation. De nombreuses garanties y sont représentées.

Ces données représentent près d'un million de lignes pour la partie portefeuille, chacune correspondant à une image contrat, c'est-à-dire à l'historique d'une police entre deux modifications contractuelles, soit suite à un sinistre, soit à la demande du client. Avec la donnée de l'exposition de chaque image contrat au sein du portefeuille, l'ampleur de l'historique à disposition est estimée à environ un demi-million d'années de contrat, pour près de 400 000 contrats distincts. Chaque image contrat sera considérée comme une observation à part entière dans nos modélisations ultérieures, que ce soit pour la problématique du coût tout comme pour celle de la fréquence.

La base des sinistres contient les caractéristiques de moins de 200 000 sinistres, survenus à plus de 100 000 assurés. Cette base va nous permettre de construire les deux variables d'intérêt pour nos travaux : la sévérité de chaque sinistre d'une part, et la fréquence des sinistres de chaque assuré d'autre part. Notons par ailleurs que les sinistres recensés dans cette base correspondent à diverses garanties. Dans le cadre de notre étude, nous allons nous intéresser aux garanties susceptibles de couvrir les sinistres les plus coûteux :

- la Responsabilité Civile (RC) Matérielle ;
- la Responsabilité Civile (RC) Corporelle.

La base des polices contient 150 variables fournissant des caractéristiques diverses, essentiellement liées au véhicule assuré, mais également relatives au client ou au contrat. À ces variables s'ajoute une quarantaine de caractéristiques des sinistres survenus au cours de la période d'étude, qui ne pourront malheureusement pas être exploitées dans le cadre de ce mémoire à caractère prédictif.

Précisons enfin que le caractère confidentiel des données exploitées nous interdit de détailler outre mesure les caractéristiques des bases mais que nous nous attacherons à conserver la portée statistique des résultats obtenus lors de ces travaux.

4.2 Gestion des variables explicatives

Tout d'abord, les variables explicatives doivent faire l'objet de divers retraitements statistiques préalables à la phase de modélisation : présélection des facteurs de risque, encodage des variables qualitatives, conversion des formats de régresseurs clés. Nous décrivons ci-après notre démarche vis-à-vis de ces différents enjeux.

4.2.1 Présélection des facteurs de risque

Dans un premier temps, face au grand nombre de facteurs de risque en présence, nous écartons empiriquement les nombreuses variables qui ne nous paraissent pas apporter de pouvoir explicatif substantiel. Cette liste d'exclusion comprend, entre autres, des caractéristiques du véhicule trop spécifiques, ou redondantes avec des critères plus généraux[21], et des variables présentant beaucoup de valeurs manquantes. Le lecteur remarquera également que la variable de sexe n'est pas conservée du fait de la réglementation en matière de segmentation tarifaire qui interdit d'utiliser ce critère comme différenciation du risque. Ces omissions résultent en la présélection d'une vingtaine de variables, présentées dans le tableau 3. Celles-ci comprennent une grande diversité de caractéristiques relatives à l'assuré, à la police, et au véhicule.

Table 3 – Variables présélectionnées

Catégorie	Variables
Assuré	• Âge • Ancienneté du permis • Statut marital • Code CSP • Nombre de sinistres antérieurs • Coefficient Réduction Majoration (CRM) • CRM précédent
Police	• Formule produit • Origine du contrat • Canal de distribution • Enfant d'assuré • Conduite accompagnée
Véhicule	• Ancienneté du véhicule • Classe SRA • Groupe SRA • Réparation SRA • Kilométrage • Mode d'acquisition • Usage

[21] I est ici fait allusion aux variables `Classe SRA` et `Groupe SRA` qui caractérisent respectivement la classe de prix et la dangerosité intrinsèque du véhicule et procurent à ce titre une information bien plus pertinente que les variables plus spécifiques, comme les caractéristiques du moteur, de la carrosserie ou encore des suspensions. Nous reviendrons sur ces variables plus loin (partie 4.2.3).

Le Coefficient Réduction Majoration (CRM) correspond à l'indicateur de bonus-malus, qui évolue en fonction de l'historique personnel de sinistralité de l'assuré. Ce coefficient est appliqué multiplicativement à la prime d'assurance, afin d'offrir à l'assuré une réduction ou majoration de celle-ci selon son comportement. Un assuré sans historique débute normalement avec un CRM de 1. En cas de bon historique, son CRM peut baisser jusqu'à 0,5, divisant ainsi par deux sa prime d'assurance. Dans le cas contraire, il peut augmenter jusqu'à 3.

En dépit de notre présélection, les valeurs manquantes se sont révélées assez nombreuses pour certaines des variables retenues, notamment le statut marital et l'ancienneté du permis. Dans nos modèles, nous avons choisi d'écarter les observations présentant au moins une valeur manquante. Cette simplification peut paraître abusive, mais elle est acceptable en rappelant que la base d'étude contient environ 1 million d'observations. De plus, il n'y a pas de raison de penser qu'il y ait un réel biais de sélection vis-à-vis de ces valeurs manquantes. Affiner les modèles construits nécessiterait quoi qu'il en soit de conserver ces observations écartées et de trouver une méthode susceptible de les intégrer au sein de l'estimation des modèles.

4.2.2 Variables qualitatives : modalités de référence

Afin d'exploiter les variables qualitatives non ordinales dans des modèles économétriques, celles-ci sont préalablement binarisées par la construction d'une indicatrice pour $(m-1)$ modalités, la dernière modalité servant de **modalité de référence** à partir de laquelle sont interprétés les effets marginaux. Le tableau 4 présente la liste des modalités de l'ensemble des variables catégorielles, et la modalité de référence est précisée en caractères gras. Par défaut, celle-ci est la première encodée lors de la conversion en variable qualitative.

4.2.3 Caractéristiques SRA

Comme évoqué précédemment, les caractéristiques SRA (Sécurité et Réparation Automobiles) ont été spécialement créées pour la tarification assurantielle et synthétisent de nombreux facteurs de risque liés au véhicule. Trois indicateurs permettent la classification de tout véhicule :

- Le **Groupe SRA**, qui représente la dangerosité intrinsèque du véhicule, contribue à la détermination des tarifs des garanties Responsabilité Civile, et est compris entre les valeurs 20 à 50
- La **Classe de prix SRA**, qui représente la valeur à neuf du véhicule, permet l'estimation du coût de sinistres matériels en cas de perte totale, et est compris entre les valeurs A à V, plus HC (Hors Classe)
- La **Classe de Réparation SRA**, qui représente le coût à la réparation des pièces détachées du véhicule, convient pour les autres types de sinistres matériels, et est compris entre les valeurs A à ZE, plus HC (Hors Classe)

Ces caractéristiques ordinales sont des variables clés dans un processus tarifaire et requièrent à ce titre un soin particulier. En outre, afin de limiter le nombre de paramètres du modèle de tarification, il est de mise de convertir les deux d'entre elles qui sont qualitatives en variables numériques. Cette conversion peut correspondre à l'attribution classique d'un rang à chaque modalité selon l'ordre alphabétique sous-jacent. Il en résulte des scores quantitatifs du risque associé à chaque véhicule. Cependant, la conversion de l'éventail de modalités en une échelle numérique

TABLE 4 – Modalités des variables catégorielles

Variable	Modalités
Statut marital	• **Marié(e)** • Célibataire • Inconnu • Concubin(e) • Divorcé(e) • PACS • Séparé(e) • Veuf(ve)
Code CSP	• **Salarié** • Retraité • Fonctionnaire et assimilé • ...
Formule produit	• **Formule 1** • Formule 2 • Formule 3 • Formule 4
Origine du contrat	• **Migration** • Natif
Canal de distribution	• **Agent** • Courtier • Autre
Enfant d'assuré	• **Non** • Oui
Conduite accompagnée	• **Non** • Oui
Mode d'acquisition	• **Comptant** • Crédit • Nantissement • Crédit bail • Autre
Usage	• **Tous usages** • Tournées/Livraison • Garage mort • Taxi • Transport Public de Marchandises (TPM) • Ambulance-VSL (Véhicule Sanitaire Léger)

arbitraire implique que la différence entre deux modalités successives vaut systématiquement 1. Ceci peut paraître abusif à l'égard d'un modèle économétrique, qui vise quantifier l'impact sur la sinistralité d'une variation élémentaire selon, par exemple, les variables SRA. Or, il n'y a pas de raison particulière de penser que le passage de la Classe A à la Classe B soit équivalent au passage de la Classe U à la Classe V en termes d'effet sur la sinistralité. Et c'est pourtant l'hypothèse qui est admise avec cette conversion numérique. Toutefois, cet écueil n'en est pas un si nous souhaitons par la suite, comme c'est le cas dans notre approche, identifier puis intégrer au sein de la modélisation les effets de seuils/ruptures selon chacune des variables explicatives sélectionnées. Cette démarche, qui est détaillée dans notre stratégie économétrique, est censée prendre en compte les éventuelles variations brutales de la sinistralité selon l'évolution du facteur de risque considéré. Cela permettra ainsi d'effacer en partie l'effet de normalisation induit par la conversion numérique

des variables SRA.

4.3 Analyse des variables expliquées

Après cette première phase de prise en main des variables explicatives, nous nous intéressons aux deux variables d'intérêt : fréquence et sévérité des sinistres. Après la description de la méthode d'écrêtement des valeurs extrêmes, nous réalisons quelques analyses descriptives permettant de de mieux appréhender le phénomène de sinistralité automobile, en amont de la modélisation.

4.3.1 Écrêtement des valeurs extrêmes

Comme nous l'avons mentionné dans notre revue de littérature, les sinistres des garanties Responsabilité Civile présentent généralement une distribution à queue épaisse. La présence de sinistres « graves », en particulier en RC Corporelle, peut nuire à la qualité des modèles élaborés. Il convient donc d'écrêter ces sinistres extrêmes à l'aide de la théorie des valeurs extrêmes. De nombreuses méthodes permettant de déterminer le seuil d'écrêtement idéal sont disponibles dans la littérature. Comme ce sujet ne représente pas le cœur d'étude de ce mémoire, nous allons nous limiter à l'utilisation de deux approches, qui se démarquent par leur simplicité et leur efficacité.

Dans un premier temps, la théorie des valeurs extrêmes indique que la distribution conditionnelle des excès résiduels, c'est-à-dire des observations « au-delà d'un seuil », se comporte asymptotiquement comme une loi de Pareto Généralisée, dont la fonction de répartition s'écrit :

$$H_{\xi,\mu,\sigma}(x) = \begin{cases} 1 - \left(1 + \dfrac{\xi(c-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & \text{pour } \xi \neq 0 \\ 1 - \exp\left(-\dfrac{x-\mu}{\sigma}\right) & \text{pour } \xi = 0 \end{cases} \tag{32}$$

avec $x \geqslant \mu$ quand $\xi \geqslant 0$, et $\mu \leqslant x \leqslant \mu - \frac{\sigma}{\xi}$ quand $\xi < 0$, où $\mu \in \mathbb{R}$, $\sigma > 0$, et $\xi \in \mathbb{R}$. Supposons dans la suite $\mu = 0$ sans perte de généralités puisqu'il ne s'agit que d'un paramètre de position.

Plus précisément le théorème de Pickands [22] ci-après affirme, sous certaines conditions, que la distribution conditionnelle des excès résiduels, définie formellement par l'expression $F_u(x) = \mathbb{P}(X - u \leqslant x | X > u) = 1 - \frac{S(x+u)}{S(u)}$, où S désigne la fonction de survie de la distribution, converge vers une loi GPD lorsque la taille de l'échantillon, et donc le seuil, tendent vers l'infini.

Théorème. *Une distribution F appartient au domaine d'attraction d'une loi généralisée des valeurs extrêmes (GEV) désignée $G_{\xi,\sigma}$ si et seulement si, pour $u < x_+$, où x_+ est le point extremum droit du support de F, formellement $x_+ = \sup\{x \in \mathbb{R} : F(x) < 1\}$, il existe une constante de normalisation positive $a(u)$ telle que :*

$$\lim_{u \uparrow x_+} F_u\left(a(u)x\right) = H_{\xi,\sigma}(x) \tag{33}$$

Ce théorème fait donc le lien entre l'estimation de la distribution du maximum des observations via les lois GEV, et l'estimation de la distribution conditionnelle des excès résiduels via les lois GPD. Il en découle la possibilité d'approcher cette distribution conditionnelle par une GPD estimée par maximum de vraisemblance.

Dans ce contexte, il convient de calculer les estimateurs des deux paramètres d'une GPD ajustée sur les montants des sinistres supérieurs à un seuil d'écrêtement u et ce, pour un large éventail de seuils possibles. Les estimateurs obtenus sont représentés graphiquement en fonction du seuil considéré en figure 16. Il s'agit ainsi de déterminer à partir de quel seuil l'estimation des excès résiduels par une GPD apparaît graphiquement pertinente. Le choix du seuil idéal est le résultat d'un compromis entre la variance du modèle, qui diminue lorsque le seuil est bas car davantage d'observations sont ainsi retenues pour estimer les paramètres, et son biais, qui diminue lorsque le seuil est élevé car cela permet de se rapprocher des hypothèses asymptotiques sous-jacentes. Avec ces précisions, une approche convenable consiste à choisir le seuil le plus bas qui fournit les mêmes estimations que tous les seuils au-dessus de celui-ci. Aussi, il est possible d'identifier sur cette figure un comportement relativement stationnaire des estimateurs au-delà du seuil de 78 000, indiqué par les pointillés.

FIGURE 16 – Estimateurs de la GPD selon le seuil considéré (RC Corporelle)

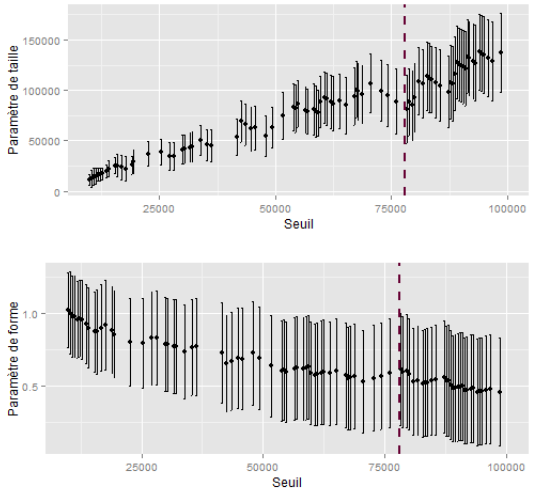
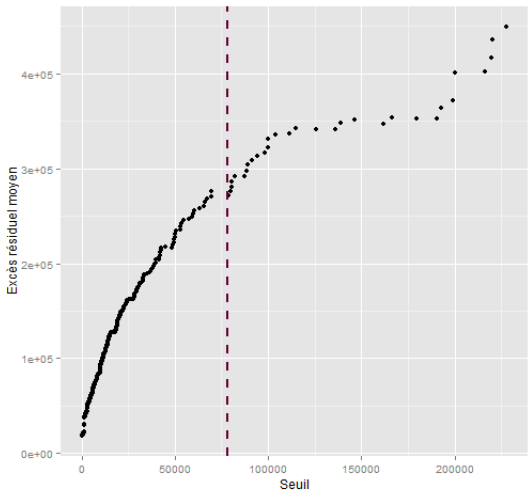


FIGURE 17 – Courbe de l'excès résiduel moyen en fonction du seuil considéré (RC Corporelle



Afin de corroborer ce choix, il est possible de recourir à une second méthode graphique pour déterminer le seuil d'écrêtement idéal. La figure 17 représente la courbe de l'**excès résiduel moyen** en fonction du seuil considéré, c'est-à-dire la courbe de l'espérance de la distribution conditionnelle évoquée plus haut, quantité qui peut s'écrire sous la forme $\mathbb{E}\left[X-u|X>u\right]$. Avec le cadre précédent, il est établi que la variable aléatoire $X_u = [X-u|X>u]$ suit une loi de Pareto généralisée.

Par ailleurs, il est possible de montrer que si X est une variable aléatoire distribuée selon une loi GPD de paramètres (ξ, σ), alors la variable aléatoire $Y_v = [X-v|X>v]$ jouit d'une propriété de stabilité :

$$Y_v \sim \text{GPD}(\xi, \sigma + \xi v) \tag{34}$$

Et il vient que si $\xi < 1$, alors pour tout $v < x_+$,

$$\mathbb{E}\left[X-v|X>v\right] = \frac{\sigma+\xi v}{1-\xi} \tag{35}$$

En appliquant cette propriété à la variable X_u définie plus haut qui suit bien une loi de

Pareto généralisée de paramètres (ξ, σ), alors pour un $x \geqslant 0$, la variable $Y_v = [X_u - v | X_u > v] = [X - (u+v) | X > u + v]$ suit une loi de Pareto généralisée $(\xi, \sigma + \xi v)$ et :

$$\mathbb{E}\left[X - (u+v) | X > (u+v)\right] = \frac{\sigma + \xi v}{1 - \xi} \tag{36}$$

Cela permet alors de constater que l'espérance conditionnelle des excès résiduels est une fonction affine du surplus v lorsque celui-ci est positif et, *a fortiori*, du seuil $u + v$. Ce résultat fournit donc une alternative permettant d'identifier empiriquement le seuil idéal. Il s'agit ainsi de déterminer le seuil à partir duquel la courbe de l'espérance conditionnelle se comporte de manière affine. En traçant le seuil précédent de 78 000 sur ce graphe, celui-ci apparaît relativement cohérent avec cette seconde approche. Ceci permet donc de maintenir le premier choix et d'écarter l'ensemble des sinistres dépassant ce seuil. Cela représente 58 observations, toutes afférentes à la garantie RC Corporelle.

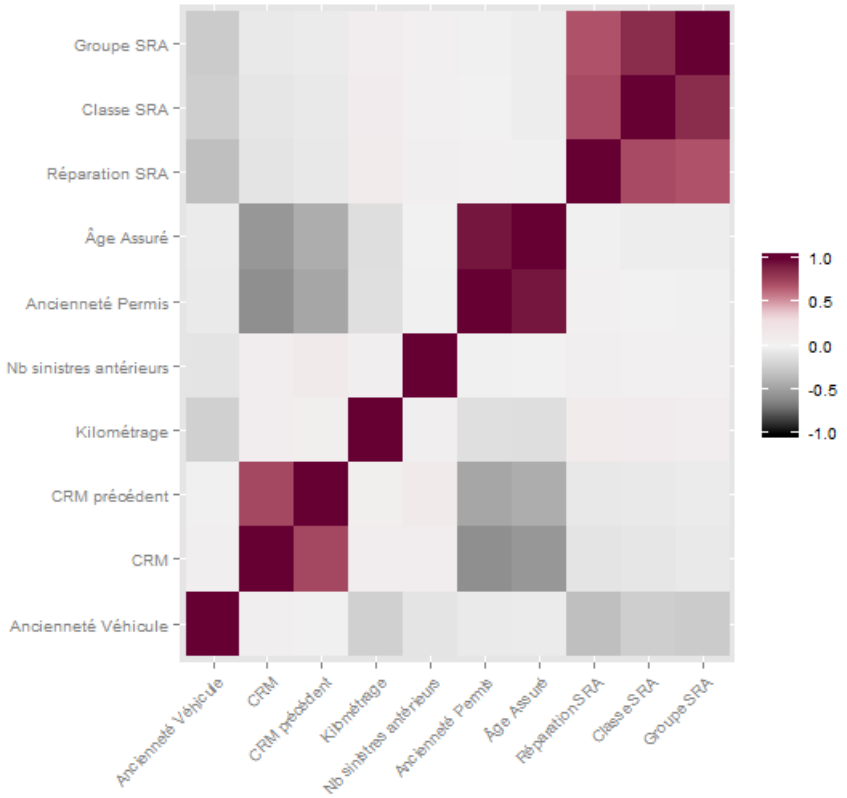
Nous avons réalisé une étude similaire pour la garantie RC Matérielle, mais celle-ci ne présentait pas de comportement susceptible d'être modélisé par la théorie des valeurs extrêmes : aucun seuil pertinent d'écrêtement n'était identifiable. Ceci s'explique par l'absence de sinistres particulièrement graves pour cette garantie. L'ensemble des sinistres relatifs à cette garantie ont donc été conservés pour la suite des travaux.

4.3.2 Corrélations entre variables explicatives

Après ces différents retraitements statistiques, étudions désormais la structure sous-jacente des données. Les modèles économétriques qui seront bâtis par la suite sont notamment connus pour performer de manière médiocre en présence de variables fortement corrélées. En effet, dans une telle situation, ces modèles quantifient mal les effets individuels propres à chacune des variables. Les coefficients estimés sont alors plus incertains, et les erreurs standards associées augmentent. Afin d'éviter cette perte de précision, il est recommandé d'identifier et de traiter ces éventuels problèmes de colinéarité. L'approche classique consiste à examiner la matrice de corrélations des prédicteurs. Cependant, celle-ci n'est calculable que pour les variables numériques. Déterminons-la pour les facteurs de risque à caractère quantitatif de notre présélection[22]. Une façon de représenter graphiquement cette matrice est de recourir à une thermocarte (*heatmap*) qui produit une visualisation en dégradés de couleurs, affichée ci-dessous (figure 18).

[22] Nous utilisons donc, pour cette analyse, les variables SRA numérisées.

FIGURE 18 – *Heatmap* des corrélations entre variables numériques



Trois blocs de corrélations importants apparaissent, mais ils semblent relativement intuitifs :

- les trois variables SRA sont fortement corrélées deux à deux ;
- l'âge de l'assuré et l'ancienneté du permis présentent une corrélation quasi-totale ;
- le CRM actuel et le CRM de l'année précédente sont aussi très corrélés entre eux.

Par ailleurs, nous observons également une corrélation négative entre les variables du second bloc d'une part, et celles du dernier bloc d'autre part. Ce phénomène n'est pas étonnant : le CRM a tendance a baisser avec l'âge, car la majorité des conducteurs ne présentent plus, avec le temps, de comportement à risque. En définitive, outre ces quelques blocs minoritaires, aucun phénomène de colinéarité généralisée n'est identifiable.

Afin de compléter cette première analyse, nous souhaitons intégrer à un graphique de ce type les variables qualitatives. Pour ce faire, il est possible de discrétiser l'ensemble des variables quantitatives[23], de les associer aux variables qualitatives, puis de mesurer leur inter-dépendance deux à deux à travers la statistique de test du χ^2. Cette statistique de test, basée sur le tableau de contingence des deux variables considérées, est définie par l'expression mathématique suivante :

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)}{\frac{n_{i.}n_{.j}}{n}} \tag{37}$$

où n_{ij} est le nombre d'observations présentant la $i^{ème}$ valeur pour la première variable et la $j^{ème}$

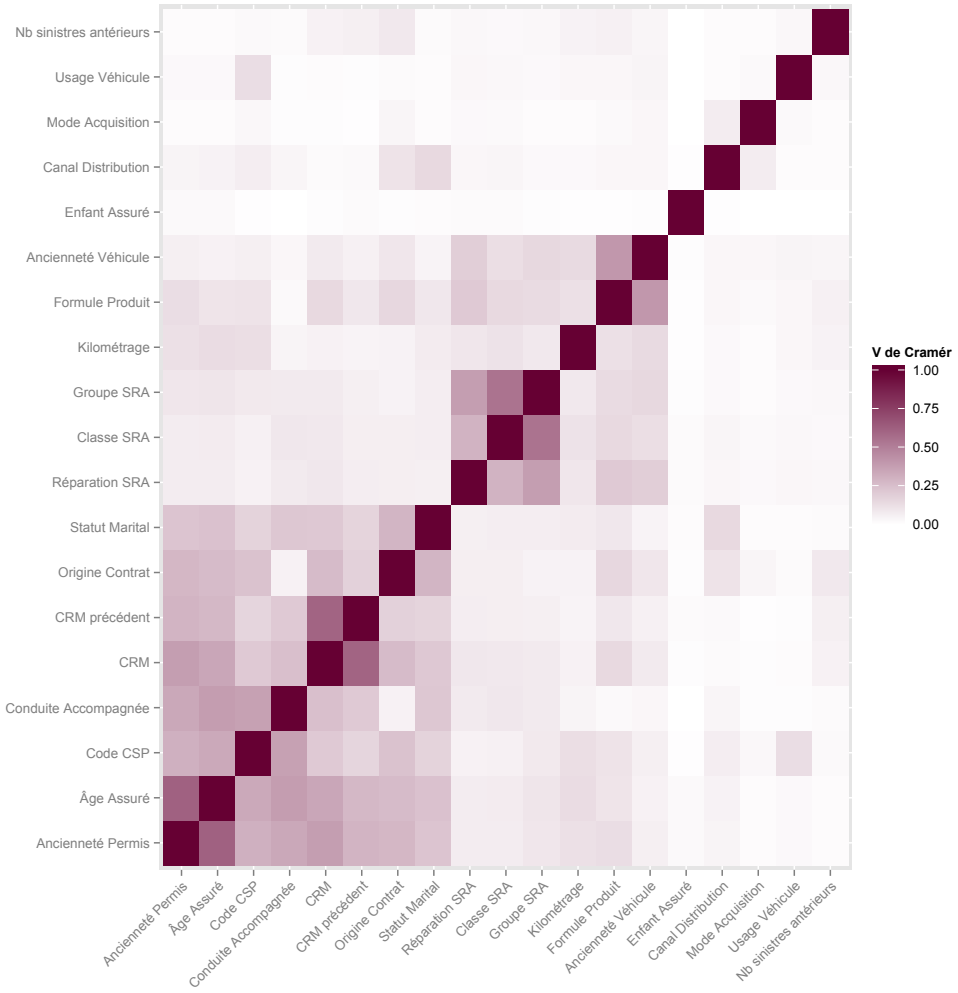
[23] Pour cette seconde analyse, nous conservons donc les variables SRA, initialement qualitatives.

valeur pour la seconde variable, et avec $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$, et enfin $n = \sum_{ij} n_{ij}$ est le nombre total d'observations. Cependant, cette quantité est dépendante du nombre d'observations, et une mesure absolue de l'intensité de l'association entre les deux variables, telle que le **V de Cramér**, est alors davantage pertinente. Celui-ci normalise le χ^2 par le χ^2 maximal théorique, qui correspondrait à un tableau de contingence comportant une seule valeur non nulle par ligne et par colonne :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n \cdot \min(l-1, c-1)}} \tag{38}$$

où l est le nombre de lignes et c le nombre de colonnes du tableau de contingence. La *heatmap* correspondante est présentée ci-après.

FIGURE 19 – *Heatmap* du V de Cramér



Sur ce graphique, deux grands blocs de corrélation se démarquent :

- un premier bloc constitué des variables d'âge et d'ancienneté du permis, du code CSP, des variables CRM, de l'origine du contrat et du statut marital ;
- un second bloc constitué, d'une part, des variables SRA, et d'autre part, de la formule produit, du kilométrage et de l'ancienneté du véhicule.

Le premier bloc comprend des variables plutôt liées à l'assuré, alors que le second correspond plutôt à des caractéristiques du véhicule. Malgré la certaine proximité intra-bloc constatée ici, l'ensemble de ces variables sont des critères clés du contrat d'assurance, et il parait donc difficile d'écarter arbitrairement l'une d'entre elles. Outre ces deux blocs, une poignée de variables ne présentent aucune structure de corrélation particulière. Celles-ci représentent également des régresseurs clés au sein de la modélisation de la sinistralité : soit parce qu'elles constituent de potentiels déterminants du risque ; soit en tant que variables instrumentales, visant à contrôler la corrélation des autres prédicteurs avec l'erreur du modèle.

4.3.3 Statistiques descriptives bivariées

Suite à l'analyse des corrélations entre variables explicatives, étudions désormais l'interaction que manifeste chacune d'entre elle avec les deux variables à prédire : la fréquence et le coût des sinistres. À cette fin, les graphiques ci-après représentent l'évolution de la sinistralité moyenne en fonction de la valeur de quelques facteurs de risque clés. Afin d'apprécier cette interaction au regard de la répartition de la population, ces courbes sont accompagnées des histogrammes[24] des variables explicatives.

FIGURE 20 – Coût moyen en fonction de l'`Âge Assuré`

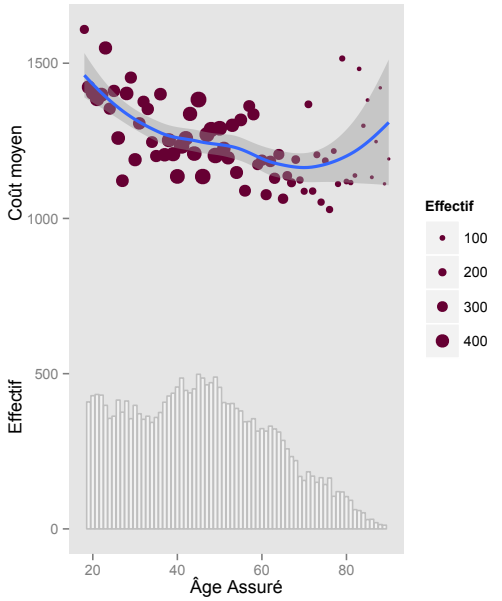
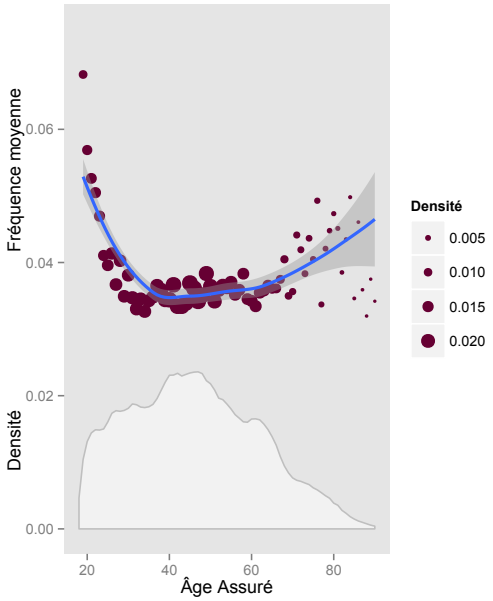


FIGURE 21 – Fréquence moyenne en fonction de l'`Âge Assuré`



L'âge de l'assuré présente un comportement parabolique vis-à-vis des deux variables expliquées. Ce phénomène assez classique traduit un risque accentué au niveau des deux segments d'âge extrêmes : les individus jeunes et âgés sont les plus enclins à subir un sinistre.

[24] Nous avons préféré représenter la densité de la distribution pour les graphiques figurant la fréquence de sinistres, car l'échelle correspondait alors mieux pour les deux quantités.

FIGURE 22 – Coût moyen en fonction de l'`Ancienneté Véhicule`

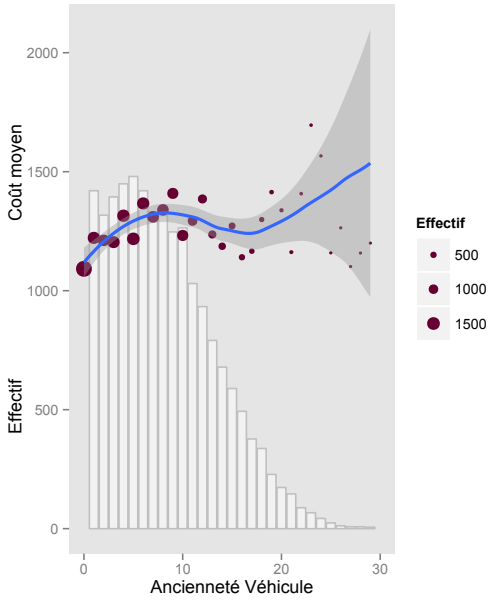
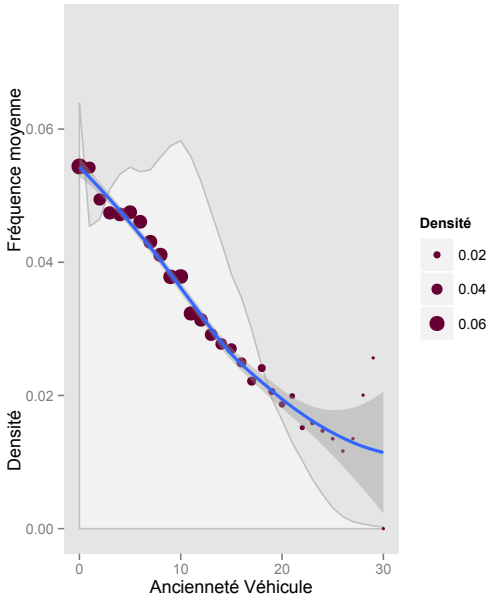


FIGURE 23 – Fréquence moyenne en fonction de l'`Ancienneté Véhicule`



Alors que le coût des sinistres augmente légèrement avec l'ancienneté du véhicule, leur fréquence est nettement décroissante. Ceci peut s'expliquer par exemple par une utilisation moindre des véhicules les plus anciens.

FIGURE 24 – Coût moyen en fonction du `Groupe SRA`

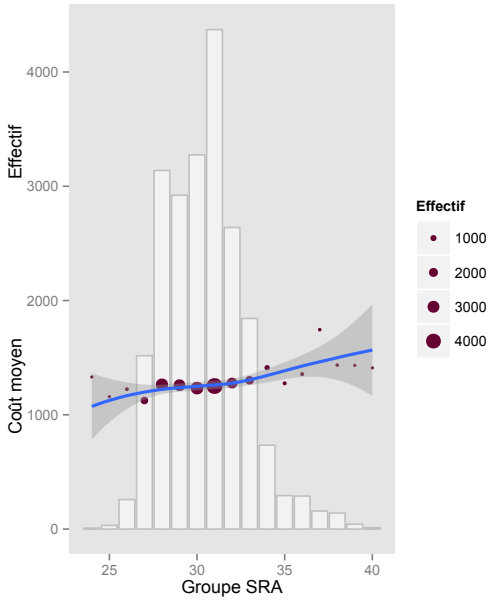
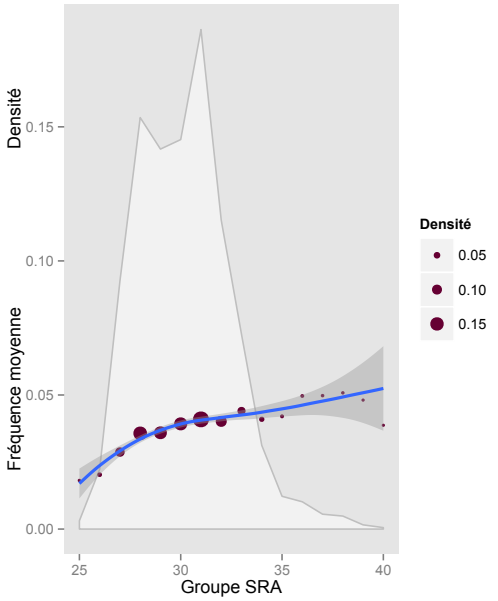


FIGURE 25 – Fréquence moyenne en fonction du `Groupe SRA`



Les courbes d'évolution de la sinistralité en fonction du Groupe SRA peuvent paraître relativement stables, mais ce constat doit être nuancé par l'échelle des ordonnées, qui indique que le risque augmente substantiellement, et de manière linéaire, avec le Groupe SRA, en comparaison

des autres graphiques.

FIGURE 26 – Coût moyen en fonction de la `Classe SRA`

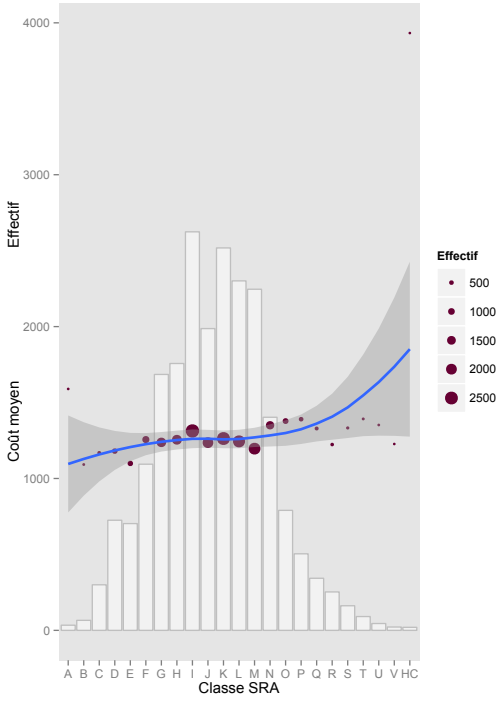
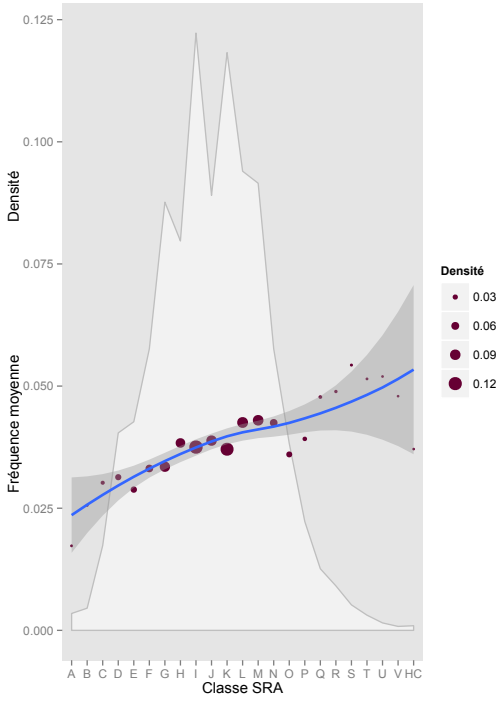


FIGURE 27 – Fréquence moyenne en fonction de la `Classe SRA`



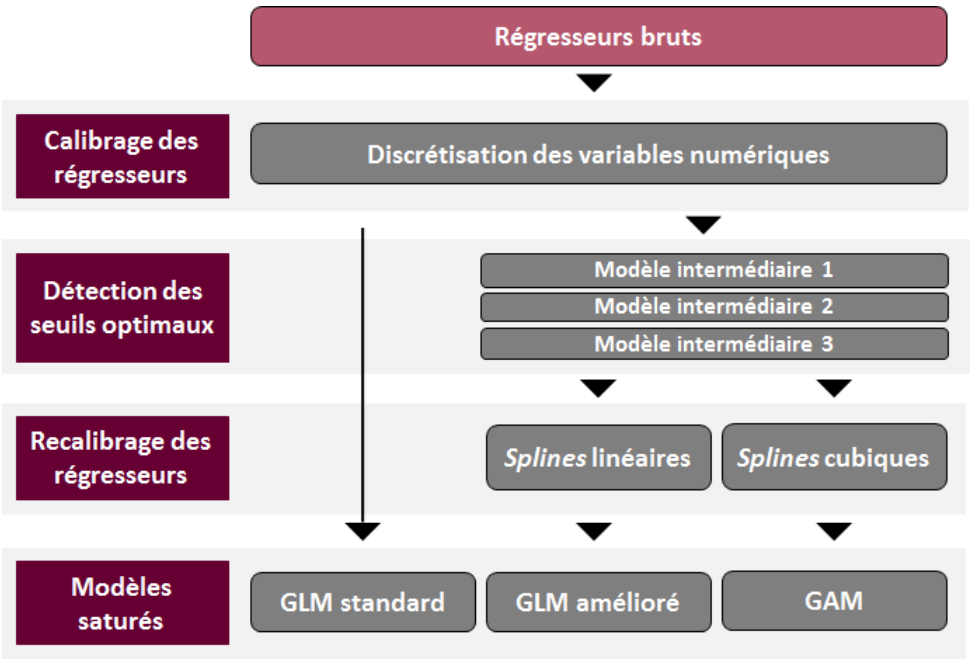
Enfin, pour la Classe SRA, seule la fréquence de sinistres présente un comportement strictement monotone, à la hausse. Le coût des sinistres apparaît quant à lui stable sur l'ensemble des valeurs de cette variable, ce qui est plutôt étonnant compte tenu du caractère financier de la classification SRA.

5 Stratégie économétrique

La stratégie économétrique détaillée dans cette partie décrit avec détails un processus opérationnel capital pour les actuaires en charge de la tarification automobile. Le calibrage des régresseurs sélectionnés exige en pratique des ressources en temps et en moyens importants pour établir des tarifs précis et exhaustifs. Cette méthodologie comprend l'identification des seuils optimaux, et la spécification des degrés adéquats des polynômes employés, qui sont des paramètres cruciaux, déterminant la qualité du modèle final. Le temps passé à observer le comportement du modèle selon différents paramétrages fournit à l'actuaire une expérience précieuse de la sinistralité, ce qui lui permet ainsi de parfaire son modèle avec son propre *a priori*. L'usage de *splines* présenté ici permet donc, à travers le calibrage afférent, de mieux apprécier le phénomène étudié, en plus d'améliorer potentiellement le modèle tarifaire existant.

Comme nous l'avons précisé dans le cadre de la présentation du portefeuille d'étude (partie 4.1), les deux garanties qui nous intéressent dans ce mémoire sont les Responsabilités Civiles (RC) Corporelle et Matérielle. Alors que la première concerne des sinistres généralement plus graves, mais également plus rares, la seconde est représentée par un échantillon plus large. Dans la suite du mémoire, nous nous proposons donc de focaliser nos analyses sur cette garantie RC Matérielle qui offre davantage de données d'étude et constitue ainsi une base plus appropriée pour produire des modèles pertinents. Cette partie décrit la stratégie suivie lors du calibrage des modèles économétriques, à savoir le GLM standard, le GLM amélioré, et le modèle GAM. Le GLM standard correspond au modèle classique appliqué aux variables continues discrétisées de manière arbitraire. Le GLM amélioré utilise des modèles intermédiaires afin de calibrer méthodiquement le nombre et la localisation des n^{o} uds de *splines* linéaires. Enfin, ces mêmes nœuds sont exploités pour le paramétrage du GAM, employant des *splines* cubiques naturelles. Les modèles produits à ce stade sont dits saturés car ils sont basés sur l'ensemble des variables présélectionnées. Ils seront ensuite élagués à travers des procédures de sélection de variables postérieures (cf. partie 3). Le schéma suivant résume notre approche de calibrage.

FIGURE 28 – Démarche de la phase de calibrage des modèles économétriques saturés



5.1 GLM standard

Dans un premier temps, nous élaborons le modèle GLM composé dit standard. Celui-ci correspond à la construction des modèles linéaires généralisés pour la fréquence et le coût à partir des variables brutes de retraitements supplémentaires, hormis la discrétisation des variables continues. Comme constaté dans le cadre de la revue de littérature, cette discrétisation constitue une véritable pratique de marché pour deux raisons. Premièrement, il existe rarement de relation de dépendance linéaire entre les facteurs de risque et la prime pure (Ohlsson [27]). En particulier, les impacts des variables explications sont généralement **non monotones** et peuvent présenter des **effets de seuils, d'extrêmes et de convexité**. Construire des variables catégorielles permet donc de prendre en compte ces impacts non linéaires. Deuxièmement, cette procédure produit des tarifs simples, homogènes par classe de risque, directement utilisables dans des politiques de souscription.

Les variables continues qui ont été discrétisées au sein de la modélisation sont classifiables en trois catégories :

- les variables d'ancienneté : l'âge de l'assuré, l'ancienneté du véhicule et celle du permis ;
- les critères SRA : la **Classe** de prix, le **Groupe**, et la classe de **Réparation** ;
- les indicateurs de bonus-malus : le CRM, et le CRM précédent

Afin de limiter le nombre total de paramètres au sein du modèle, le nombre de seuils de discrétisation est restreint à une dizaine pour chaque variable continue d'intérêt. Cette méthode revient à utiliser une fonction en escalier (ou *spline* de degré 0) et produit donc des effets « en plateaux ». Les résultats graphiques de cette première méthodologie sont présentés en partie 7. Mais les graphiques 29 à 32 ci-dessous illustrent déjà l'évolution des coefficients des régression pour des modèles similaires.

5.2 GLM amélioré

Le modèle précédent peut être amélioré en calibrant soigneusement les effets de seuils des variables continues. La discrétisation opérée précédemment permet d'observer l'évolution de l'impact d'une variable au sein du modèle. En effet, en représentant graphiquement les valeurs des coefficients estimés par le modèle pour toutes les classes d'une même variable, il est alors possible de repérer une éventuelle tendance. Comme ces coefficients correspondent aux effets sur la variable latente du modèle, leur valeur n'est pas directement interprétable en termes d'impact sur la variable réponse. Pour cela, il faudrait baser les analyses sur les effets marginaux, qui diffèrent de ces coefficients par l'application de la fonction de lien inverse ; la fonction logarithmique dans cette situation. En revanche, les effets de seuils subsistent avec cette transformation strictement monotone, ce qui justifie à cette fin l'utilisation des coefficients sous-jacents au modèle.

Les graphiques des coefficients estimés associés à quelques variables discrétisées d'intérêt sont présentés plus bas. Dans le cadre de cette démarche de calibrage plus fin, nous avons souhaité obtenir une vision plus détaillée des tendances qu'avec le modèle précédent. Pour ce faire, nous avons donc démultiplié les seuils de discrétisation pour certaines variables dont l'ensemble de définition était étendu, telles que l'âge. Toutefois, pour éviter de sur-paramétrer le modèle et risquer ainsi d'altérer sa convergence, les variables d'intérêt sont discrétisées les unes après les autres, par catégorie (ancienneté, SRA et bonus-malus). Pour chaque catégorie de variables, les régresseurs concernés sont donc discrétisés puis employés dans un GLM intermédiaire. À l'issue de cette modélisation intermédiaire, la représentation graphique des coefficients associées à une même variable explicative doit permettre d'identifier les seuils pertinents correspondants. Enfin, les informations récoltées au cours des trois modèles intermédiaires sont finalement exploités dans le cadre d'un dernier GLM. Cette démarche aboutit donc à la production du modèle que nous nommons « GLM amélioré ».

5.2.1 Sévérité

Dans les graphiques suivants (figures 29, 30, 31 et 32), sont représentés, pour la variable de sévérité, les coefficients issus des trois modèles intermédiaires sous forme de nuage de points, ainsi que les intervalles de confiance associés. La taille des points correspond à l'effectif de la classe d'individus sous-jacente. Enfin, une courbe de tendance polynomiale est également tracée, pondérée par les expositions totales des observations dans chaque classe. Ces graphiques permettent ainsi le constat d'effets d'extrêmes liés à l'âge de l'assuré (figure 29), ou encore d'un pic d'impact pour le groupe SRA (figure 32).

FIGURE 29 – Coefficients associés à l'`Âge de l'assuré` (sévérité)

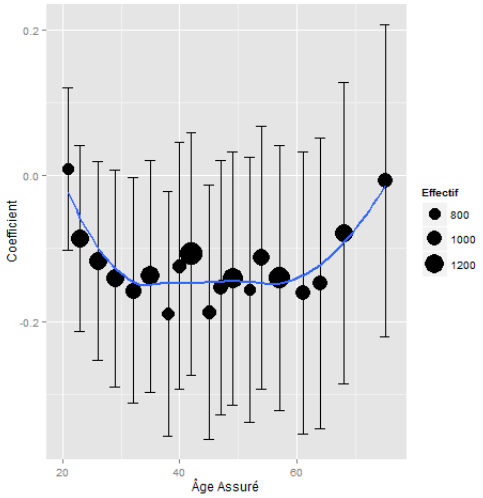


FIGURE 30 – Coefficients associés à l'`Ancienneté du véhicule` (sévérité)

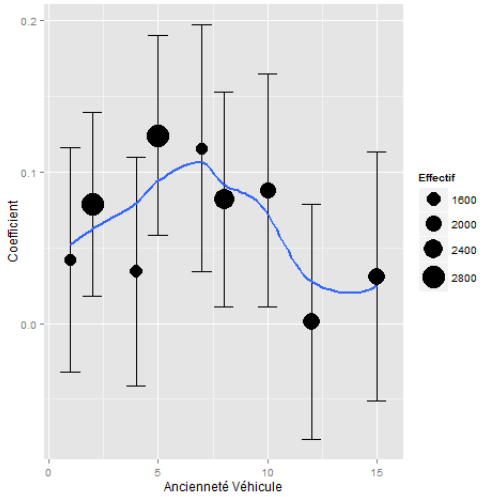


FIGURE 31 – Coefficients associés à la `Classe SRA` (sévérité)

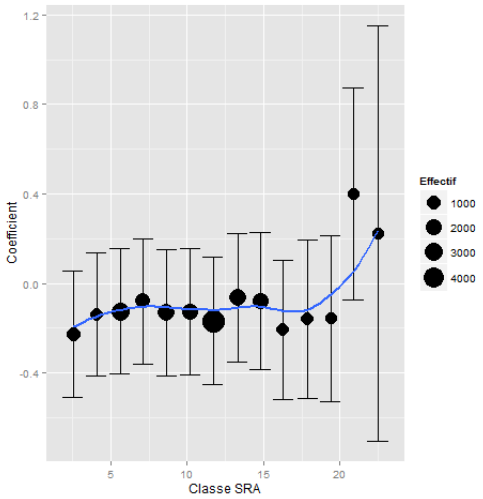
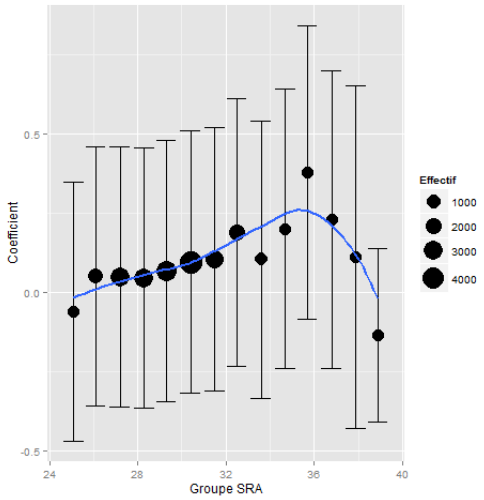


FIGURE 32 – Coefficients associés au `Groupe SRA` (sévérité)



Ces graphiques, qui correspondent aux modèles intermédiaires aboutissant au GLM amélioré, illustrent néanmoins aussi le comportement du GLM standard. Chaque point représenté correspond au coefficient de régression d'un segment de risque. Et l'emploi d'indicatrices pour encoder l'appartenance à ces segments de risque conduit à l'obtention de prédictions « en plateaux », formant ainsi une ligne brisée (cf partie 7).

Afin de prendre en compte les divers effets non linéaires, il est possible d'exploiter ces différents graphiques pour identifier, de manière certes empirique, les seuils de discrétisation les plus pertinents. Par exemple, l'âge de l'assuré semble présenter des ruptures de tendances aux âges 31 et 59 : entre ces deux âges, les coefficients estimés sont stables, ce qui indique un effet strictement monotone sur ce segment. Pour l'ancienneté du véhicule, l'identification est moins évidente. Nous choisissons finalement deux nœuds de discrétisation aux points 6 et 12. Un unique nœud apparaît pertinent dans les cas de la Classe et du Groupe SRA, respectivement les points 18 et 36.

Les autres variables numériques d'intérêt ont également fait l'objet d'une démarche similaire. Les graphiques correspondants sont présentés en annexe (partie A.1).

Après avoir identifié des points seuils au sein du domaine de définition de la variable considérée, il est alors possible de les spécifier dans l'entraînement du modèle, en ayant recours à des transformations en *splines* de degré 1 avec un ou deux nœuds de discontinuité. Nous nous limiterons en effet au choix de deux nœuds par souci de parcimonie. Il est de toute manière rare d'observer davantage de ruptures pour des variables tarifaires. Conceptuellement, l'utilisation de ces *splines* simples revient à inclure dans le modèle, pour une variable X_j et un seuil s, les transformations de la forme suivante :

$$(s - X_j)^+ \text{ et } (X_j - s)^+ \tag{39}$$

Cette amélioration permet ainsi d'estimer un coefficient de régression pour chacun des deux ou trois intervalles alors spécifiés. Même si les *splines* de base réellement employées au sein de l'algorithme d'estimation sont en fait différentes de celles de l'expression 39, comme nous le mentionnions précédemment (partie 2.3.3), l'esprit de calibrage est bien le même : capturer les effets de discontinuité tout en conservant le canevas linéaire. La combinaison des différentes transformations d'une même variable explicative est représentée ci-après, pour l'âge de l'assuré (figure 33), et l'ancienneté du véhicule (figure 34), accompagnée des intervalles de confiance.

FIGURE 33 – *Spline* linéaire associée à l'`Âge de l'assuré` (sévérité)

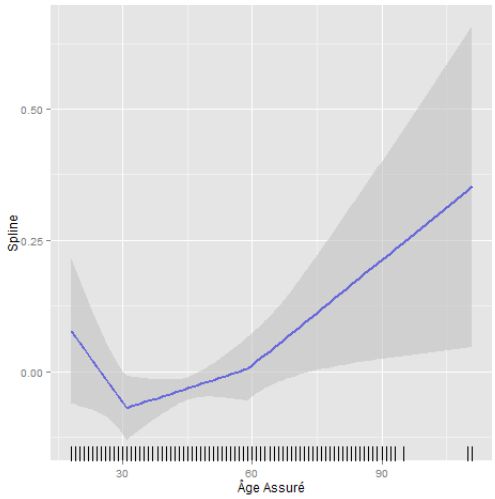
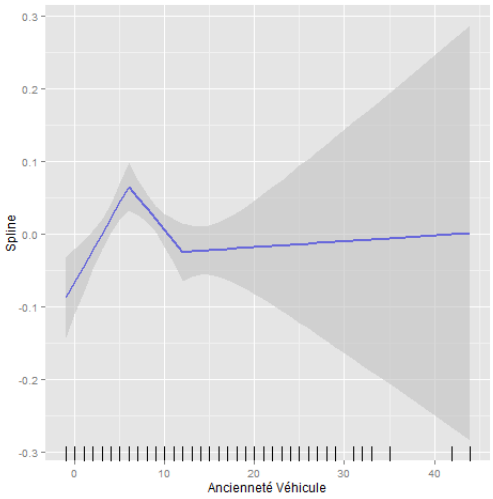


FIGURE 34 – *Spline* linéaire associée à l'`Ancienneté du véhicule` (sévérité)



Les comportements des courbes représentées par ces graphiques sont cohérents avec les premières intuitions basées sur les modèles intermédiaires précédents. Toutefois, ces comportements apparaissent davantage réguliers que les effets « en plateaux » produits par le modèle standard. Dans cette situation, le calibrage des nœuds permet non seulement de réduire le nombre de points de discrétisation mais aussi de conserver une tendance linéaire sur chacun des intervalles construits. Cette extension de l'approche classique, en apparence plus complexe, simplifie en réalité la mise en œuvre opérationnelle en réduisant le nombre de paramètres à estimer.

5.2.2 Fréquence

De même, nous reproduisons une démarche similaire pour la modélisation de la variable de fréquence des sinistres. Dans ce second chantier, la plupart des variables numériques envisagées pour la discrétisation n'ont pas présenté de comportement discontinue particulier. Les figures 35, 36 ou encore 38 illustrent des tendances davantage linéaires. Par conséquent, il est raisonnable de considérer que les effets relatifs à ces variables explicatives soient estimables par un unique coefficient de régression. De ce fait, seules trois variables ont nécessité, de notre point de vue, un traitement spécifique : la classe SRA, et les deux indicateurs de CRM.

FIGURE 35 – Coefficients associés à l'`Âge de l'assuré` (fréquence)

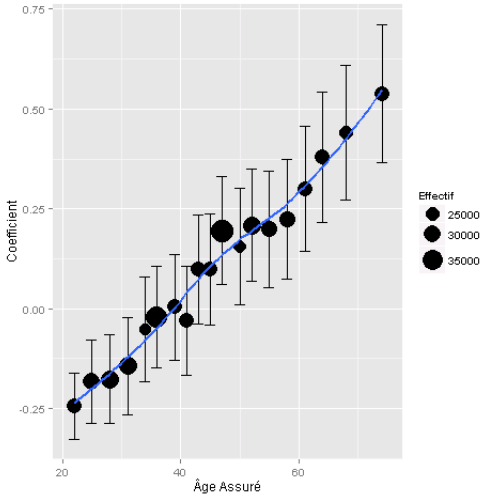


FIGURE 36 – Coefficients associés à l'`Ancienneté du véhicule` (fréquence)

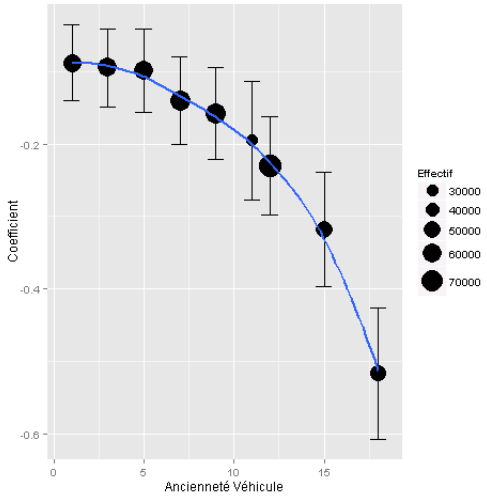


FIGURE 37 – Coefficients associés à la `Classe SRA` (fréquence)

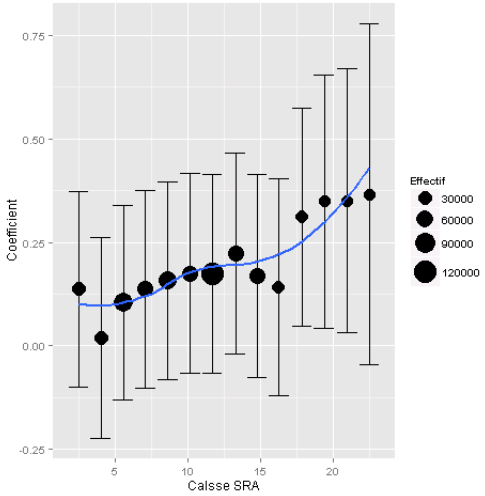
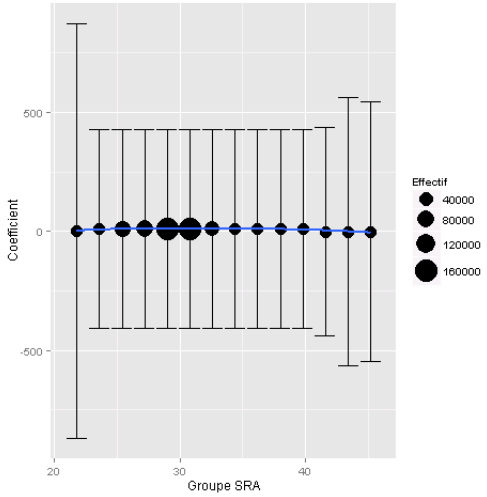


FIGURE 38 – Coefficients associés au `Groupe SRA` (fréquence)



Pour la Classe SRA, nous fixons un nœud au point 15. Pour les deux variables de CRM, dont les figures correspondantes sont données en annexe (partie A.1), le nœud choisi a pour valeur 0,75. Les *splines* résultantes de calibrage sont représentées ci-après (figures 39, 40 et 41).

FIGURE 39 – *Spline* linéaire associée à la `Classe SRA` (fréquence)

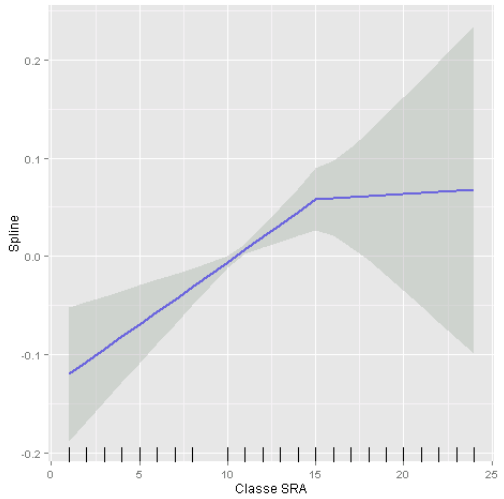


FIGURE 40 – *Spline* linéaire associée au `CRM` (fréquence)

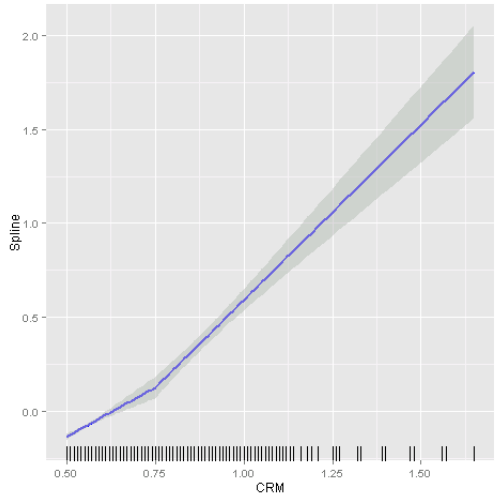
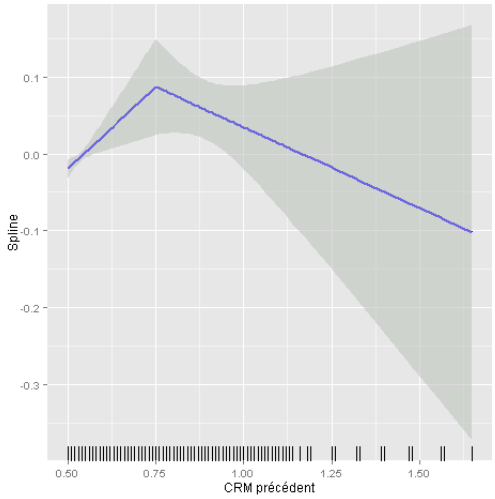


FIGURE 41 – *Spline* linéaire associée au `CRM précédent` (fréquence)



De même que pour la variable de sévérité, ces *splines* permettent une meilleure prise en compte de l'évolution non linéaire des impacts sur la variable de fréquence. Seule la modélisation du CRM ne semble pas être nettement améliorée à travers l'utilisation d'une *spline* : ses deux composantes présentent une tendance linéaire quasi-identique.

5.3 GAM

L'approche précédente, qui consiste à modifier légèrement le modèle GLM standard par la spécification de nœuds de discontinuité, entre déjà dans le cadre plus large des GAM, sans en utiliser tous les atouts. Afin d'obtenir un modèle encore plus flexible, il est envisageable de recourir à des *splines* plus complexes pour chacune des variables continues calibrées précédemment. Il est classique d'employer des *splines* cubiques naturelles, qui ont été introduites en partie 2.3.3. Le choix de transformations polynomiales de degré 3 s'explique par leur caractère populaire. À ce

stade, il est déjà important de prendre conscience du fort risque de sur-apprentissage planant sur des modèles aussi flexibles appliqués à des données qui, comme cela est prévisible au regard des graphiques précédents, ne présentent pas de relation de dépendance particulière à des degrés élevés. Nous souhaitons toutefois expérimenter la force des GAM à leur paroxysme à des fins de comparaisons. Enfin, l'utilisation de *splines* naturelles, définies en partie 2.3.4, permet de réduire légèrement le nombre de degrés de liberté, ce qui contribue à compenser en partie la volatilité accrue découlant de cette approche.

De même que pour les discrétisations des parties précédentes, l'usage de *splines* nécessite de spécifier des nœuds aux points de rupture de la relation de dépendance. Dans ce nouveau modèle, il n'y a pas de raison de recourir à des nœuds différents que pour le GLM amélioré. Il convient donc de conserver les résultats des travaux précédents lors de la spécification du modèle GAM.

5.3.1 Sévérité

Les graphiques suivants représentent les *splines* de quelques facteurs clés, issues du modèle additif employé, dans le cas de la sévérité. Ces courbes offrent un comportement évidemment plus régulier que les paramétrages précédents, ce qui est sans doute davantage pertinent au regard du comportement réel du phénomène de sinistralité. Toutefois, cette amélioration s'accompagne d'une volatilité augmentée, qu'il conviendra d'examiner à l'aide de mesures de test indépendantes.

FIGURE 42 – *Spline* cubique naturelle associée à l'`Âge de l'assuré` (sévérité)

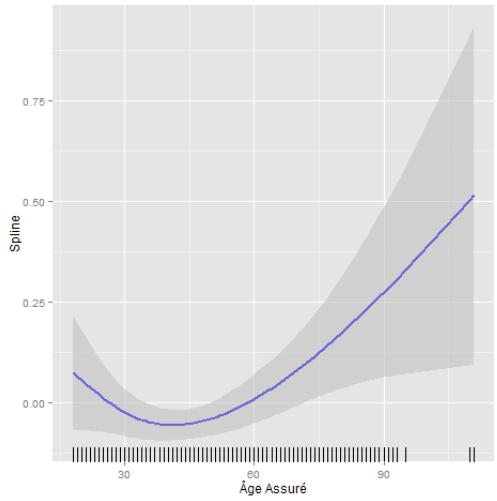
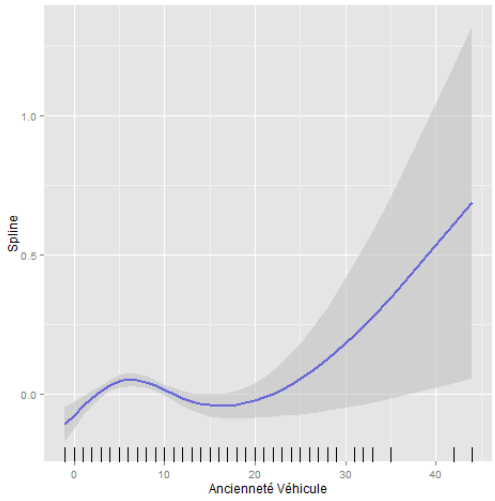


FIGURE 43 – *Spline* cubique naturelle associée à l'`Ancienneté du véhicule` (sévérité)



À part l'apport de régularité additionnelle, cette dernière approche économétrique ne semble pas fournir d'amélioration substantielle dans la modélisation de la sinistralité.

5.3.2 Fréquence

Enfin, les trois graphiques ci-dessous présentent les *splines* associées aux variables numériques transformées au sein du modèle de fréquence. Ces variables sont les mêmes que celles retenues au cours du calibrage du GLM amélioré : `Classe SRA`, `CRM`, et `CRM précédent`.

FIGURE 44 – *Spline* cubique naturelle associée à la `Classe SRA` (fréquence)

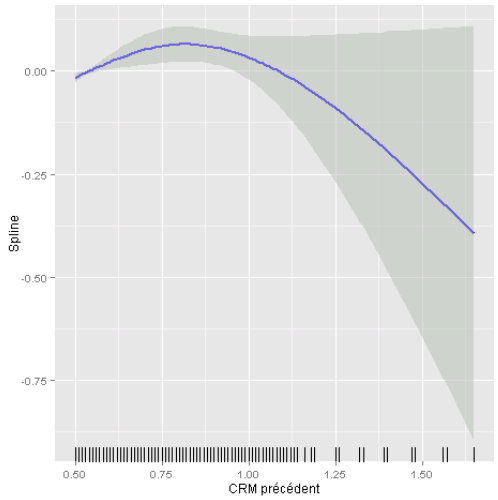


FIGURE 45 – *Spline* cubique naturelle associée au `CRM` (fréquence)

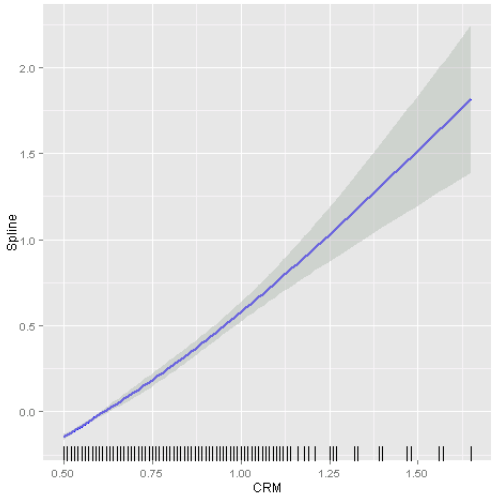
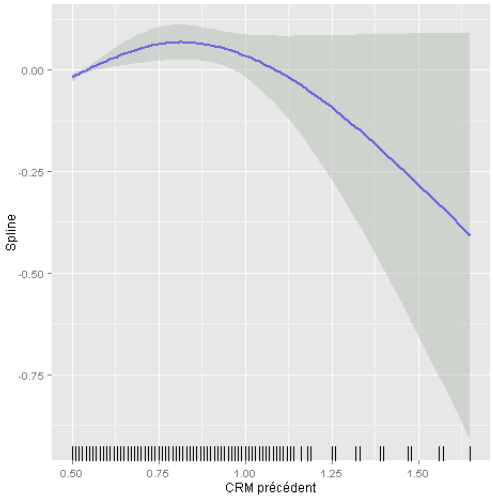


FIGURE 46 – *Spline* cubique naturelle associée au `CRM précédent` (fréquence)



En définitive, ces diverses méthodologies modifient la structure des données en amont des modèles économétriques. Elles se différencie selon le critère de la flexibilité, et aboutissent pour cette raison à des estimateurs de forme différente. Suite à cette première phase de calibrage, il convient d'écarter les facteurs de risque les moins pertinents afin de produire une segmentation simple et robuste.

6 Validation des modèles

Les techniques de sélection de variables employées ici peuvent être considérées comme autant d'outils permettant aux actuaires responsables de la tarification d'identifier les facteurs déterminants de la sinistralité. Non seulement les sélections obtenues doivent venir valider les variables tarifaires en place, mais elles peuvent également être exploitées à des fins d'analyse. En effet, il est important pour les opérationnels de mieux connaître les facteurs de risque susceptibles d'impacter significativement la sinistralité. Toute déviance de la rentabilité peut alors être pilotée de manière plus efficace en examinant en détails l'évolution de chacune de ces variables cernées. Dans cette situation, la cause de la dérive peut être plus facilement identifiée, et traitée plus rapidement à travers un remaniement des tarifs, soit en intégrant une variable manquante dans la segmentation en vigueur, soit en ré-évaluant les niveaux tarifaires liés à une variable existante.

Suite à la phase de calibrage des régresseurs en entrée des modèles économétriques, il est capital d'affiner ceux-ci par une revue de la pertinence des divers prédicteurs employés. À cette fin, nous mettons ici en œuvre la régularisation du Lasso, présenté en partie 3.3, afin d'écarter les variables les moins explicatives des modèles alors bâti. Cette démarche doit permettre de définir une sélection de facteurs à conserver, et produire ainsi des modèles plus parcimonieux, dits « optimaux », qui constitueront les **modèles définitifs** sur lesquels se basent les différentes analyses comparatives de la partie 7. Il nous paraît également important de valider la sélection retenue par deux autres approches alternatives : la procédure *stepwise forward*, et la forêt aléatoire. Dans un premier temps, nous décrivons la mise en place du Lasso, que nous illustrons par quelques graphes clés associés au modèle de fréquence. Puis nous présentons les points saillants de la procédure *stepwise* à travers l'exemple du modèle de sévérité. Nous fournissons ensuite la hiérarchisation des régresseurs issue de la forêt, et sur laquelle se base la méthode de sélection associée. Enfin, nous comparons les différentes sélections obtenues et évaluons *a posteriori* la pertinence du Lasso.

6.1 Lasso : exemple du modèle de fréquence

Afin d'aboutir à la sélection principale, nous menons une régression Lasso en **sur-couche** des modèles économétriques spécifiés lors de la partie 5. Cela signifie que nous conservons le calibrage des régresseurs réalisé précédemment et appliquons simplement l'opérateur Lasso au sein du processus d'estimation. Dans ce processus, chaque indicatrice et chaque *spline* élémentaire est considérée comme une variable à part entière et est donc, à ce titre, susceptible d'être écartée par la procédure de sélection. Illustrons la mise en œuvre de cette méthode par quelques résultats clés issus du modèle de fréquence. Comme nous le mentionnions en partie 3.3, le choix de la valeur du paramètre d'apprentissage λ demeure un véritable sujet. Le procédé de validation croisée permet d'obtenir deux valeurs d'intérêt que nous souhaitons désormais comparer. Les évolutions de l'erreur en fonction des valeurs de λ pour les trois modèles économétriques de la variable de fréquence sont tracées par les figures 47,49 et 51. Les intervalles de confiance de ces estimateurs y sont représentés. La droite verticale pleine indique la valeur de $\lambda_{Dev}^{\min}$, alors que la droite en pointillés désigne la valeur de λ_{Dev}^{1se}. À droite de ces graphiques sont représentées les évolutions du

nombre de coefficients non nuls, en fonction de λ. Cette courbe coupe systématiquement l'axe des abscisses en $\lambda_{\max}$, par construction de la suite des λ.

FIGURE 47 – Déviance moyenne estimée en fonction de λ (GLM standard)

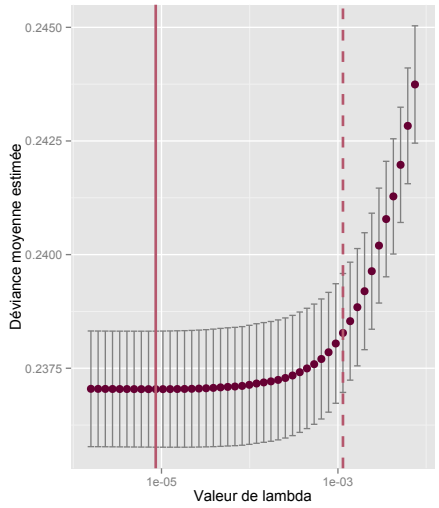


FIGURE 48 – Nombre de coefficients non nuls en fonction de λ (GLM standard)

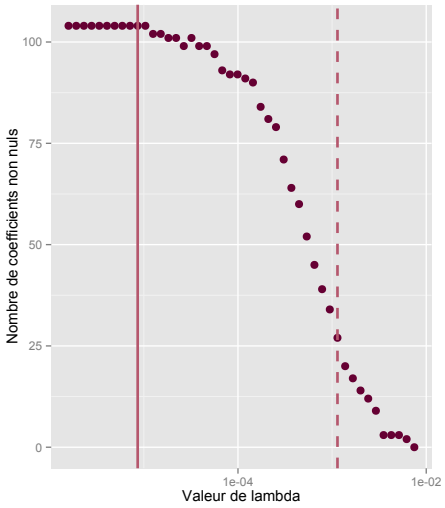


FIGURE 49 – Déviance moyenne estimée en fonction de λ (GLM amélioré)

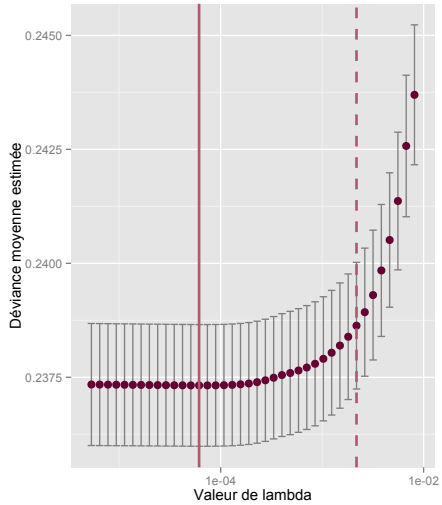


FIGURE 50 – Nombre de coefficients non nuls en fonction de λ (GLM amélioré)

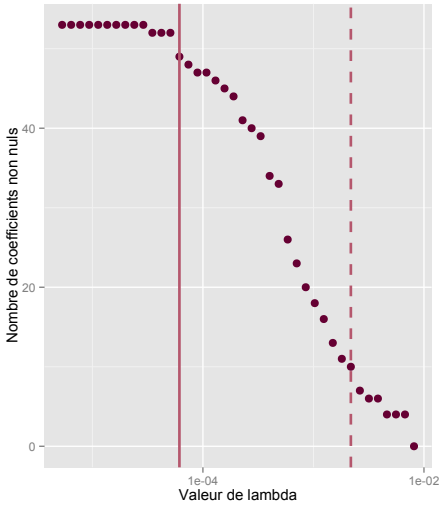


FIGURE 51 – Déviance moyenne estimée en fonction de λ (GAM)

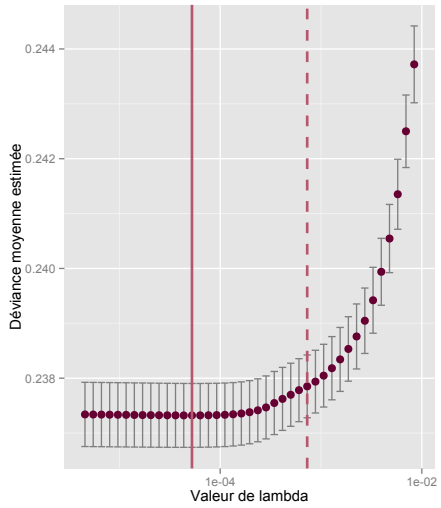
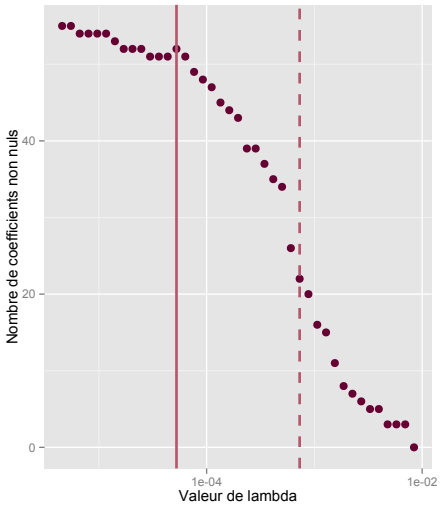


FIGURE 52 – Nombre de coefficients non nuls en fonction de λ (GAM)



Il convient de constater que le choix de λ_{Dev}^{1se} conduit effectivement à obtenir un modèle bien plus parcimonieux qu'avec $\lambda_{Dev}^{\min}$. Dans la suite de nos travaux, il nous paraît important de comparer les résultats produits par ces deux possibilités, puisque nous n'avons pas, à ce stade, d'argument particulier en faveur de l'une ou l'autre de ces valeurs.

6.2 *Stepwise* : exemple du modèle de sévérité

Afin d'illustrer la procédure *stepwise forward* qui a été employée, nous présentons ici quelques graphiques relatifs à son application au modèle de sévérité (coût). La valeur de l'AIC est forcément décroissante au cours de cette procédure, car c'est le critère de sélection de la variable la plus pertinente à chaque itération. Mais il est intéressant d'examiner le courbe d'évolution de cette quantité en fonction des étapes successives, afin d'apprécier le caractère convergeant de cette procédure.

FIGURE 53 – Evolution de l'AIC au cours de la procédure *stepwise* (GLM standard)

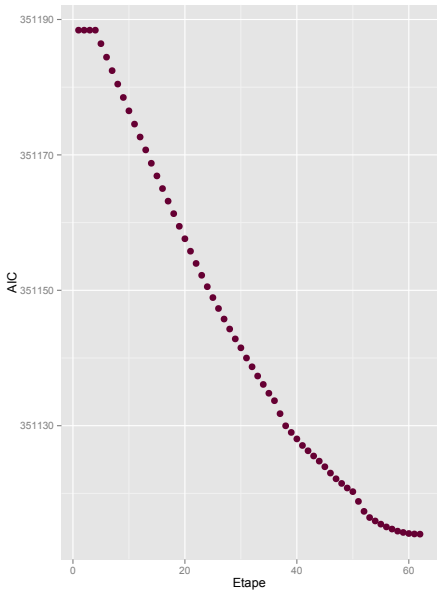
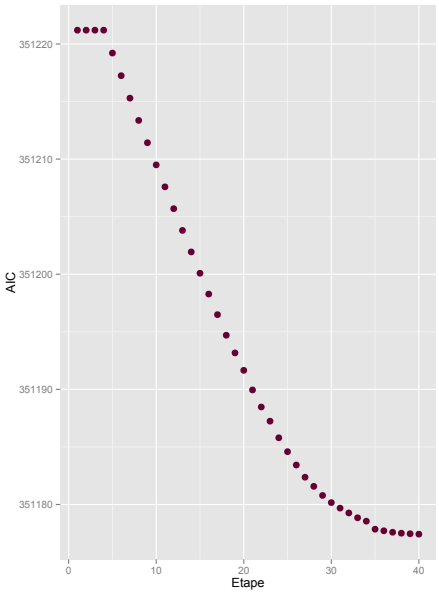
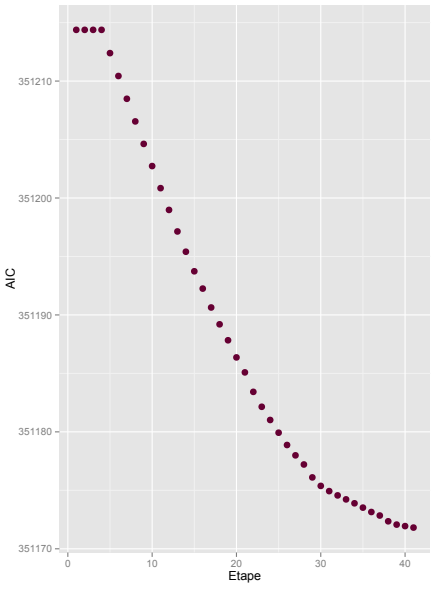


FIGURE 54 – Evolution de l'AIC au cours de la procédure *stepwise* (GLM amélioré)



Ces différents graphiques mettent en évidence une certaine stabilisation de l'AIC au niveau des dernières étapes de la procédure. Ceci indique donc que le critère d'arrêt du *stepwise* n'est pas anodin, et que la valeur finale de cette mesure correspond selon toute vraisemblance à un minimum global sur l'ensemble des combinaisons de variables existantes.

FIGURE 55 – Evolution de l'AIC au cours de la procédure *stepwise* (GAM)



En complément de ces figures, et de l'ensemble des analyses de la partie comparative suivante, le lecteur pourra se reporter en annexe, dans le tableau 11, pour les résultats exhaustifs de la procédure *stepwise* appliquée aux trois modèles économétriques. Outre les valeurs des estimateurs et leur intervalle de confiance, il y trouvera également diverses mesures de performance, donc l'AIC

et la déviance. Toutefois, ces différents résultats ne seront pas analysés outre mesure dans le cadre de cette étude, car ils sont souvent sujets à l'erreur d'interprétation en régression *stepwise*. En effet, cette technique invalide systématiquement les hypothèses qui lui sont généralement attribuées et elle est donc, à ce titre, vivement critiquée (partie 3.2.2).

6.3 Forêt aléatoire

Enfin, les résultats du *random forest* sont également exploitables afin de produire une sélection de variable alternative. Cette sélection est basée sur la mesure d'importance par permutation introduite en partie 3.4, elle-même calculée à partir des erreurs *out-of-bag*. Les figures 56 et 57 représentent l'évolution de cette erreur en fonction du nombre d'arbres composant la forêt.

FIGURE 56 – Erreur *out-of-bag* de la forêt aléatoire en fonction du nombre d'arbres (coût)

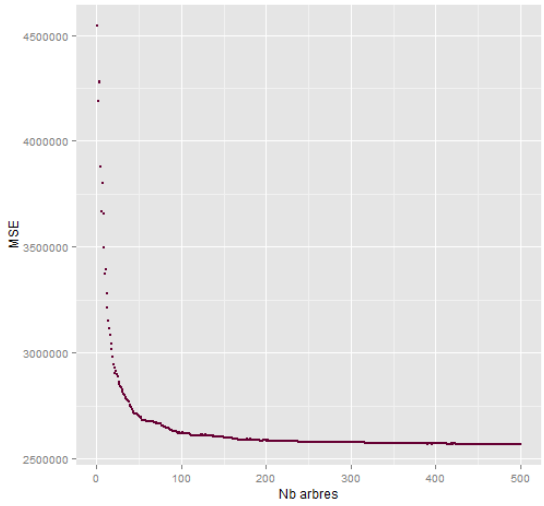
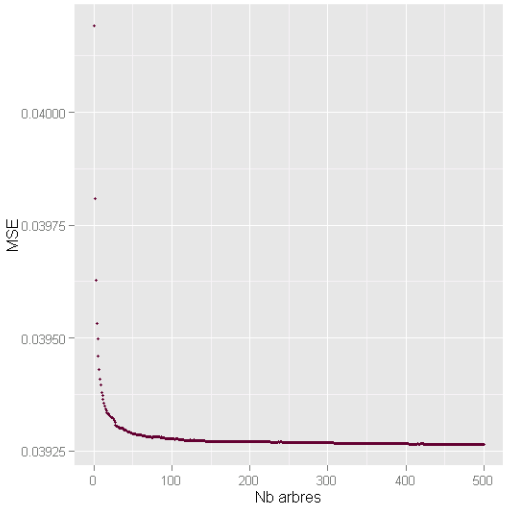


FIGURE 57 – Erreur *out-of-bag* de la forêt aléatoire en fonction du nombre d'arbres (fréquence)



Ces deux graphiques affichent une stabilisation de l'erreur *out-of-bag* au-delà de quelques centaines d'arbres. Ils justifient donc que le choix de 500 arbres conduit à produire un modèle robuste, et remplit donc l'objectif de limiter la volatilité associée. Nous présentons maintenant la valeur de l'importance de chaque prédicteur impliqué dans les modèles de sévérité et de fréquence.

FIGURE 58 – Importance des variables au sein de la forêt aléatoire (sévérité)

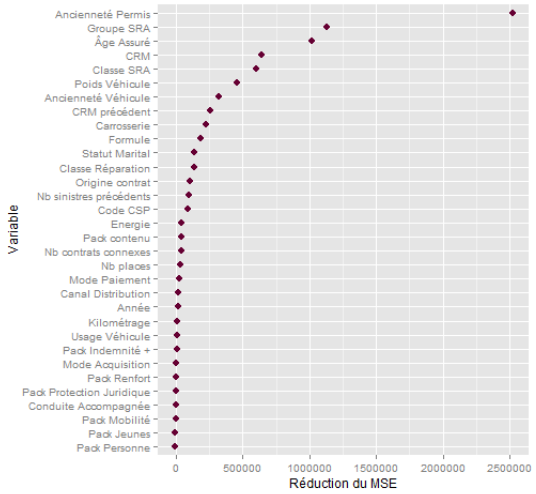
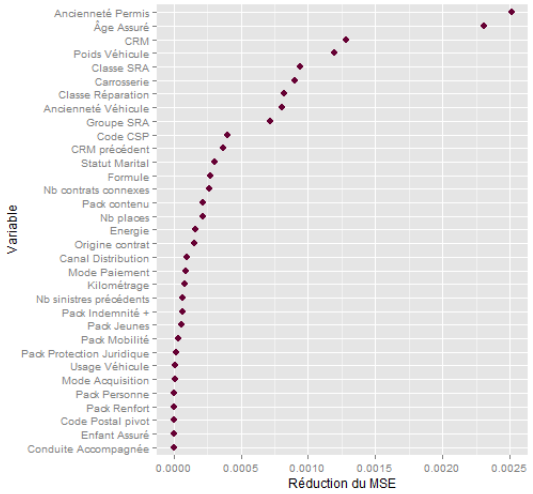


FIGURE 59 – Importance des variables au sein de la forêt aléatoire (fréquence)



Ces deux figures mettent en exergue une certaine disparité d'importance parmi les variables explicatives. En effet, certains facteurs clés se démarquent nettement du reste des régresseurs en termes de contribution à la formation de la forêt : l'`Âge Assuré` et l'`Ancienneté Permis`, le `Groupe` et la `Classe SRA`, le `CRM`.

6.4 Comparaison des sélections

Enfin, nous avons désormais l'ensemble des éléments requis pour déterminer trois sélections différentes. Même si nous retiendrons le Lasso pour des raisons théoriques, il nous paraît important de valider ce choix par les résultats des deux méthodes alternatives. Nous présentons ci-après les tableaux des variables sélectionnées par chacune des trois approches. Ces résultats ne sont donnés que pour l'exemple du GLM amélioré, par souci de concision. Pour le Lasso, les sélections résultant des deux valeurs de λ possibles sont fournies à titre indicatif. Précisons toutefois que le choix de $\lambda_{Dev}^{\min}$ est généralement privilégié. Pour le *random forest*, il est nécessaire de spécifier un niveau d'importance minimum permettant de filtrer les régresseurs. Ici le seuil peu filtrant de 0 a été considéré, ce qui a permis d'exclure une poignée de facteurs.

6.4.1 Sévérité

Le tableau ci-dessous indique les facteurs de risque sélectionnés par chaque approche pour le modèle de sévérité. Comme chaque facteur de risque est potentiellement décliné en plusieurs régresseurs au sein des modèles économétriques, via les mécanismes de transformation en indicatrices ou *splines*, il n'est pas possible de retranscrire exhaustivement les variables élémentaires conservées par ces méthodes. Par souci de clarté, nous avons choisi d'étiqueter les facteurs de risque par une croix ($\times$) dès lors qu'au moins l'une de leurs composantes était conservée par la procédure considérée. C'est la raison pour laquelle les résultats suivants donnent la fausse impression que les procédures employées sont peu sélectives.

TABLE 5 – Sélection de variables pour la sévérité

Catégorie	Variable	*Stepwise*	Lasso		*Random Forest*
			$\lambda^{\min}_{Dev}$	λ^{1se}_{Dev}	
Client	Âge Assuré	×	×		×
	Ancienneté Permis	×	×		×
	Statut Marital	×	×		×
	Code CSP	×	×		×
	Nb sinistres antérieurs	×	×		×
	CRM	×	×		×
	CRM précédent	×	×		×
Police	Formule Produit	×	×		×
	Origine Contrat	×	×		×
	Canal Distribution	×	×		×
	Conduite Accompagnée				
	Année Contrat	×	×		×
Véhicule	Ancienneté Véhicule	×	×		×
	Classe SRA	×	×		×
	Groupe SRA	×	×		×
	Réparation SRA				×
	Kilométrage	×	×		×
	Mode Acquisition				×
	Usage Véhicule	×			×

Tout d'abord, le Lasso utilisant la valeur de λ^{1se}_{Dev} produit une sélection étonnamment radicale : aucun prédicteur n'est conservé. Ce résultat nous semble exagéré. Les autres approches aboutissent dans l'ensemble à des sélections similaires. Il est vrai que le choix d'un seuil d'importance peu filtrant pour le *random forest* ne permet pas d'écrémer sensiblement les variables explicatives disponibles, ce qui nuit à la qualité de la comparaison. Néanmoins, ce tableau indique qu'aucune variable clé, comme l'âge ou l'ancienneté, n'a été complètement écartée par ces méthodes, ce qui constitue une conclusion importante.

6.4.2 Fréquence

Le modèle de fréquence introduit deux nouvelles variables explicatives par rapport au modèle de sévérité : une indicatrice précisant si un enfant est couvert par la police ; et le code postal pivot, qui indique si l'assuré est domicilié dans le département majoritaire du portefeuille d'étude, à savoir 45, unique facteur géographique exploité dans ce modèle.

TABLE 6 – Sélection de variables pour la fréquence

Catégorie	Variable	*Stepwise*	Lasso		*Random Forest*
			$\lambda_{Dev}^{\min}$	λ_{Dev}^{1se}	
Client	Âge Assuré	×	×		×
	Ancienneté Permis	×	×	×	×
	Statut Marital	×	×	×	×
	Code CSP	×	×		×
	Nb sinistres antérieurs	×	×	×	×
	CRM	×	×	×	×
	CRM précédent	×	×	×	×
	Code Postal pivot		×		×
Police	Formule Produit	×	×	×	×
	Origine Contrat	×	×		×
	Canal Distribution	×	×		×
	Enfant Assuré		×		
	Conduite Accompagnée		×		
Véhicule	Ancienneté Véhicule	×	×	×	×
	Classe SRA	×	×		×
	Groupe SRA	×	×	×	×
	Réparation SRA		×		×
	Kilométrage	×	×	×	×
	Mode Acquisition		×		×
	Usage Véhicule	×	×		×

Comme pour le modèle de sévérité, le Lasso utilisant le λ alternatif produit la sélection la plus parcimonieuse et écarte de manière étonnante l'âge de l'assuré. En revanche, les trois autres méthodes semblent peu ou prou équivalentes. Le choix du Lasso basé sur le $\lambda_{Dev}^{\min}$ ne nous paraît donc pas aberrant. Aussi, nous validons la sélection issue de cette approche pour l'élaboration des modèles finaux susceptibles d'être étudiés. Ces modèles finaux sont le résultat d'une ré-exécution du processus d'estimation classique des modèles économétriques, c'est-à-dire sans l'utilisation de l'opérateur Lasso, basé sur les variables sélectionnées ici.

7 Comparaison des résultats

Les résultats présentés dans cette partie illustrent les applications opérationnelles des outils développés dans cette étude. Les écarts de prédictions révélés par les indicateurs macroscopiques ou les graphiques univariés mettent en évidence les segments d'assurés qui méritent une attention particulière. En effet, lorsque les modèles diffèrent substantiellement pour une catégorie tarifaire donnée, cela indique que les prédictions sont instables et qu'il faut donc veiller à maintenir des tarifs conservateurs sur ces segments. Cette multiplicité des modèles offre ainsi un *benchmark* utile pour valider les niveaux tarifaires existants ou, à l'inverse, corriger des sources de risque localisées. De plus, en complément des modèles économétriques, la forêt aléatoire s'accompagne d'outils pratiques très avantageux pour compléter le processus tarifaire existant. Par exemple, les graphiques tridimensionnels permettent de détecter d'éventuelles interactions importantes, qui peuvent être ensuite spécifiées manuellement au sein du modèle économétrique principal. Par ailleurs, comme il s'agit d'une méthode robuste en présence de variables fortement corrélées, elle offre des prédictions témoins pour déceler un défaut de modélisation du GLM vis-à-vis de certains facteurs de risque. Ces différents indicateurs peuvent tout à fait être synthétisés et utilisés de manière récurrente par les équipes opérationnelles. Ils apportent des informations profitables pour corriger d'éventuelles lacunes du modèle tarifaire en place.

Suite aux différentes phases de calibrage des modèles retenus (partie 5), et de sélection de variables (partie 6), nous avons abouti à l'élaboration de modèles finaux susceptibles d'être comparés. Une large gamme d'outils permet d'évaluer la qualité de tels modèles : les mesures de performance prédictive, les graphiques d'impacts univariés, les représentations tridimensionnelles d'éventuelles interactions, etc. Afin d'effectuer ces comparaisons en toute impartialité, il convient d'appliquer ces outils sur une **base de test indépendante**, à savoir la seconde moitié des données initiales.

Afin d'évaluer la performance prédictive des différents modèles, nous choisissons deux mesures d'erreur traduisant des réalités distinctes. En premier lieu, le RMSE (*Root Mean Square Error*) constitue la fonction de perte quadratique classique, incontournable dans toute analyse comparative. En second lieu, la rentabilité technique reflète les intérêts commerciaux de la tarification, et constitue donc un indicateur macroscopique pertinent au regard des problématiques opérationnelles.

7.0.3 Erreur quadratique

Examinons tout d'abord les erreurs résiduelles à l'aide de la mesure de la racine de la moyenne des carrés des résidus, ou RMSE (*Root Mean Square Error*) :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{40}$$

La charge totale espérée correspond au produit de la sévérité espérée par la fréquence espérée. Afin d'avoir une appréciation des implications du choix du λ lors du calibrage du Lasso, nous avons choisi de présenter les erreurs associées aux deux valeurs de λ envisagées précédemment.

TABLE 7 – Erreur résiduelle quadratique (RMSE)

Variable	Calibrage Lasso	GLM Standard	GLM Amélioré	GAM	*Random Forest*
Sévérité	$\lambda_{\min}$	1 752,26	1 749,88	1 751,89	1 766,30
	λ_{1se}	1 755,18	1 754,01	1 754,93	
Fréquence	$\lambda_{\min}$	0,199 026	0,196 383	0,196 365	0,197 653
	λ_{1se}	0,198 943	0,196 514	0,196 408	
Charge totale	$\lambda_{\min}$	430,81	429,10	428,97	429,91
	λ_{1se}	430,67	429,12	428,99	

Nous constatons tout d'abord que la forêt aléatoire performe légèrement moins bien que les modèles économétriques et ce, de manière systématique. Ceci indique donc qu'un modèle aussi complexe n'est pas forcément plus pertinent dans le cadre de la tarification automobile. Rappelons toutefois que ces résultats sont largement conditionnés par les données exploitées, et les paramètres de calibrage utilisés. Il convient donc de nuancer les conclusions formulées à partir de ces différentes valeurs. Dans l'ensemble, les mesures de RMSE apparaissent **particulièrement proches**, mais cela provient sans doute du volume important des données de test sur lesquelles elles sont moyennées. Pour le modèle de sévérité, le GLM amélioré semble sur-performer dans une très faible mesure les deux autres modèles économétriques. Pour le modèle de fréquence, celui-ci présente une RMSE tout à fait comparable avec celle du GAM, et toutes deux sont inférieures à celle du GLM standard. Cependant, au regard de la charge totale de sinistres, c'est le GAM qui se démarque. En définitive, les deux modèles économétriques les plus flexibles apparaissent légèrement plus performants que la méthode classique. Ces résultats tendent donc à favoriser les modèles issus du calibrage méthodique des régresseurs. Cependant, insistons sur le fait que ces valeurs sont **extrêmement similaires**, et qu'il est donc impossible de produire des analyses tranchées à partir de ces résultats. De surcroît, l'ordre de grandeur du RMSE n'est pas réellement interprétable opérationnellement dans le cadre concret de la tarification automobile, ce qui limite encore davantage les conclusions formulées. Afin de pallier ces lacunes, complétons ces premiers résultats comparatifs par l'étude des S/P.

7.0.4 Rentabilité technique

Une alternative à la mesure du RMSE est celle de la rentabilité technique à l'aide du ratio sinistres/primes (*loss ratio*), qui représente un indicateur davantage opérationnel, susceptible de fournir de meilleures intuitions sur les performances relatives des différents modèles au regard de la problématique de tarification :

$$S/P = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \hat{Y}_i} \tag{41}$$

Cet indicateur économique permet de juger si les modèles tarifaires proposés ici conduisent à une gestion pérenne du portefeuille automobile, à travers des estimations fiables de la prime pure. Il convient d'examiner ici si les S/P suivants sont proches du niveau théorique parfait de 100%.

Table 8 – S/P technique

Variable	Calibrage Lasso	GLM Standard	GLM Amélioré	GAM	*Random Forest*
Charge totale	$\lambda_{\min}$	101,61%	98,75%	116,15%	97,14%
	λ_{1se}	103,07%	99,08%	115,50%	

Avec cette seconde mesure, le GAM s'éloigne fortement de tous les autres modèles. Vient ensuite la forêt aléatoire qui est légèrement moins performante que les deux derniers modèles économétriques, au coude à coude. Aussi, en tenant compte de ces deux tableaux de résultats, nos conclusions demeurent mitigées, mais le GLM amélioré apparaît globalement plus stable, et relativement performant. Il est important de noter que les deux mesures d'erreur utilisées produisent des résultats relativement différents, ce qui démontre qu'**il n'existe pas d'indicateur idéal**. Par exemple, le RMSE ne reflète pas l'écart considérable du GAM, de plus de 10 points de pourcentage, observé avec le S/P. Les conclusions opérationnelles sont donc également très limitées pour cette seconde mesure d'erreur. Mais l'étude demeure néanmoins intéressante afin d'apprécier le caractère volatil de ces résultats. Face à ces défauts, il est capital de comparer les caractéristiques des différents modèles à un niveau plus microscopique, sur des segments de risque précis, à travers les impacts univariés.

7.1 Impacts univariés

Après ces indicateurs macroscopiques, il convient d'étudier les divergences de prédiction entre les différents modèles selon des segments spécifiques, afin d'avoir une vision plus précise des phénomènes en présence. À cette fin, nous présentons ci-après les graphes de l'évolution des valeurs prédites en fonction d'une variable clé donnée, toutes choses étant égales par ailleurs. Le caractère *ceteris paribus* de ces graphes est garanti par la génération d'un profil médian vis-à-vis de toutes les autres variables, selon les deux règles suivantes :

- la valeur médiane pour les variables quantitatives ;
- la modalité la plus fréquente pour les variables qualitatives.

Le profil médian ainsi considéré est décrit par le tableau suivant.

TABLE 9 – Caractéristiques du profil médian

Catégorie	Variables	Valeur
Assuré	Âge	45 ans
	Ancienneté du permis	13 ans
	Statut marital	Marié(e)
	Code CSP	Salarié
	Nombre de sinistres antérieurs	0
	Coefficient Réduction Majoration (CRM)	0,5
	CRM précédent	0,5
Police	Formule produit	Formule 3
	Origine du contrat	Migration
	Canal de distribution	Agent
	Enfant d'assuré	Non
	Conduite accompagnée	Non
Véhicule	Ancienneté du véhicule	9 ans
	Classe SRA	Classe I
	Groupe SRA	Groupe 30
	Réparation SRA	Classe L
	Kilométrage	0–5000 km
	Mode d'acquisition	Autre
	Usage	Tous usages

7.1.1 Sévérité prédite pour un profil médian fictif

Les modèles de sévérité présentent les plus grandes divergences en termes d'impacts univariés. Sur chacun des graphiques suivants, sont représentées les courbes des valeurs prédites par les quatre modèles en fonction des valeurs d'une variable d'intérêt.

FIGURE 60 – Sévérité prédite en fonction de l'`Âge de l'Assuré`

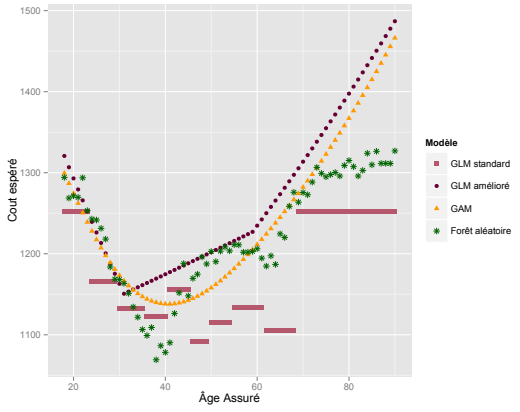
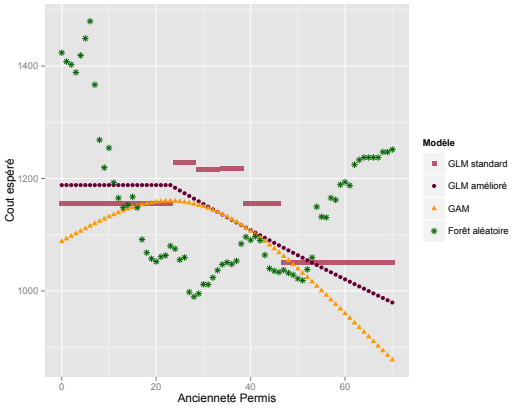


FIGURE 61 – Sévérité prédite en fonction de l'`Ancienneté du Permis`



Alors que les comportements des valeurs prédites par les différents modèles sont comparables

selon l'âge de l'assuré (figure 60), une importante disparité apparaît selon l'ancienneté du permis (figure 61), entre la forêt aléatoire et les modèles économétriques. En effet, le nuage de points associé à la méthode du *machine learning* présente une parabole inversée par rapport aux trois autres tracés. Cette singularité est en réalité l'illustration d'un phénomène fréquent, imputable aux GLM, en présence de variables fortement corrélées. En effet, ces modèles sont généralement incapables de déterminer les impacts marginaux individuels associés à deux régresseurs très corrélés entre eux, tels que l'âge et l'ancienneté. Ces deux prédicteurs présentent une corrélation supérieure à 0,9. Et le comportement de la sinistralité vis-à-vis de ces deux facteurs de risque est tout à fait similaire, comme cela est prévisble : il s'agit d'une parabole tournée vers le haut, indiquant un risque accru au niveau des populations extrêmes. Toutefois, les modèles économétriques ont tendance à attribuer arbitrairement l'impact global des deux variables corrélées à l'une d'entre elles seulement, en l'occurrence l'âge de l'assuré. Pour compenser cet effet ainsi exagéré, les impacts de la seconde variable, l'ancienneté du permis, sont souvent inversés par rapport à la réalité, ce qui produit les courbes contradictoires de la figure 61. En revanche, la forêt aléatoire semble bien intégrer les impacts relatifs à chacun des deux régresseurs concernés, sans être brouillée par leur corrélation élevée. Ce phénomène se retrouve également dans le cadre du modèle de fréquence, à travers les figures 60 et 61.

FIGURE 62 – Sévérité prédite en fonction de l'`Ancienneté du Véhicule`

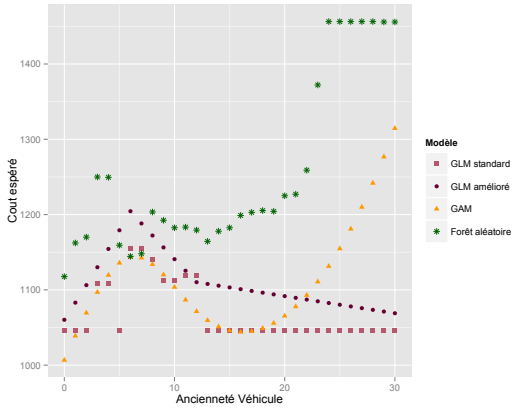
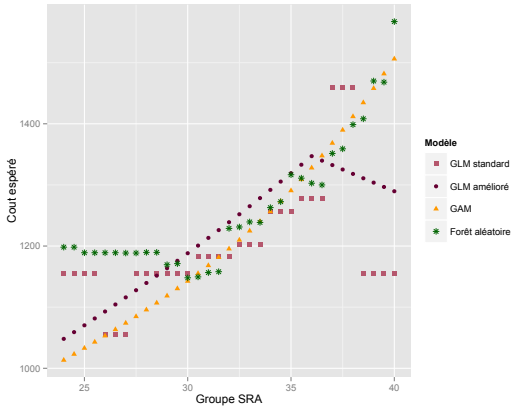


FIGURE 63 – Sévérité prédite en fonction du `Groupe SRA`



Le graphique 62 illustre de nouvelles disparités. Sur le tracé de l'ancienneté du véhicule, le modèle GAM est considérablement plus volatil que tous les autre modèles. Ce phénomène peut s'expliquer par la procédure de sélection de variable qui a conservé l'ensemble des composantes de la *spline* associée à cette variable. En conséquence, le comportement résultant est davantage flexible – et donc également moins robuste – que celui des autres modèles. Cet exemple montre donc l'importance de la sélection de variable préliminaire. Sur ce même graphique, la forêt aléatoire semble fournir un bon compromis entre la rigidité manifeste des deux GLM et la flexibilité du GAM, puisqu'il modélise bien une sévérité espérée accrue pour les véhicules les plus anciens. En revanche, les nuages de points du graphique 63 ne présentent pas de grande hétérogénéité.

FIGURE 64 – Sévérité prédite en fonction du `CRM`

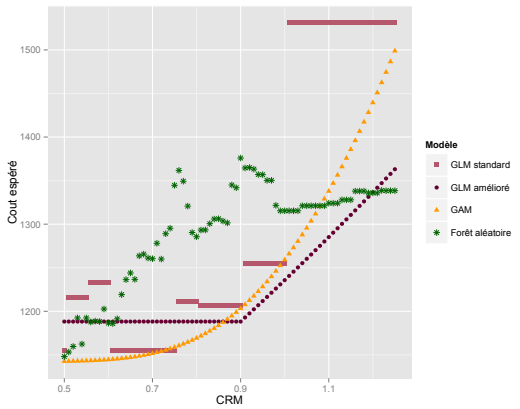
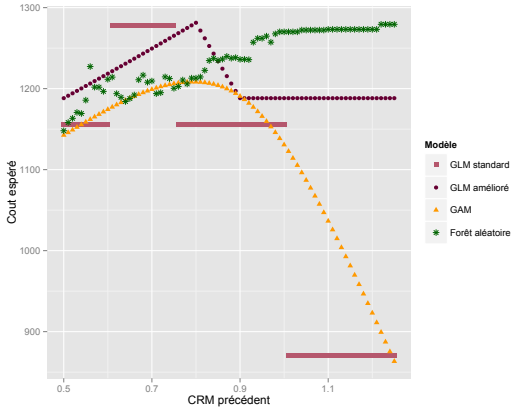


FIGURE 65 – Sévérité prédite en fonction du `CRM précédent`



Enfin, les graphiques liés au CRM présentent des divergences semblables à celles constatées pour les variables d'âge et d'ancienneté : alors que les trois modèles économétriques se comportent de manière analogue, la forêt aléatoire suit une tendance vraisemblablement inversée par rapport à ces derniers. Une analyse similaire à celle de l'âge peut ainsi être formulée : les modèles économétriques réagissent mal vis-à-vis de ces deux variables fortement corrélées que sont les CRM de deux années consécutives. Par conséquent, ce type de méthodes ne parvient pas à modéliser correctement les impacts réels propres à chacun de ces deux régresseurs, et tendent à estimer des effets opposés. En particulier, la forêt aléatoire offre un tracé plus cohérent avec la réalité sur la figure 65 : l'augmentation du risque avec le CRM de l'année précédente, et traduit donc un malus précédent élevé. De plus l'évolution des valeurs prédites par cette méthode de *machine learning* est plus stable que pour les autres modèles, ce qui semble aussi raisonnable puisque cette variable retardée devrait apporter, en toute logique, moins de pouvoir explicatif que son homologue actuel. Alors que les modèles économétriques mettent l'emphase sur le CRM actuel et accentue ainsi la pente de la courbe associée, au détriment d'une estimation erronée de l'effet dû au CRM précédent, la forêt abouti à une estimation mesurée de chacune de des deux variables, offrant ainsi un modèle plus réaliste.

Par ailleurs, malgré le comportement globalement réaliste du *random forest*, celui-ci se caractérise par une certaine volatilité vis-à-vis du CRM. En effet, les valeurs prédites par ce modèle non orthodoxe varient beaucoup entre les CRM de 0,6 et 0,9, et n'affichent pas de tendance lisse comme celles produites par les modèles économétriques. Cet inconvénient de la forêt aléatoire est une véritable difficulté au regard des exigences tarifaires en matière de cohérence. **Il s'agit d'un écueil qui discrédite fortement l'applicabilité opérationnelle de cette technique alternative.**

7.1.2 Fréquence prédite pour un profil médian fictif

En parallèle de ces analyses relatives à la modélisation de la sévérité des sinistres, étudions les graphiques analogues pour la prédiction de la fréquence, variable autrement plus symbolique de la tarification automobile.

FIGURE 66 – Fréquence prédite en fonction de l'`Âge de l'Assuré`

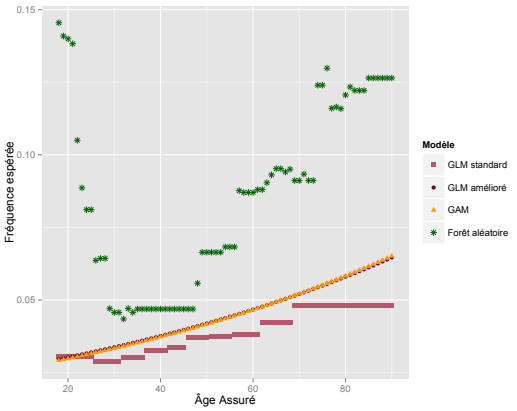
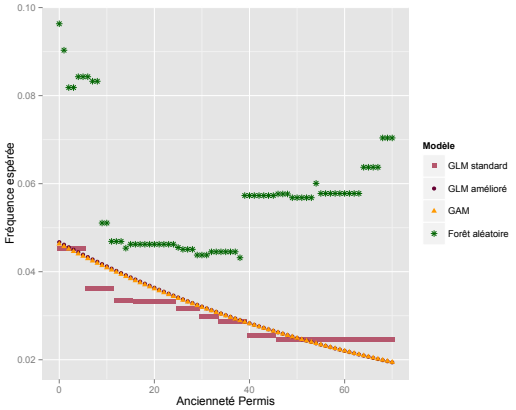


FIGURE 67 – Fréquence prédite en fonction de l'`Ancienneté du Permis`



Les constats formulés à partir de ces deux premiers graphiques sont équivalents à ceux de la sévérité de la partie précédente, mais ici le phénomène divergeant entre les modèles économétrique et la méthode de *machine learning* est encore plus prononcé. Alors que les GLM, et les GAM dans une moindre mesure, sont flattés pour leur robustesse intrinsèque, celle-ci apparaît excessive dans cette situation. Seule la forêt aléatoire parvient à capturer les effets d'extrêmes propres aux variables d'âge et d'ancienneté. Cela peut s'expliquer par l'absence de nœuds de discrétisation pour ces régresseurs au sein du GLM amélioré et du GAM. En effet, il convient de rappeler que les tendances des coefficients issus des modèles intermédiaires de calibrage (partie 5.2) ne présentaient pas de ruptures évidentes pour le modèle de fréquence (figure 35 et 88), ce qui ne justifiait ainsi pas l'introduction de nœuds internes particuliers pour ces deux prédicteurs. Le comportement extrêmement rigide observé *in fine* sur les deux graphes 66 et 67 ci-dessus résulte donc en partie de la démarche optée lors du calibrage préliminaire.

Il faut toutefois nuancer les propos précédents. Si les modèles économétriques sont impactés par la forte corrélation entre ces deux facteurs de risque, c'est aussi parce que celle-ci est substantielle et avérée. Dans ce sens, le couplage quasi-systématique de ces deux effets est parfaitement concevable, y compris pour des données de test. Aussi, les estimations globales résultants de ces approches classiques ne sont pas considérablement éloignés de la réalité. Insistons sur le fait que les graphiques ci-dessus ne représentent l'évolution de la sinistralité qu'en faisant varier un seul facteur de risque, les autres régresseurs étant maintenus à leur valeur médiane. Cette évolution ne tient pas compte des interactions entre plusieurs variables, et ne traduit donc pas la réalité des risques dans leur globalité. Pour ce faire, il conviendra d'examiner quelques interactions clés ultérieurement.

FIGURE 68 – Fréquence prédite en fonction de l'`Ancienneté du Véhicule`

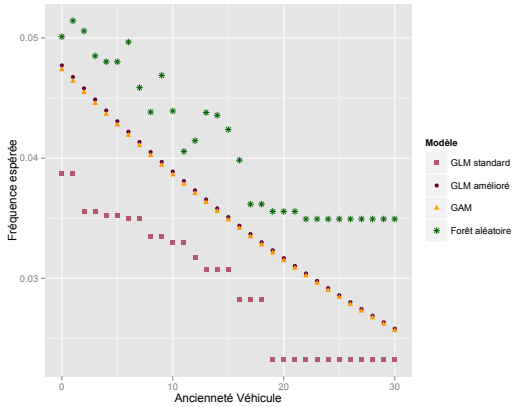
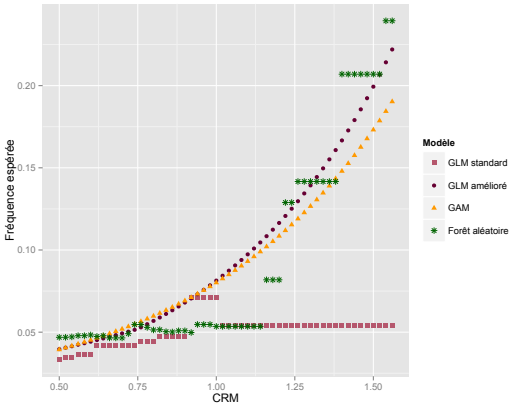


FIGURE 69 – Fréquence prédite en fonction du `CRM`



Dans certaines situations précédentes, la forêt aléatoire démontrait une meilleure flexibilité, à juste titre, dans la modélisation des effets d'extrêmes. Cette flexibilité peut également apparaître exacerbée dans d'autres cas, tels qu'à travers les figures 70 et 71 ci-dessous, conduisant ainsi à une variance superflue. En effet, le pic extrême de fréquence prédite pour les faibles valeurs du Groupe SRA est peu crédible, et prouve ainsi que **cette méthode de *machine learning* demeure, malgré son caractère ensembliste, sensiblement volatile.** Par conséquent, **la forêt ne peut produire des tarifs raisonnablement commercialisables** sans procédure de lissage supplémentaire.

FIGURE 70 – Fréquence prédite en fonction du `Groupe SRA`

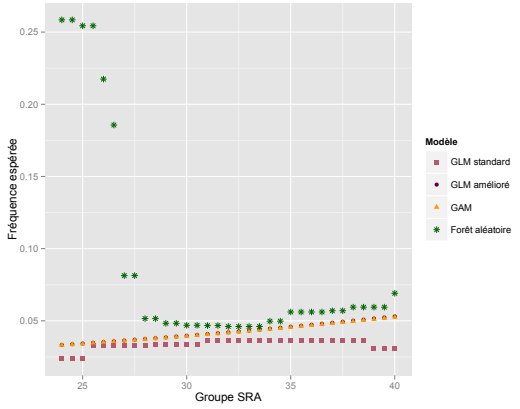
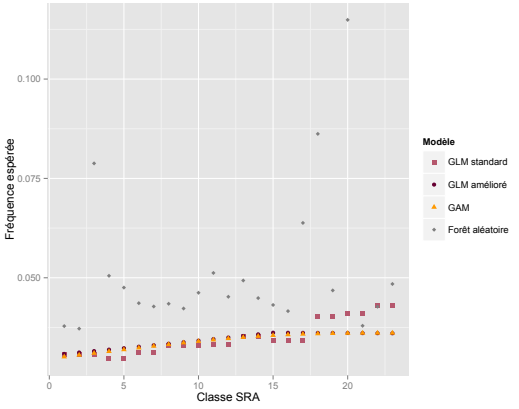


FIGURE 71 – Fréquence prédite en fonction de la `Classe SRA`



7.1.3 Sinistralité prédite pour un profil simulé

Les valeurs médianes utilisées pour la réalisation des graphiques précédents produisent parfois des profils de risque fictifs irréalistes. Par exemple, l'ancienneté du permis médiane de 13 ans qui est employée n'est pas une valeur possible pour l'ensemble des assurés âgés de moins de 31 ans ($18 + 13$). L'intérêt de recourir à un profil médian s'explique par la volonté d'observer les effets marginaux propres à une unique variables, toutes choses étant égales par ailleurs. Mais cette approche conduit parfois à étudier des profils incohérents, ce qui n'apporte pas d'information

pertinente. Afin d'analyser l'évolution de la sinistralité prédite sur l'ensemble du spectre des âges possibles, tout en conservant des valeurs cohérentes vis-à-vis des autres variables, nous proposons de tracer un graphique alternatif, pour un profil simulé réaliste. Le profil simulé correspond au vieillissement d'un jeune conducteur, ayant reçu son permis à 18 ans, et n'ayant aucun sinistre au cours de sa vie. Les variables évoluant de pair avec l'âge de ce profil sont : l'ancienneté du permis qui débute à 0, le CRM qui débute à 1 puis est multiplié chaque année par 0,95 en l'absence de sinistre jusqu'à la valeur plancher de 0,5, et le CRM précédent qui évolue en fonction. Les deux figures suivantes donnent la sévérité et la fréquence prédites de ce profil au cours du temps.

FIGURE 72 – Sévérité prédite au cours du temps pour un profil simulé réaliste

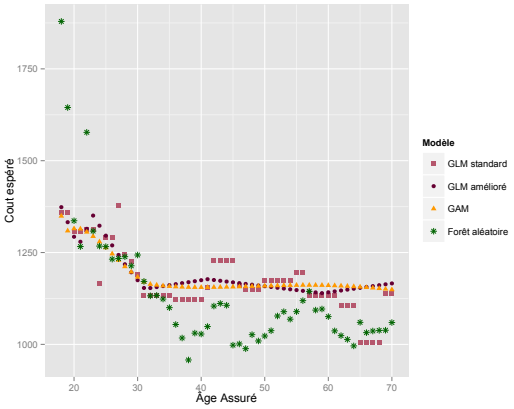
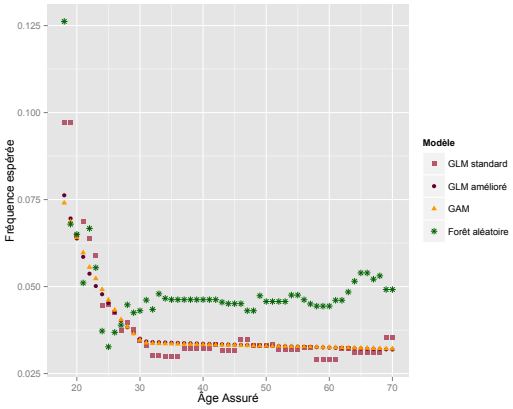


FIGURE 73 – Fréquence prédite au cours du temps pour un profil simulé réaliste



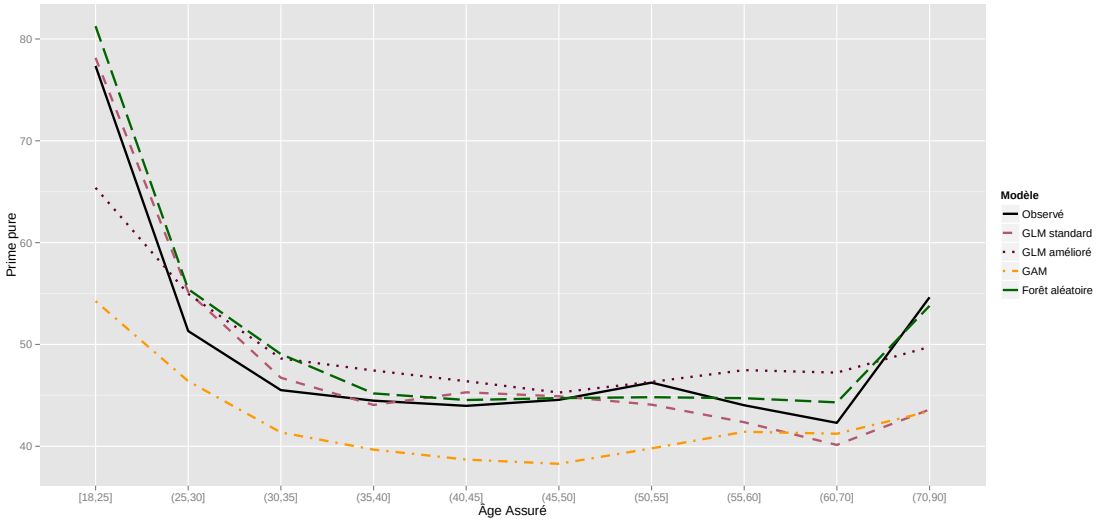
Les courbes représentées sur ces deux graphiques semblent davantage cohérentes que celles des figures précédentes, en particulier au niveau des âges extrêmes. Notons d'abord que la sinistralité semble se stabiliser au bout d'un certain temps. Cela peut s'expliquer par le fait que le CRM stagne à la valeur de 0,5 au bout de 13 années sans sinistre. Par ailleurs, de fortes divergences entre forêt aléatoire et modèles économétriques apparaissent sur l'ensemble du spectre de l'âge, ce qui démontre une nouvelle fois l'incertitude liée aux tarifs produits par cette méthode de *machine learning*. Les trois modèles économétriques semblent assez équivalents en termes de prédiction pour ce profil simulé. Néanmoins, insistons sur le fait que le gain de cohérence offert par ces graphiques s'accompagne aussi d'une perte d'information vis-à-vis des impacts marginaux individuels propres à chaque variable. De nombreux segments de risque ne sont donc pas représentés sur ces figures.

7.1.4 Prime pure globale prédite pour le portefeuille de test

Nous proposons maintenant une dernière alternative graphique susceptible de corriger les défauts des figures précédentes. Celles-ci sont limitées dans leur interprétation pour deux raisons. Premièrement, elles sont propres soit à la sévérité, soit à la fréquence des sinistres, ce qui ne permet pas d'apprécier correctement les impacts univariés sur la modélisation de la prime pure globale. Deuxièmement, elles sont valables pour un profil assuré médian, qui n'est pas réaliste pour certaines valeurs de la variable expliquée considérée. Par exemple, faire évoluer l'âge de l'assuré tout en maintenant l'ancienneté du permis à sa valeur médiane ne reflète pas la réalité du portefeuille. Afin de pallier ces deux lacunes, nous présentons ici un graphique (figure 74) susceptible de donner une intuition sur les niveaux de prime pure moyenne qui serait déterminée par chacun des mo-

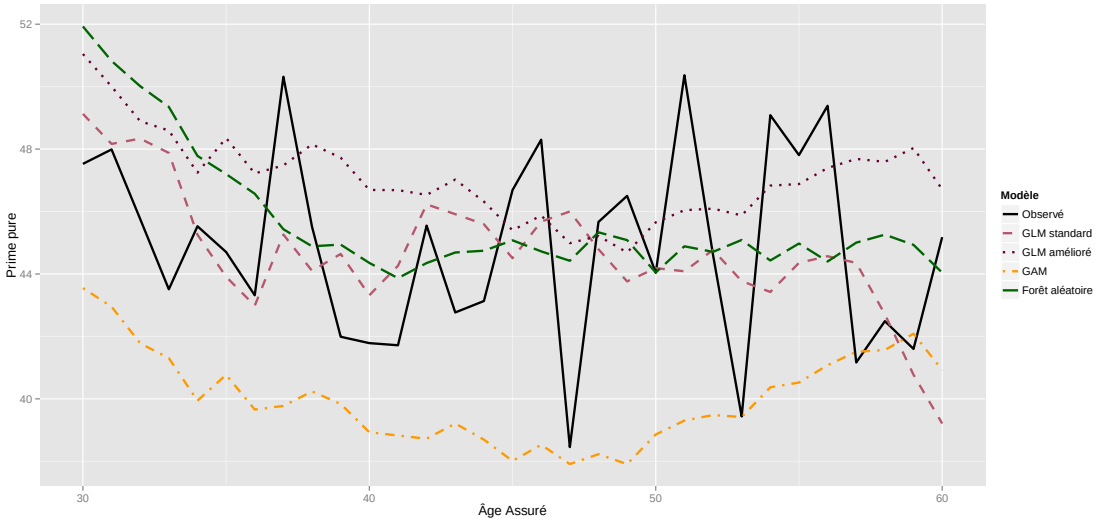
dèles pour la base de test. Cependant, les montants moyens de prime donnés ici sont spécifiques au portefeuille étudié, et ne permettent pas de formuler de conclusions générales sur les impacts imputables à l'âge. Par souci de lisibilité, les primes pures ont été moyennées par classe d'âge, de longueur 5 ans.

FIGURE 74 – Prime pure moyenne du portefeuille de test par classe d'`Âge de l'Assuré`



Ce graphique, plus proche de la réalité des tarifs, montrent que les divergences en termes d'impacts univariés pour la sévérité ou la fréquence de la sinistralité n'ont pas de conséquence démesurée sur les niveaux tarifaires. La réconciliation de la sévérité et de la fréquence d'une part, et l'application des modèles sur un portefeuille de test réel d'autre part, conduisent à produire des primes pures relativement cohérentes avec la sinistralité observée. Alors que le GAM produit des tarifs systématiquement inférieurs, les autres modèles semblent assez semblables en termes de tarification : il apparaît moins de 5 euros de différence sur la plupart des segments, pour une prime pure avoisinant les 45 euros. Cependant, il faut nuancer cette interprétation par le caractère moyenné des courbes qui masque entre autres la volatilité de la forêt aléatoire, et par la subjectivité du portefeuille considéré. Afin d'illustrer ces deux limites, il est préférable de fournir un zoom du graphique précédent sur les âges intermédiaires (30–60 ans), et en conservant la précision d'un point par âge.

FIGURE 75 – Prime pure moyenne du portefeuille de test en fonction de l'`Âge de l'Assuré`



Les courbes de ce second graphique, non moyennées par classe, illustrent parfaitement l'importante volatilité des sinistres observés sur le portefeuille de test, en comparaison des prédictions théoriques. Ce constat démontre donc l'une des limites d'une telle analyse pour laquelle l'ensemble des facteurs de risque évoluent de manière arbitraire. Le second problème, lié à la volatilité des prédictions de la forêt, n'est pas observable ici, mais cela peut s'expliquer par le fait que ces valeurs demeurent encore très agrégées, puisqu'elles correspondent à des moyennes de prime pure par âge. Toutefois, comme nous l'avons vus à travers les graphiques univariés précédents, la forêt aléatoire conserve bien un caractère irrégulier selon les effets marginaux d'une variable donnée. Comme notre problématique suppose d'établir une grille tarifaire par rapport à chacun des facteurs de risque d'intérêt, ce manque de régularité demeure un défaut invalidant cette méthode à des fins opérationnelles.

Enfin, concluons cette partie relative aux impacts univariés en précisant qu'il n'existe pas de graphique idéal, capable de retranscrire l'intégralité des effets modélisés par les différentes méthodes. Le recours à ces graphiques multiples et disparates nous semble ainsi indispensable afin d'appréhender les phénomènes en présence dans toute leur complexité.

7.2 Interactions

Les impacts univariés représentés précédemment ont démontré la nécessité de recourir à des outils graphiques plus complexes afin d'apprécier l'évolution du risque selon plusieurs facteurs déterminants clés. En effet, le caractère *ceteris paribus* des figures précédentes fournit une vision artificielle de la sinistralité pour un profil médian, qui ne reflète qu'imparfaitement la diversité des profils de risque en présence. Nous avons choisi pour chacune de deux variables expliquées (sévérité et fréquence), une interaction d'intérêt.

7.2.1 Sévérité

En premier lieu, nous nous proposons d'examiner avec plus de précisions l'interaction supputée entre `Âge de l'assuré` et `Ancienneté du permis`, dans le cadre du modèle de sévérité.

FIGURE 76 – Interactions `Âge` x `Ancienneté Permis` au sein du GLM standard (sévérité)

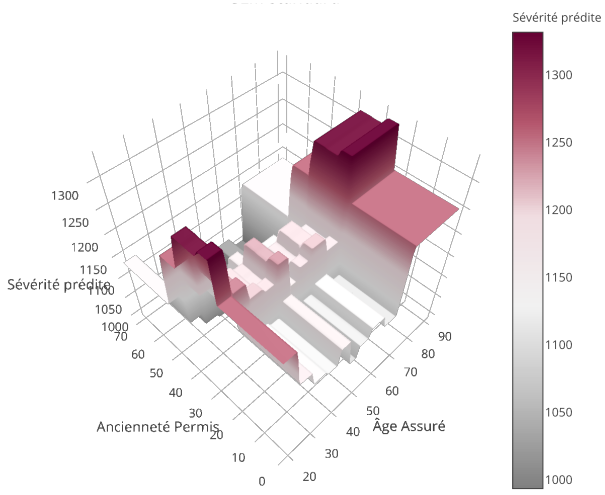


FIGURE 77 – Interactions `Âge` x `Ancienneté Permis` au sein du GLM amélioré (sévérité)

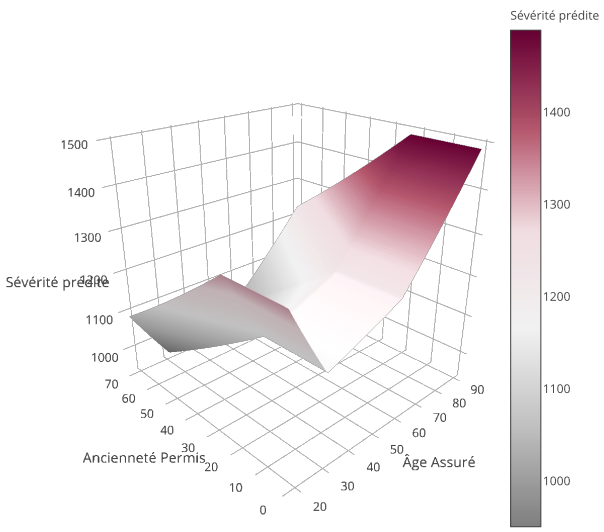
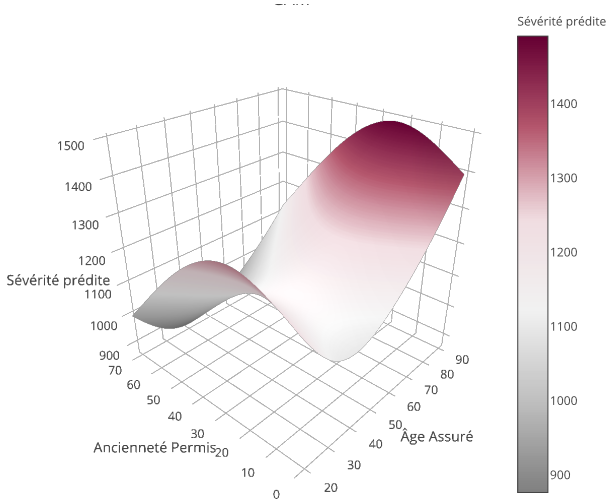
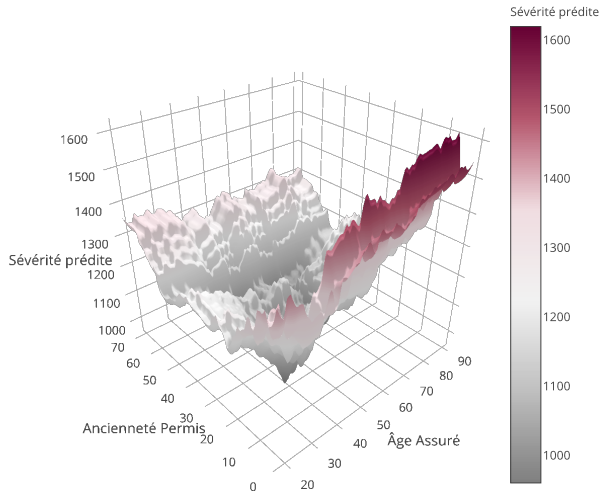


FIGURE 78 – Interactions **Âge** x **Ancienneté Permis** au sein du GAM (sévérité)



Les figures 76, 77 et 78 confirment que les trois modèles économétriques partagent globalement le même effet vis-à-vis de ces deux variables clés, comme l'indiquaient les impacts univariés analysés précédemment : un comportement parabolique selon chacune d'entre elles, avec des orientations opposées. Ces représentations bidimensionnelles n'apportent pas réellement d'information supplémentaire par rapport aux tracés unidimensionnels : elles ne correspondent qu'aux croisement des deux tendances considérées.

FIGURE 79 – Interactions **Âge** x **Ancienneté Permis** au sein du *Random Forest* (sévérité)



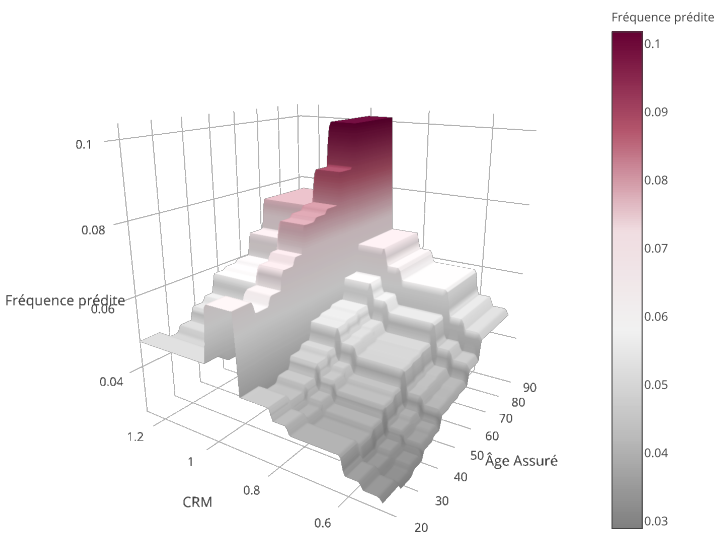
En revanche, la forêt aléatoire (figure 79) met en relief la présence d'une interaction particulière entre ces deux variables. D'une part, le caractère parabolique de l'effet de l'âge ne se retrouve que pour un profil médian vis-à-vis de l'ancienneté, à travers le « puits de risque » au centre de l'espace. D'autre part, l'âge a un impact relativement stable au niveau des segments extrêmes de l'ancienneté, comme l'atteste l'apparition de crêtes planes sur les côtés sud-est et nord-ouest du graphique. Ceci indique que l'augmentation du risque dû à l'ancienneté prédomine pour ces

cas. En particulier, si un conducteur a obtenu son permis récemment, celui-ci à une sinistralité espérée très élevée et ce, quelque soit son âge, ce qui paraît cohérent. Globalement, l'effet de l'ancienneté conserve une tendance homogène pour l'ensemble des profils de risque liés à l'âge. La forêt contribue donc à capturer certaines interactions, contrairement aux trois autres modèles, et permet ainsi **la modélisation des phénomènes multivariés complexes**.

7.2.2 Fréquence

Concernant le modèle de fréquence, nous avons retenu la représentation d'une interaction moins évidente *a priori* : le CRM d'une part, l'une des trois seules variables qui ont été converties en *splines* dans ce modèle de fréquence, et l'âge de l'assuré, qui demeure un facteur de risque majeur pour la tarification automobile.

FIGURE 80 – Interactions `CRM` x `Âge Assuré` au sein du GLM standard (fréquence)



Avec ce nouvel exemple, le GLM standard (figure 80) se démarque sensiblement avec l'apparition d'un pic de risque pour un CRM de 1. Cet effet est sans doute imputable à d'autres variables non représentées sur ce graphe, telles que l'ancienneté du contrat, car ce niveau de CRM est celui attribué aux nouveaux clients d'un portefeuille. Dans l'ensemble, les constats formulés pour le modèle de sévérité sont toujours pertinents ici : il n'y a pas d'interaction particulière modélisée par défaut par le GLM.

FIGURE 81 – Interactions `CRM` x `Âge Assuré` au sein du GLM amélioré (fréquence)

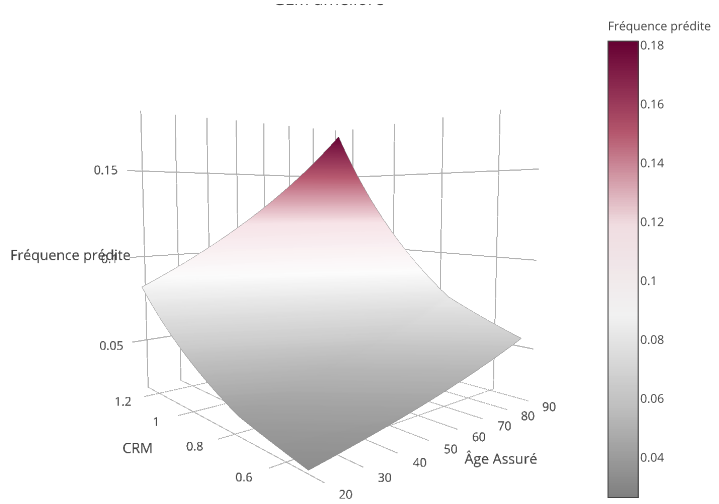
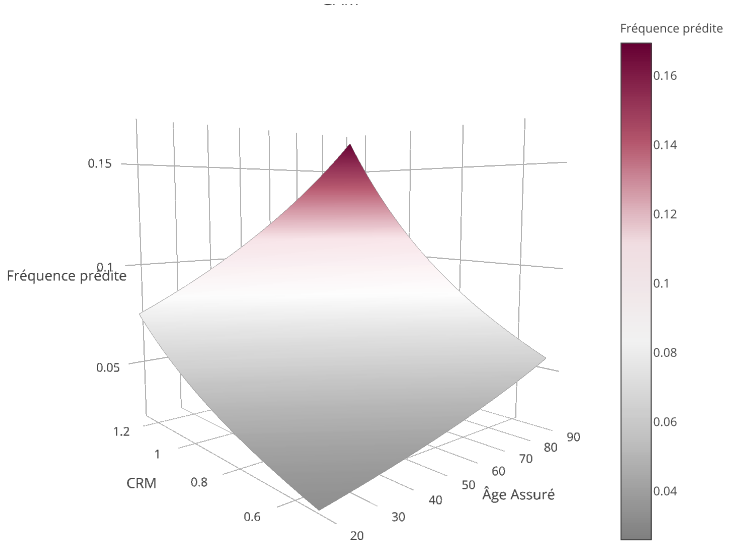
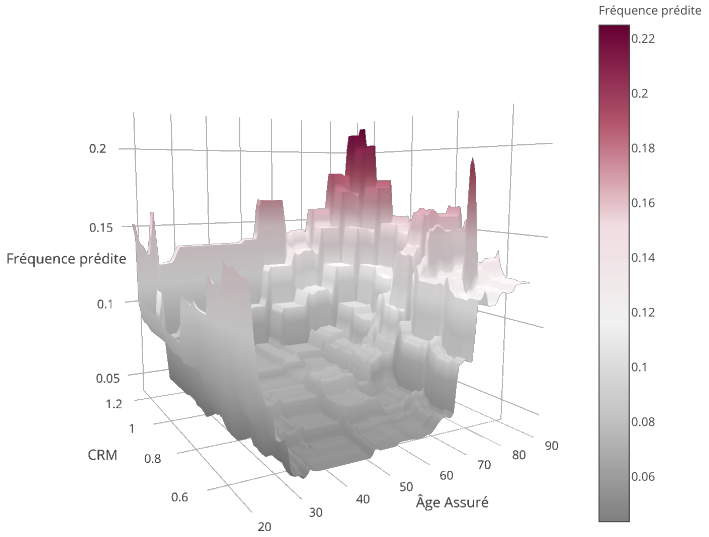


FIGURE 82 – Interactions `CRM` x `Âge Assuré` au sein du GAM (fréquence)



Les deux extensions successives du GLM (figures 81 et 82) produisent une surface de risque extrêmement lisse. Cela s'explique par l'absence de nœuds introduits lors du calibrage de l'âge de l'assuré pour le modèle de fréquence. Et les nœuds liés au CRM ne figent pas de rupture significativement de l'impact associé. En réalité, les comportement observés ici ne sont que le reflet des tendances absolument linéaires qui caractérisaient les modèles économétriques intermédiaires utilisés pour le calibrage préliminaire.

FIGURE 83 – Interactions `CRM` x `Âge Assuré` au sein du *Random Forest* (fréquence)



Enfin, le *random forest* (figure 83) démontre une nouvelle fois son aptitude à détecter des interactions non triviales. Tout d'abord, précisons que le coin sud du graphique correspond à des profils impossibles : les jeunes conducteurs débutent toujours avec un CRM de 1. Il convient donc d'ignorer cette partie de l'espace, qui présente d'ailleurs un comportement curieux. Au-delà de cette région, la forêt aléatoire prédit bien une augmentation du risque avec l'âge, ainsi qu'avec le CRM, tout comme le font les autres modèles (coin nord). Cependant, à l'inverse des modèles économétriques, cette hausse est bien plus abrupte lorsque ces deux facteurs croissent simultanément. En effet, nous constatons que le profil moyen au regard de ces deux variables – âge moyen et CRM moyen – présente une fréquence prédite au plus bas, en comparaison des autres profils potentiels. Cette particularité n'est pas reproduite par les trois autres modèles, qui affichent une pente homogène sur l'ensemble de leur surface de risque. Avec ce constat, nous pouvons conclure que les modèles économétriques ont tendance à simplement sommer, à tort, le risque dû à la combinaison de ces deux facteurs, alors que ceux-ci paraissent davantage substituables qu'additionnels. Autrement dit, selon la forêt, une augmentation de la fréquence espérée ne peut s'expliquer que par un âge élevé, ou par un CRM élevé, mais pas par l'association d'un âge moyen et d'un CRM moyen. Cette interprétation nous semble davantage réaliste, mais cet avantage ne nous paraît pas suffisant pour compenser le défaut principal de la forêt aléatoire en matière d'applicabilité opérationnelle.

Conclusion

Cette étude développe l'utilisation d'outils dépassant le cadre de la statistique traditionnelle pour répondre à divers enjeux récurrents des problématiques tarifaires en assurance non-vie. Ces différentes techniques, issues de l'apprentissage statistique, fournissent des gains de performances informatique et prédictive, et améliorent même parfois l'appréhension de l'utilisateur vis-à-vis des données étudiées, sans pour autant perdre en interprétabilité. Ces outils sont de réels atouts en pratique lors d'éventuelles révisions tarifaires pour mieux comprendre le phénomène de sinistralité et pallier d'éventuelles lacunes du modèle tarifaire existant.

La démarche proposée dans cette étude inclut en premier lieu une **méthodologie pratique de calibrage des régresseurs**, au sein du modèle GLM classique, afin de mieux prendre en compte le **caractère non linéaire** du phénomène de sinistralité. Plus précisément, cette approche emploie des *splines*, transformations polynomiales par morceaux, pour capter les effets de non-monotonie, de convexité, de seuils et d'extrêmes. De plus, le calibrage méthodique de ces *splines* a également pour objectif de mieux définir les classes tarifaires commercialisées grâce à l'identification graphique des seuils idéaux.

Ensuite, cette étude propose le recours à une procédure de pénalisation, le Lasso, pour aboutir à une **sélection de variable** permettant de réduire la dimension du modèle final. Cette méthode a été choisie pour résoudre les défauts computationnels et statistiques de la procédure classique, le *stepwise*. La technique employée constitue donc une manière additionnelle de cibler les variables tarifaires les plus importantes, tout en respectant les contraintes opérationnelles que sont le temps de calcul et la précision des tarifs calculés. Il est par exemple envisageable d'employer cette approche en tant qu'approximation du modèle principal, afin de permettre un pilotage dynamique du portefeuille automobile, dans les situations d'urgence.

Enfin, l'approche de ce mémoire se conclut par l'**analyse des interactions**, grâce à l'usage d'une forêt aléatoire, qui se distingue des modèles économétriques en intégrant automatiquement des impacts croisés entre variables explicatives. Cette spécificité de la forêt peut donc s'exploiter en pratique en amont du processus tarifaire, pour identifier les éventuelles interactions à spécifier manuellement au sein du modèle de tarification principal utilisé par la compagnie d'assurance. De manière générale, ce mémoire souhaite montrer que le soin et l'effort apportés au calibrage méthodique des variables tarifaires exploitées, au sein de modèles classiques, permet de produire des tarifs plus robustes que les méthodes automatisées sur lesquelles l'actuaire a peu d'emprise.

L'approche proposée par ce mémoire s'est accompagnée d'une **mise en œuvre pratique** sur des données réelles. Tout d'abord, nous avons comparé les résultats des différents modèles économétriques, afin d'apprécier la pertinence de l'usage de *splines* introduit ici. Les résultats montrent que les *splines* linéaires, formant ainsi un GLM dit « amélioré », fournit les meilleures performances prédictives sur base de test. Le GAM, complexifiant davantage ce modèle par l'emploi de *splines* cubiques naturelles, ne présente pas toujours les résultats escomptés. Bien qu'il talonne de près le GLM amélioré en termes de mesures macroscopiques, la flexibilité qu'il apporte apparaît souvent excessive à travers les impacts univariés. *A contrario*, le GLM standard, qui parvient péniblement à capturer des effets non linéaires grâce à la discrétisation de ses variable numériques,

conserve un **caractère haché et inégal**, impropre à la modélisation adéquate du phénomène de sinistralité. Finalement, le GLM amélioré se révèle être un bon compromis entre la souplesse des *splines* et la robustesse de la linéarité.

Ensuite, nous comparons ces résultats avec ceux de la forêt aléatoire, méthode de *machine learning*. Celle-ci offre des performances compétitives, et se démarque même considérablement en présence de variables fortement corrélées. En effet, elle modélise avec succès les impacts propres à chaque facteur de risque, sans être compromise par la manifestation d'effets croisés. Malgré ces atouts indéniables, elle présente pourtant un **comportement irrégulier**, et ne permet donc pas de produire des tarifs cohérents sans lissage supplémentaire. Sur certains segments, elle atteint même une volatilité extrême. Ces défauts nuisent donc sérieusement à la robustesse du modèle tarifaire. En effet, les contraintes opérationnelles imposées par la conduite du *business as usual* (BAU) requièrent de privilégier des **approches parcimonieuses pour des raisons d'interprétabilité**. Notamment, toute volatilité non facilement explicable n'est pas acceptable pour une direction technique qui doit rendre des comptes intelligibles auprès des comités de direction. De plus, les analyses quotidiennes réalisées par les actuaires en charge de la tarification sont bien plus exploitables lorsqu'elles se basent sur des modèles économétriques pour lesquels **les effets propres à chaque variable sont quantifiables**. C'est pourquoi nous privilégions, au terme de cette étude, l'extension intermédiaire du GLM, avec des *splines* linéaires, qui nous semble à la fois simple et mieux ajustée à la réalité de la sinistralité. Toutefois, il faut garder à l'esprit que ces résultats sont conditionnés par la **qualité des données**, propres à un seul assureur, et les opérations préliminaires réalisées sur celles-ci, telles que l'écrêtement des sinistres graves ou l'omission des valeurs manquantes.

En définitive, la démarche de calibrage par *splines*, et les outils que sont le Lasso et la forêt aléatoire, présentés dans ce mémoire, ne viennent pas concurrencer l'approche traditionnelle de tarification, mais représentent selon nous des compléments, provenant de la *data science*, pour rendre les tarifs commerciaux plus robustes. L'usage de *splines* simples permet d'identifier les points de rupture et de modéliser les effets non-linéaires tout en conservant la structure paramétrique du modèle classique. Le Lasso offre une alternative efficiente pour sélectionner les variables les plus pertinentes sans biais statistique. Et la forêt aléatoire permet de détecter les interactions dans l'optique de les spécifier au sein du modèle final. Dans l'ensemble, ces outils d'aide à la tarification peuvent améliorer la robustesse du processus tarifaire tout en demeurant **opérationnellement applicables**.

Bibliographie

[1] C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu, and B. White. Probabilistic estimation based data mining for discovering insurance risks. *IEEE Intelligent Systems*, 14 :49–58, 1999.

[2] Richard Bellman. *Adaptive control processes : a guided tour*. Princeton University Press Princeton, N.J, 1961.

[3] Noureddine Benlagha, Michel Grun-Réhomme, and Olga Vasechko. Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes. *Revue MODULAD*, 47(39), 2009.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] Thomas Bouché. Modèle de propension des assurés par rapport aux risques de sinistres corporels graves en assurance automobile. Mémoire d'actuariat, EURIA, 2014.

[6] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[7] Leo Breiman. Random forests. *Mach. Learn.*, 45(1) :5–32, October 2001.

[8] Arthur Charpentier. Exposure with binomial reponses, February 2013. URL : http://freakonometrics.hypotheses.org/3318.

[9] A. Christmann. *On a Strategy to Develop Robust and Simple Tariffs from Motor Vehicle Insurance Data*. Technical report : Sonderforschungsbereich Komplexitätsreduktion in Multivariaten Datenstrukturen. Univ., SFB 475, 2004.

[10] Arnak S. Dalalyan. Apprentissage et data mining. ENSAE ParisTech 3ème année.

[11] Carl DeBoor. A practical guide to splines. 1978.

[12] Charles Dugas, Yoshua Bengio, Nicolas Chapados, Pascal Vincent, Charles Dugas, Yoshua Bengio, Nicolas Chapados, Pascal Vincent, Germaln Denoncourt, and Christian Fournier. Statistical learning algorithms applied to automobile insurance ratemaking. *In Casualty Actuarial Society Forum-Arlington*, pages 179–213, 2003.

[13] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 04 2004.

[14] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. Technical report, Annals of Applied Statistics, 2007.

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22, 2010.

[16] Andrew Gelman and Jennifer Hill. Missing-data imputation. In *Data Analysis Using Regression and Multilevel/Hierarchical Models*, pages 529–544. Cambridge University Press, 2006. Cambridge Books Online.

[17] Guillaume Gonnet. Etude de la tarification et de la segmentation en assurance automobile. Mémoire d'actuariat, ISFA, 2010.

[18] M. Goovaerts. Beard r. e., pentikäinen t. and pesonen e. (1984). risk theory (3rd edition). chapman & hall ltd. *ASTIN Bulletin*, 15 :69–70, 4 1985.

[19] Franck Harrell. Stepwise regression problems, 1996. URL : http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction.* Springer, 2 edition, 2009.

[21] Charlotte Huther. Tarification et suivi de portefeuille en assurance non-vie : Analyse comparative de modèles de prévisions en assurance santé. Mémoire d'actuariat, ISFA, 2014.

[22] James Pickands III. Statistical Inference Using Extreme Order Statistics. *Annals of Statistics*, 3 :119–131, 1975.

[23] Alan Julian Izenman. *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning.* Springer Publishing Company, Incorporated, 1 edition, 2008.

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning : With Applications in R.* Springer Publishing Company, Incorporated, 2014.

[25] Xavier Milhaud. Segmentation et modélisation des comportements de rachat en assurance vie. Mémoire d'actuariat, ISFA, 2011.

[26] Karl P. Murphy, Michael J. Brockman, Peter K. W. Lee, Karl P Murphy, Michael J Brockman, and Peter K W Lee. 107 using generalized linear models to build dynamic pricing systems for personal lines insurance, 2000.

[27] E. Ohlsson and B. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models.* EAA Series. Springer Berlin Heidelberg, 2010.

[28] Antoine Paglia. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. Mémoire d'actuariat, EURIA, 2010.

[29] Antoine Paglia and Martial Phelippe-Guinvarc'h. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin Français d'Actuariat*, 11(22) :49–81, 2011.

[30] Virginie Pouna Siewe. Modèles additifs généralisés : Intérêts de ces modèles en assurance automobile. Mémoire d'actuariat, ISUP, 2010.

[31] Terry M. Therneau and Elizabeth J. Atkinson. An introduction to recursive partitioning using the rpart routines. divsion of biostatistics 61, 1997.

[32] Robert Tibshirani. The lasso page. URL : http://statweb.stanford.edu/~tibs/lasso.html.

[33] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58 :267–288, 1994.

Liste des tableaux

Table des figures

Liste des algorithmes

Glossaire

AIC — *Akaike Information Criterion* : Mesure de performance d'un modèle économétrique basée sur sa vraisemblance et pénalisée par le nombre de ses degrés de liberté. 45, 47–49, 51, 79, 80, 107, 110

apprentissage automatique — *machine learning* : Ensemble d'algorithmes issus de l'informatique produisant des modèles prédictifs. 16, 23, 110

apprentissage statistique — *statistical learning* : Ensemble de techniques auto-apprenantes à fondements statistiques produisant des modèles prédictifs. 3, 16, 23, 24, 37, 40, 41, 101, 110

CART — *Classification And Regression Tree* : Arbre décisionnel construit à partir d'une classification hiérarchique descendante, et permettant de classer les observations en des groupes homogènes vis-à-vis des variables explicatives. 13, 20, 26, 37–41, 109, 110

CRM — Coefficient de Réduction Majoration : Indicateur de bonus-malus, permettant de prendre en compte l'historique personnel de l'assuré en matière de sinistralité au sein de la tarification de sa prime pure. 57, 58, 63, 64, 69, 73–75, 82, 88, 90, 92, 93, 98–100, 107, 108, 110

forêt aléatoire — *random forest* : Méthode ensembliste agrégeant plusieurs arbres décisionnels décorrélés. 2, 4, 6, 13, 14, 16–18, 24–27, 41, 42, 44, 46, 54, 55, 77, 81, 82, 85–87, 89–95, 97, 100–102, 107, 109, 110, 122

GAM — *Generalized Additive Model* : Modèles généralisant les GLM par la relaxation de l'hypothèse de linéarité. 2, 4, 5, 7–9, 14, 15, 20, 23, 24, 27, 29, 32, 33, 36, 37, 40, 41, 68, 74, 75, 80, 86, 87, 89, 91, 94, 97, 99, 101, 107, 108, 110, 115, 117, 119

GLM — *Generalized Linear Model* : Modèle économétrique pseudo-linéaire recourant à une hypothèse de distribution de probabilité de la variable de sortie afin d'en estimer l'espérance. 2, 3, 5–7, 9, 14–16, 20, 21, 23–25, 27–29, 31–33, 41, 45, 53, 68–71, 74, 75, 80, 82, 85–87, 89, 91, 96, 98, 99, 101, 102, 107, 108, 110, 112, 113, 115, 117, 118

Lasso — *Least Absolute Shrinkage and Selection Operator* : Méthode de pénalisation d'un modèle économétrique ordinaire par la norme 1 des coefficients, et produisant ainsi un modèle robuste en présence de données de test indépendantes. 2, 4, 5, 8, 13, 14, 17, 25, 26, 36, 44, 46, 49–54, 77, 82–87, 101, 102, 106, 110, 122, 123

MARS *Multivariate Adaptive Regression Spline* : Procédure d'élaboration d'un modèle additif composé de splines simples et permettant les interactions. 13, 26, 27, 37, 38, 40, 41, 109, 110

modèle collectif Modèle de sinistralité basé sur une hypothèse de distribution composée fréquence-coût. 19, 28, 110

prime pure Espérance de la charge totale de sinistres. 14, 18, 19, 27, 28, 86, 93–95, 107, 110

spline Transformation polynomiale par morceaux respectant des contraintes de régularité utile en régression linéaire. 3, 4, 6–8, 13, 15, 18, 23, 27, 33–38, 40, 41, 50, 68, 69, 74, 75, 82, 89, 98, 101, 102, 106, 110, 113, 115

SRA Sécurité et Réparation Automobiles : Organisme de sécurité routière recensant et classifiant l'ensemble des véhicules existant selon des critères de dangerosité et de coût de remplacement/réparation. 57–60, 63, 64, 66, 67, 69–71, 73, 75, 82, 88, 89, 92, 107, 110

stepwise Procédure itérative de sélection de variables. 4, 13, 14, 24–26, 37, 44, 46–50, 54, 77, 79–81, 84, 101, 107, 109, 110, 119, 122

A Graphiques complémentaires

A.1 Modèles intermédiaires du GLM amélioré

Les figures 84 à 91 représentent les courbes des coefficients estimés à partir des modèles intermédiaires, conduisant à bâtir le GLM amélioré. Ces modèles intermédiaires ont été calibrés sur les variables discrétisées arbitrairement, selon la pratique commune.

A.1.1 Sévérité

FIGURE 84 – Coefficients associés à l'`Ancienneté du permis` (modèle de coût)

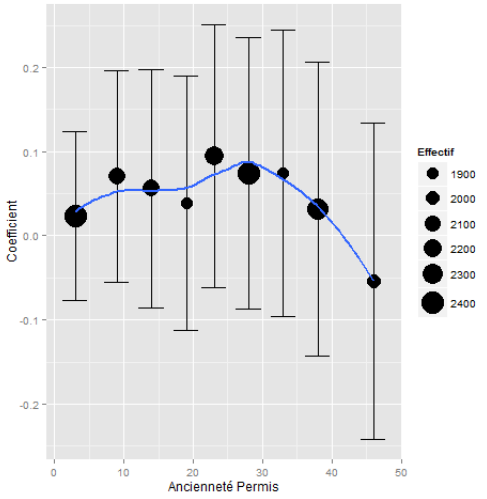


FIGURE 85 – Coefficients associés à la `Réparation SRA` (modèle de coût)

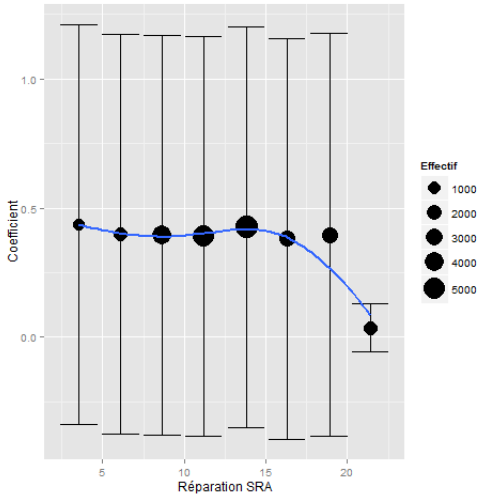


FIGURE 86 – Coefficients associés au `CRM` (modèle de coût)

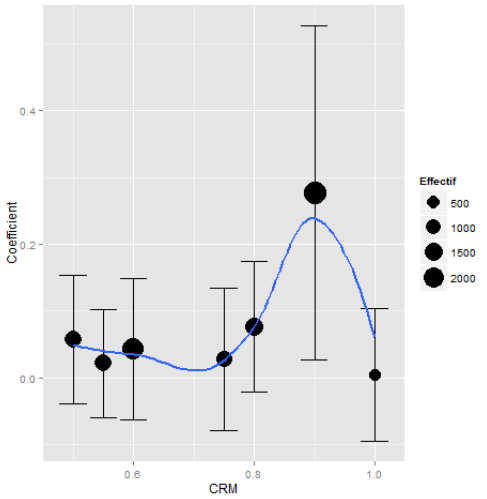
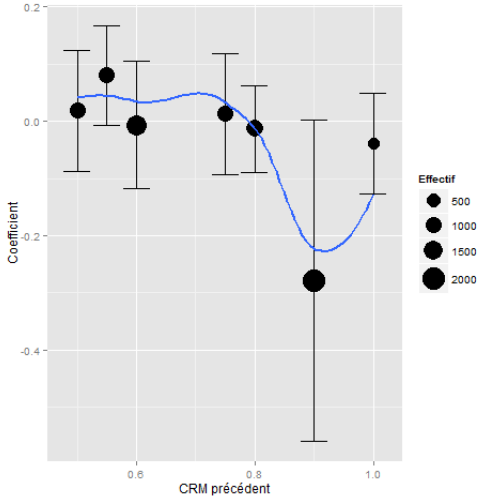


FIGURE 87 – Coefficients associés au `CRM précédent` (modèle de coût)



A.1.2 Fréquence

FIGURE 88 – Coefficients associés à l'`Ancienneté du permis` (modèle de fréquence)

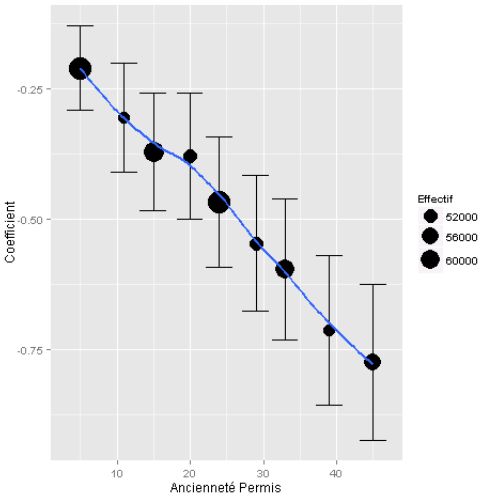


FIGURE 89 – Coefficients associés à la `Réparation SRA` (modèle de fréquence)

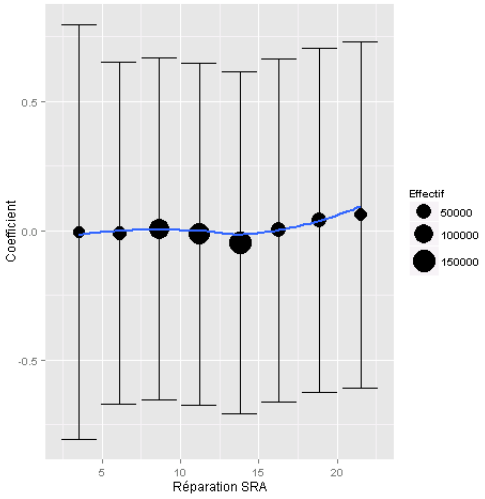


FIGURE 90 – Coefficients associés au `CRM` (modèle de fréquence)

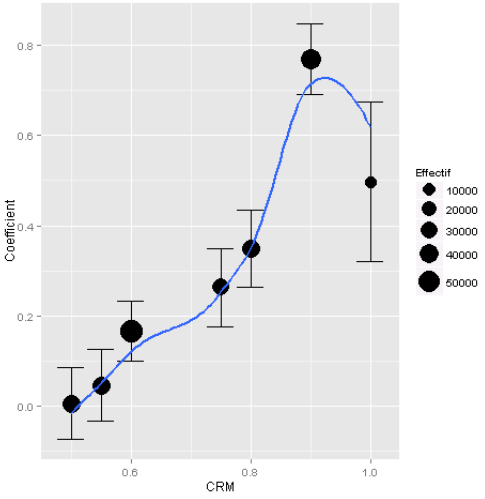
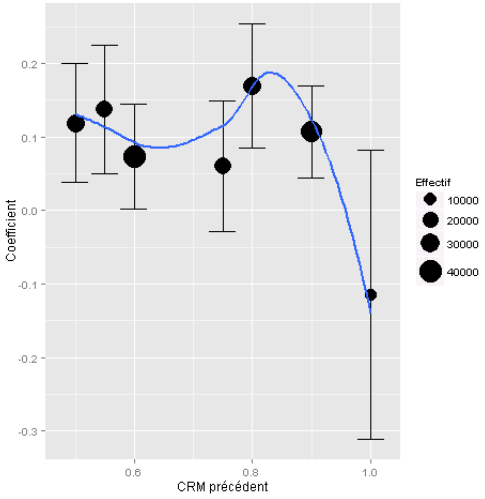


FIGURE 91 – Coefficients associés au `CRM précédent` (modèle de fréquence)



A.2 *Splines* du GLM amélioré

Les figures 92 à 97 représentent les *splines* linéaires produites par le modèle GLM amélioré, calibré sur les variables discrétisées méthodiquement, selon l'approche exposée en partie 5.2.

FIGURE 92 – *Spline* linéaire associée à l'`Ancienneté du permis` (modèle de coût)

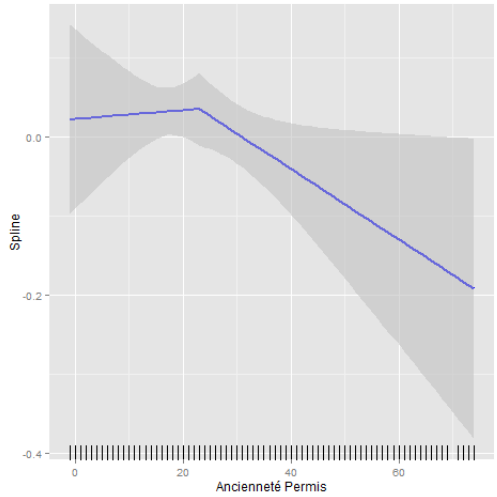


FIGURE 93 – *Spline* linéaire associée à la `Réparation SRA` (modèle de coût)

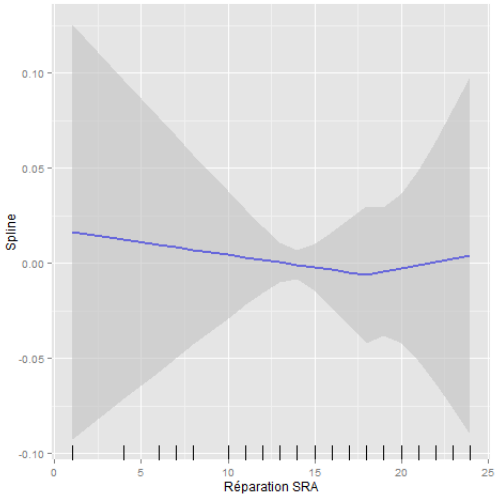


FIGURE 94 – *Spline* linéaire associée à la `Classe SRA` (modèle de coût)

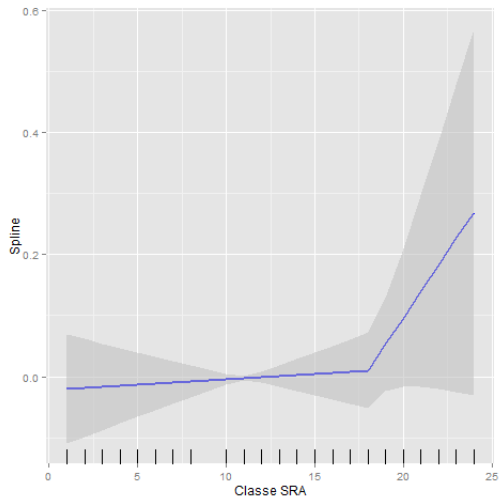


FIGURE 95 – *Spline* linéaire associée au `Groupe SRA` (modèle de coût)

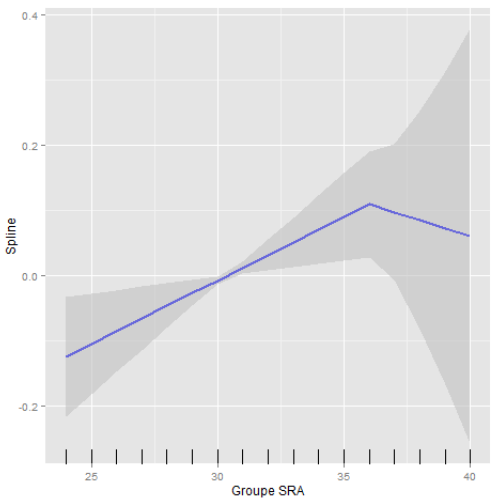


FIGURE 96 – *Spline* linéaire associée au `CRM` (modèle de coût)

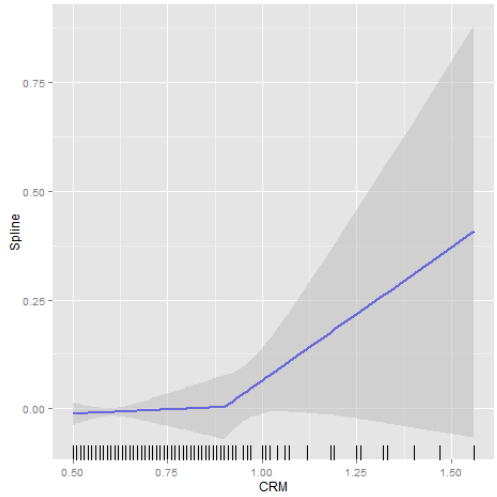
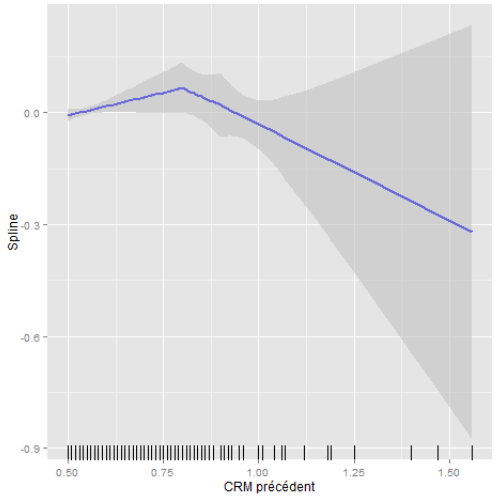


FIGURE 97 – *Spline* linéaire associée au `CRM précédent` (modèle de coût)



A.3 *Splines* du GAM

Les figures 98 à 103 représentent les *splines* cubiques naturelles produites par le modèle GAM, calibré sur la même discrétisation des variables que pour le GLM amélioré.

FIGURE 98 – *Spline* cubique naturelle associée à l'`Ancienneté du permis` (modèle de coût)

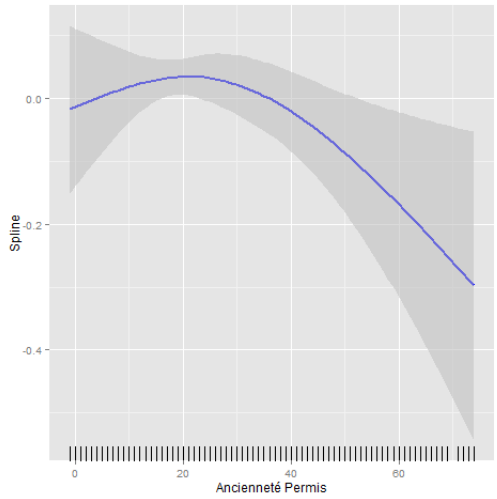


FIGURE 99 – *Spline* cubique naturelle associée à la `Réparation SRA` (modèle de coût)

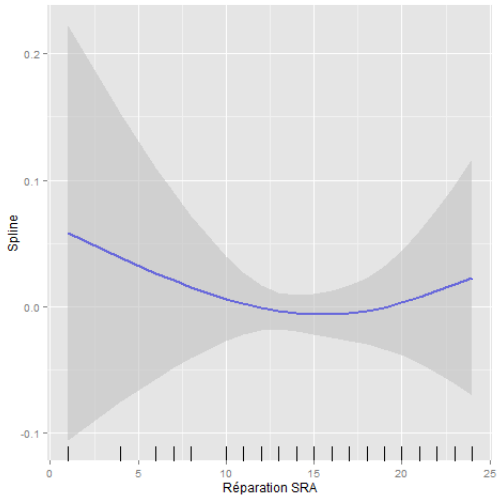


FIGURE 100 – *Spline* cubique naturelle associée à la `Classe SRA` (modèle de coût)

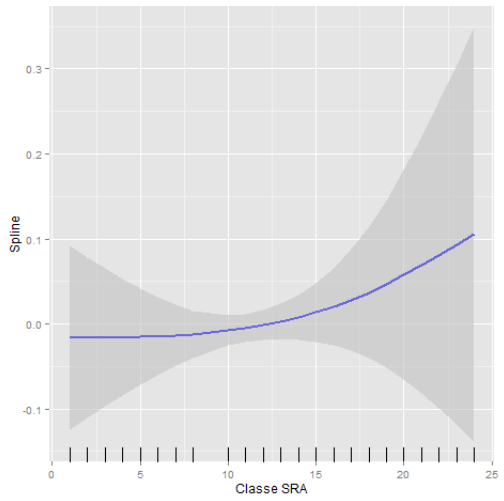


FIGURE 101 – *Spline* cubique naturelle associée au `Groupe SRA` (modèle de coût)

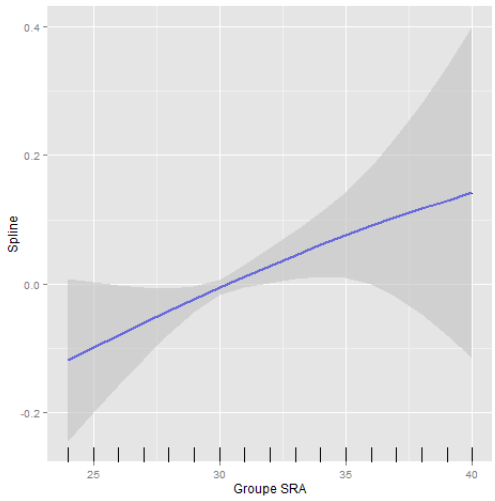


FIGURE 102 – *Spline* cubique naturelle associée au `CRM` (modèle de coût)

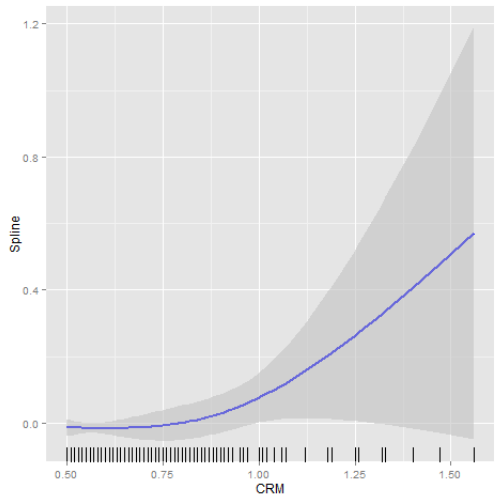
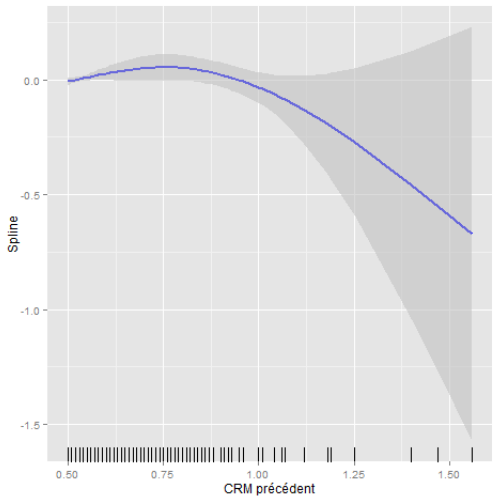


FIGURE 103 – *Spline* cubique naturelle associée au `CRM précédent` (modèle de coût)



B Interprétation des coefficients de régression

L'interprétation des résultats produits par les modèles généralisés est toutefois non triviale. En effet, les coefficients estimés dans le cadre de ces régressions doivent se rapporter à l'échelle de la variable linéaire sous-jacente Y_i^*, aussi appelée **variable latente** :

$$Y_i^* = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \tag{42}$$

Mais l'utilisation d'une fonction de lien non identitaire transforme cette structure linéaire et interdit alors d'interpréter directement ces coefficients comme cela est possible en régression linéaire classique. Plus précisément, ces coefficients ne sont plus les effets *ceteris paribus*[25] sur la variable

[25] Toutes choses étant égales par ailleurs.

réponse d'une variation élémentaire des prédicteurs auxquels ils correspondent. Dans cette situation, trois approches alternatives permettent d'interpréter les résultats des modèles GLM et GAM de manière appropriée :

- les **effets marginaux** permettent de quantifier l'impact additif d'une **variation infinitésimale** des prédicteurs, et sont donc l'équivalent des coefficients de la régression classique pour les modèles généralisés
- les **effets additifs** correspondent à l'impact additif sur la variable réponse d'une **variation unitaire** des prédicteurs
- les **effets multiplicatifs** correspondent à l'impact multiplicatif sur la variable réponse d'une **variation unitaire** des prédicteurs, et ils coïncident avec le rapport des cotes (*odds ratio*) pour les régresseurs binaires

B.1 Effets additifs

Plus formellement, les effets marginaux correspondent à la dérivée partielle (empirique) de la variable prédite $\hat{Y}_i$ par rapport à chacune des variables explicatives X_{ij}. En dérivant les deux membres de l'équation 4 du modèle GLM avec une fonction de lien logarithmique, il vient :

$$\frac{\partial \mathbb{E}\left[Y_i|X_i\right]}{\partial X_{ij}} = \beta_j \mathrm{e}^{X_i^t \beta} = \beta_j \hat{Y}_i \tag{43}$$

Cette quantité dépend de la variable de dérivation X_{ij} et du point de calcul $(x_{i1}, \ldots, x_{ip})$. Il est donc possible de déterminer jusqu'à $n \times p$ valeurs correspondant aux effets marginaux de chaque variable pour chaque observation. Il est néanmoins courant de calculer ces effets marginaux soit en un point moyen ou médian de la population étudiée, soit en les moyennant sur toute la population.

Cependant, cette définition n'a de réel sens que pour les variables quantitatives, puisqu'une variation infinitésimale selon le sexe est difficilement concevable. Mais cette notion peut être étendue aux variables qualitatives en calculant l'effet additif *ceteris paribus* d'une variation unitaire de l'indicatrice considérée, c'est-à-dire l'effet de passer de la modalité de référence (codée 0) à la modalité d'intérêt (codée 1) :

$$\begin{aligned}\mathbb{E}[Y_i|X_{i1}, \ldots, X_{ij}+1, \ldots, X_{ip}] - \mathbb{E}[Y_i|X_{i1}, \ldots, X_{ij}, \ldots, X_{ip}] &= \mathrm{e}^{X_i^t \beta + \beta_j} - \mathrm{e}^{X_i^t \beta} \\ &= \left(\mathrm{e}^{\beta_j} - 1\right) \cdot \hat{Y}_i \end{aligned} \tag{44}$$

Ces deux outils sont particulièrement populaires en régression logistique, qui vise à modéliser une probabilité. Dans ce cadre binaire, il parait donc plus naturel d'évaluer les effets additifs sur cette probabilité, puisque les valeurs prises par celle-ci sont bornées entre 0 et 1. En revanche, avec nos modèles de tarification, il est bien plus efficace d'exploiter la forme naturellement multiplicative des modèles employant un lien logarithmique. C'est donc cette dernière approche que nous allons privilégier dans nos interprétations ultérieures.

B.2 Effets multiplicatifs

Le caractère multiplicatif de nos modèles tarifaires garantit que les impacts multiplicatifs ont une forme très simple :

$$\frac{\mathbb{E}[Y_i|X_{i1},\ldots,X_{ij}+1,\ldots,X_{ip}]}{\mathbb{E}[Y_i|X_{i1},\ldots,X_{ij},\ldots,X_{ip}]} = \frac{\mathrm{e}^{X_i^t\beta+\beta_j}}{\mathrm{e}^{X_i^t\beta}} = \mathrm{e}^{\beta_j} \tag{45}$$

Il suffit donc d'appliquer la fonction exponentielle aux coefficients issus du modèle de régression pour obtenir ces impacts multiplicatifs sur la sinistralité espérée.

Comment interprète-t-on les erreurs standards et intervalles de confiance associés ? Tout comme les estimateurs, les intervalles de confiance peuvent être adaptés au canevas multiplicatif grâce à la transformation exponentielle de leurs bornes. Précisons néanmoins que cette dernière produit toutefois des **intervalles asymétriques**. Il est alors possible de retrouver des erreurs standards équivalentes, mais le calcul est subtil et le résultat peu informatif. À titre d'indice de significativité, nous avons employé le critère qui indique si la valeur 1 est contenue dans l'intervalle de confiance, ce qui correspond à un effet multiplicatif non statistiquement différent de l'effet neutre. Dans ce sens, sont représentés ci-après les impacts multiplicatifs relatifs à l'ensemble des variables explicatives et les intervalles de confiance correspondants, pour chacun des trois modèles économétriques élaborés.

FIGURE 104 – Effets multiplicatifs du GLM standard (sévérité)

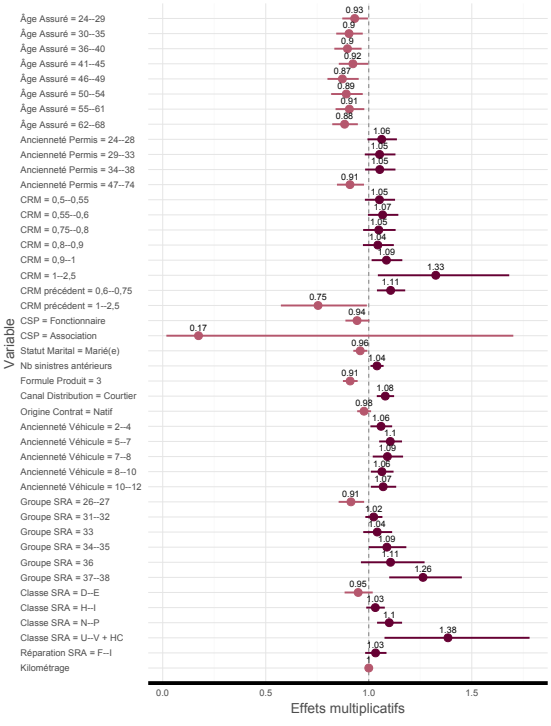


FIGURE 105 – Effets multiplicatifs du GLM amélioré (sévérité)

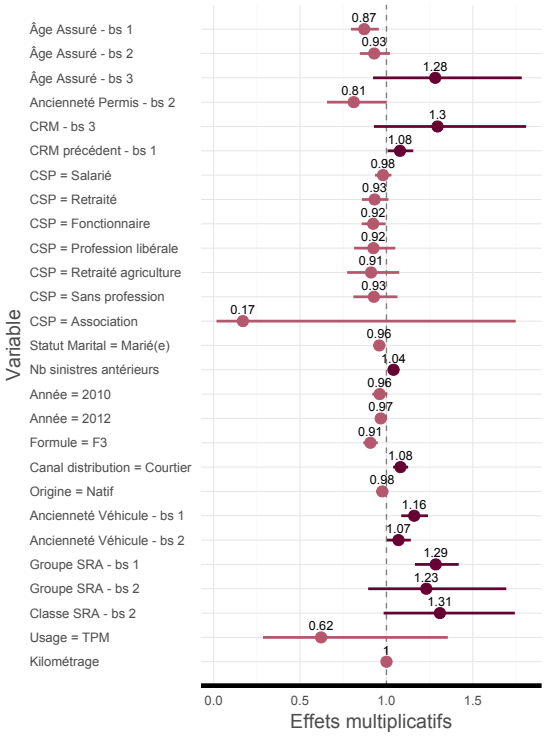
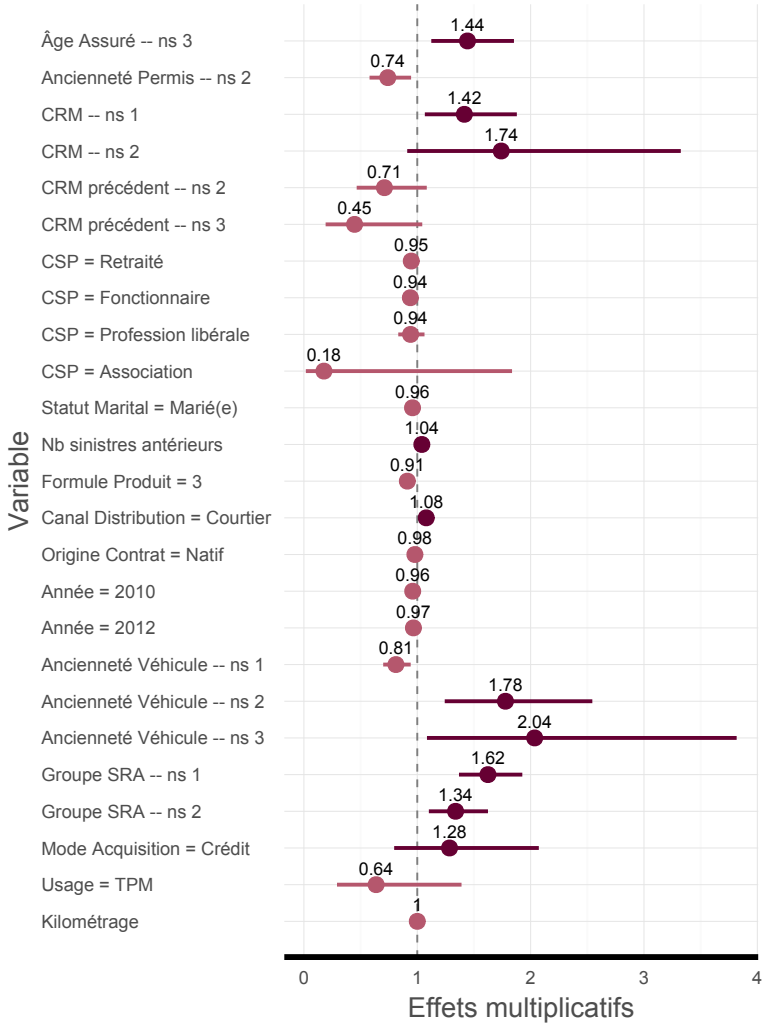


FIGURE 106 – Effets multiplicatifs du GAM (sévérité)



B.3 Tableau des coefficients

Le tableau 11 présente les coefficients estimés par les trois modèles économétriques optimaux (GLM standard, GLM amélioré, et GAM), à l'issue de la sélection de variables par le *stepwise*. Le lecteur remarquera pour les deux derniers modèles que les variables comportant la mention *bs* correspondent à des *basis splines* ou B-*splines*, soit des constituants élémentaires d'une *spline* linéaire, et les variables comportant la mention *ns* correspondent à des *natural splines*, soit des constituants élémentaires d'une *spline* cubique naturelle.

TABLE 11 – Effets multiplicatifs des modèles économétriques (variable de coût)

Variable	GLM standard	GLM amélioré	GAM
Âge Assuré = 24–29	0.93 [0.87; 1.00]*		
Âge Assuré = 30–35	0.90 [0.84; 0.97]*		
Âge Assuré = 36–40	0.90 [0.83; 0.97]*		
Âge Assuré = 41–45	0.92 [0.85; 1.00]*		

Variable	GLM standard	GLM amélioré	GAM
Âge Assuré = 46–49	0.87 [0.80; 0.95]*		
Âge Assuré = 50–54	0.89 [0.82; 0.97]*		
Âge Assuré = 55–61	0.91 [0.84; 0.98]*		
Âge Assuré = 62–68	0.88 [0.82; 0.95]*		
Âge Assuré – bs 1		0.87 [0.79; 0.96]*	
Âge Assuré – bs 2		0.93 [0.85; 1.02]	
Âge Assuré – bs 3		1.28 [0.92; 1.78]	
Âge Assuré – ns 3			1.44 [1.12; 1.85]*
Ancienneté Permis = 24–28	1.06 [0.99; 1.14]		
Ancienneté Permis = 29–33	1.05 [0.98; 1.13]		
Ancienneté Permis = 34–38	1.05 [0.98; 1.13]		
Ancienneté Permis = 47–74	0.91 [0.85; 0.98]*		
Ancienneté Permis – bs 2		0.81 [0.66; 1.00]	
Ancienneté Permis – ns 2			0.74 [0.58; 0.95]*
CRM = 0,5–0,55	1.05 [0.98; 1.13]		
CRM = 0,55–0,6	1.07 [1.00; 1.14]		
CRM = 0,75–0,8	1.05 [0.97; 1.13]		
CRM = 0,8–0,9	1.04 [0.97; 1.12]		
CRM = 0,9–1	1.09 [1.01; 1.16]*		
CRM = 1–2,5	1.33 [1.04; 1.68]*		
CRM précédent = 0,6–0,75	1.11 [1.04; 1.18]*		
CRM précédent = 1–2,5	0.75 [0.57; 0.99]*		
CRM – bs 3		1.30 [0.93; 1.81]	
CRM précédent – bs 1		1.08 [1.01; 1.16]*	
CRM – ns 1			1.42 [1.07; 1.88]*
CRM – ns 2			1.74 [0.91; 3.32]
CRM précédent – ns 2			0.71 [0.47; 1.08]
CRM précédent – ns 3			0.45 [0.19; 1.04]
CSP = Fonctionnaire	0.94 [0.89; 1.00]	0.92 [0.86; 0.99]*	0.94 [0.88; 1.00]
CSP = Association	0.17 [0.02; 1.70]	0.17 [0.02; 1.75]	0.18 [0.02; 1.84]
CSP = Salarié		0.98 [0.93; 1.03]	
CSP = Retraité		0.93 [0.86; 1.01]	0.95 [0.89; 1.01]
CSP = Profession libérale		0.92 [0.81; 1.05]	0.94 [0.83; 1.06]
CSP = Retraité agriculture		0.91 [0.77; 1.07]	
CSP = Sans profession		0.93 [0.81; 1.06]	
Statut Marital = Marié(e)	0.96 [0.93; 0.99]*	0.96 [0.92; 0.99]*	0.96 [0.92; 0.99]*
Nb sinistres antérieurs	1.04 [1.01; 1.07]*	1.04 [1.01; 1.07]*	1.04 [1.01; 1.07]*
Formule Produit = 3	0.91 [0.87; 0.95]*	0.91 [0.87; 0.95]*	0.91 [0.87; 0.96]*
Canal Distribution = Courtier	1.08 [1.04; 1.12]*	1.08 [1.04; 1.13]*	1.08 [1.04; 1.12]*
Origine Contrat = Natif	0.98 [0.94; 1.01]	0.98 [0.94; 1.01]	0.98 [0.94; 1.01]
Année = 2010		0.96 [0.92; 1.00]	0.96 [0.92; 1.00]

Variable	GLM standard	GLM amélioré	GAM
Année = 2012		0.97 [0.93; 1.00]	0.97 [0.93; 1.00]
Ancienneté Véhicule = 2–4	1.06 [1.01; 1.11]*		
Ancienneté Véhicule = 5–7	1.10 [1.05; 1.16]*		
Ancienneté Véhicule = 7–8	1.09 [1.02; 1.17]*		
Ancienneté Véhicule = 8–10	1.06 [1.01; 1.12]*		
Ancienneté Véhicule = 10–12	1.07 [1.01; 1.13]*		
Ancienneté Véhicule – bs 1		1.16 [1.09; 1.24]*	
Ancienneté Véhicule – bs 2		1.07 [1.00; 1.14]*	
Ancienneté Véhicule – ns 1			0.81 [0.70; 0.94]*
Ancienneté Véhicule – ns 2			1.78 [1.24; 2.54]*
Ancienneté Véhicule – ns 3			2.04 [1.09; 3.82]*
Groupe SRA = 26–27	0.91 [0.85; 0.98]*		
Groupe SRA = 31–32	1.02 [0.98; 1.07]		
Groupe SRA = 33	1.04 [0.97; 1.11]		
Groupe SRA = 34–35	1.09 [1.00; 1.18]*		
Groupe SRA = 36	1.11 [0.96; 1.27]		
Groupe SRA = 37–38	1.26 [1.10; 1.45]*		
Groupe SRA – bs 1		1.29 [1.16; 1.42]*	
Groupe SRA – bs 2		1.23 [0.89; 1.69]	
Groupe SRA – ns 1			1.62 [1.37; 1.93]*
Groupe SRA – ns 2			1.34 [1.10; 1.62]*
Classe SRA = D–E	0.95 [0.88; 1.02]		
Classe SRA = H–I	1.03 [0.99; 1.08]		
Classe SRA = N–P	1.10 [1.04; 1.16]*		
Classe SRA = U–V + HC	1.38 [1.08; 1.78]*		
Classe SRA – bs 2		1.31 [0.98; 1.74]	
Réparation SRA = F–I	1.03 [0.98; 1.09]		
Mode Acquisition = Crédit			1.28 [0.80; 2.07]
Usage = TPM		0.62 [0.29; 1.36]	0.64 [0.29; 1.39]
Kilométrage	1.00 [1.00; 1.00]*	1.00 [1.00; 1.00]	1.00 [1.00; 1.00]*
AIC	351114.01	351177.42	351171.82
BIC	351481.27	351408.96	351387.38
Log-vraisemblance	−175511.01	−175559.71	−175558.91
Déviance	16633.89	16700.65	16699.55
Nb observations	21673	21673	21673

∗ 1 hors de l'intervalle de confiance

C Compléments sur la sélection de variables

Les tableaux suivants viennent compléter les résultats présentés au regard des variables sélectionnées par les trois techniques étudiées dans ce mémoire. La liste des facteurs de risque retenus par le *stepwise* et le Lasso est rééditée ici, dans l'ordre de significativité – c'est-à-dire à l'aide de la *p-value* – pour la première approche, les hypothèses propres au Lasso ne permettant pas de produire un test d'importance pertinent. Enfin, les variables de la troisième colonne ont également été ordonnées selon le critère d'importance de la forêt aléatoire (*random forest*), uniquement à des fins de comparaison puisque cette dernière approche de sélection n'a pas été réellement suivie d'une seconde étape de régression économétrique. Rappelons d'ailleurs que la forêt aléatoire permet bien d'effectuer une sélection de variables, à condition de spécifier un seuil d'importance filtrant les régresseurs. Ici le seuil peu filtrant de 0 a été considéré, ce qui a permis d'exclure une poignée de facteurs. Les niveaux d'importance sont visualisables à travers les figures 58 et 59.

TABLE 12 – Sélection de variables ordonnée pour la sévérité (RC Corporelle)

	Stepwise	Lasso	*Random Forest*
1	Âge	Âge	Âge
2	Formule = F3	Classe SRA	Ancienneté Permis
3	Ancienneté Permis	Origine = Natif	Poids
4	Groupe SRA	CSP = Etudiant	Réparation SRA
5	Classe SRA	CSP = Profession annexe de l'agriculture	Groupe SRA
6	Origine = Natif	Statut Marital = Inconnu	Classe SRA
7	Réparation SRA		CRM
8	CSP = Agriculteur		Statut Marital
9	CRM précédent		Ancienneté Véhicule
10	Formule = F2		CRM précédent
11	CSP = Etudiant		Code CSP
12	Formule = F1		Nb Places
13	Ancienneté Véhicule		Formule
14	Statut Marital = Inconnu		Nb contrats connexes
15	Usage = Tournées/Livraison		Pack Contenu
16	CSP = Profession libérale moins de 10 salariés		Pack Indemnité
17	Canal = Autre		Origine
18	CSP = Artisan		Nb sinistres antérieurs
19	Mode d'acquisition = Crédit		Canal
20			AAC
21			Pack Mobilité
22			Energie

TABLE 13 – Sélection de variables ordonnée pour la sévérité (RC Matérielle)

	Stepwise	Lasso	*Random Forest*
1	Groupe SRA	Âge	Ancienneté Permis
2	Ancienneté Véhicule	Ancienneté Permis	Groupe SRA
3	Formule = F3	Ancienneté Véhicule	Âge
4	Canal = Courtier	Groupe SRA	CRM
5	Âge	Classe SRA	Classe SRA
6	Nb sinistres antérieurs	CRM	Poids
7	Statut Marital = Marié(e)	CRM précédent	Ancienneté Véhicule
8	CRM précédent	Année = 2010	CRM précédent
9	CSP = Fonctionnaire et assimilé	Formule = F3	Carrosserie
10	Kilométrage	Canal = Courtier	Formule
11	Classe SRA	Origine = Natif	Statut Marital
12	Année = 2012	Kilométrage	Réparation SRA
13	CSP = Retraité	CSP = Fonctionnaire et assimilé	Origine
14	CRM	CSP = Artisan	Nb sinistres antérieurs
15	CSP = Association	CSP = Agriculteur	Code CSP
16	Origine = Natif	Statut Marital = Marié(e)	Energie
17	Usage = TPM	Nb sinistres antérieurs	Pack Contenu
18	CSP = Retraité profession de l'agriculture		Nb contrats connexes
19	CSP = Sans profession		Nb places
20	CSP = Salarié		Mode Paiement
21			Canal
22			Année
23			Kilométrage
24			Usage
25			Pack Indemnité
26			Mode d'acquisition