Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 22/02/2017

Par : **Charles Tremblay**

Titre : **Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires.**

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

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

Membres présents du jury de la filière

Entreprise : PACIFICA
Nom : Lionel Féraud
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Laura Candas
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 :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
3, Avenue Pierre Larousse - 92245, MALAKOFF CEDEX, FRANCE

Résumé

Prédire le risque extrême est stratégique : il est d'occurence rare et sa modélisation difficile, mais il constitue un levier tarifaire majeur. Ce mémoire étudie le risque extrême du produit d'assurance professionnelle multirisque de PACIFICA - une filiale de Crédit Agricole Assurances - au cours de la période 2009-2015. Nous cherchons à prédire le risque grave lié aux incendies, avec trois objectifs : construire un modèle de prime pure grave, analyser les déterminants du risque grave pour développer la prévention, et cibler des profils de haut risque. Les modèles linéaires répondent aux deux premiers objectifs ; le troisième fait appel aux arbres de classification et de régression.

L'étude des valeurs extrêmes nous conduit à définir les graves comme des sinistres de montant supérieur à 70 000 €. Nos données sont de faible volume, et marquées par un fort déséquilibre : les incendies graves sont très rares. Pour pallier ce problème, nous utilisons un algorihme de suréchantillonnage synthétique, le SMOTE (*Synthetic Minority Over-sampling TEchnique*), qui développe les capacités d'apprentissage des arbres et des modèles linéaires par la génération aléatoire, dans l'espace des variables explicatives, d'observations sinistrées fictives, localisées entre deux sinistres graves réels.

Nous testons six modèles distincts : trois modèles linéaires (régression logistique, régression logistique modifiée et régression de Poisson) et trois algorithmes d'apprentissage statistique supervisé (arbre de classification, random forest, arbre de régression de Poisson). Les performances de ces modèles sont évaluées par l'AUC mesuré sur une base de test. Les modèles de fréquence - régression de Poisson et arbre de régression de Poisson - sont les plus performants, du fait de leur meilleure capacité à prendre en compte les variations d'exposition. La sélection des variables au sein de la régression de Poisson est assurée par la pénalisation LASSO (*Least Absolute Shrinkage and Selection Operator*), qui présente plusieurs avantages par rapport au stepwise : elle est rapide, autonome, et possède de bonnes propriétés statistiques.

Le suréchantillonnage synthétique améliore considérablement la performance des modèles. Nous atteignons un AUC de 70% pour le modèle linéaire. Le faible volume de nos données rend la sélection des variables peu stable, mais nous pouvons interpréter les plus importantes : la sinistralité attritionnelle antérieure apparaît ainsi très explicative du risque grave. L'arbre atteint après optimisation un AUC de 74%. Il s'avère plus sensible aux paramètres du suréchantillonnage que le modèle linéaire, du fait de sa grande adaptabilité à la structure des données. L'arbre permet d'identifier un groupe de haut risque au sein duquel la fréquence de graves est plus de 20 fois supérieure à la fréquence globale, et ce en moins d'une dizaine de règles de décision. Il est également en mesure de dégager des critères simples et discriminants sur de grands volumes : en seulement deux critères isole deux grands groupes tarifaires, l'un 7 fois plus risqué que l'autre.

Arbres et modèles linéaires n'apparaissent pas seulement complémentaires dans cette étude, mais en interaction. Le modèle linéaire est la meilleure méthode pour retraiter intelligemment des variables catégorielles hétérogènes, qui ne peuvent être utilisées telles quelles par les arbres. En retour, les arbres offrent une perspective non-linéaire sur les données, qui constitue un bon outil pour améliorer la spécification du modèle.

Abstract

Predicting extreme risk is a strategical issue : it rarely occurs and its modelisation is difficult, but it constitutes a major pricing lever. This study deals with extreme risk on the product of comprehensive professionnal insurance of PACIFICA, a subsidiary of the Credit Agricole - during the period 2009-2015. We aim at predicting the serious risk associated with fires, with three objectives : setting up a model of pure premium for severe losses, analysing the determinants of severe losses to develop prevention programs, and targeting high risk profiles. Linear models meet the first two objectives while the third one requires the use of classification and regression trees.

The study of extreme values leads us to define severe damages as damages which cost exceeds 70 k€. Our data are of low volume and marked by a strong unbalance ; indeed severe fires are rare. To compensate for this difficulty, we use an algorithm of synthetic over-sampling, the SMOTE (Synthetic Minority Over-sampling Technique) which develops the training faculty of trees and of linear models through the random generation, in the explanatory variables space, of fictitious damage observations, located between two real severe damages.

We put to the test six different models : three linear models (logistic regression, modified logistic regression and Poisson regression) and three algorithms of supervised statistical learning (classification tree, random forest, Poisson regression tree). The performances of these models are assessed by the AUC measured on a test group. The frequency models - Poisson regression and Poisson regression tree - are the most performing, due to their better capacity at taking into account the exposure variations. The selection of variables within the Poisson regression is provided by LASSO penalisation (Least Absolute Shrinkage and Selection Operator), which offers several advantages over the stepwise system : it is quick, autonomous, and possesses good statistical properties.

The synthetic over-sampling considerably improves the performance of the models. We reach an AUC of 70% for the linear model. The low volume of our data makes the selection of variables little stable but we can interpret the most important ones ; the previous attritionnal damages thus appear to explain severe risk quite well. Following optimisation, the tree reaches an AUC of 74%. It shows the tree is more sensitive to parameters of over-sampling than the linear model, due to its greater adaptability to the structure of the data. The tree makes it possible to identify a high-risk group inside which the frequency of severe damages is more than 20 times superior to the global frequency, and this through less than ten rules of decision. It is as well able to find simple and discriminatory criteria on great volumes : with only two criteria, one can isolate two big pricing groups, one being seven times more risky than the other.

Trees and linear models are shown to be not only complementary but in interaction. The linear model is the best method to cleverly deal with heteregeneous categorical variables, which cannot be used as such in trees. Trees offer in return a non-linear perspective on the data, which constitutes a good tool to improve model specification.

Note de synthèse

Prédire le risque extrême est stratégique : il est d'occurence rare et sa modélisation difficile, mais il constitue un levier tarifaire majeur. Les sinistres extrêmes génèrent fréquemment plus du tiers de la charge totale d'un produit d'assurance, et sont susceptibles de remettre en cause la solvabilité de l'assureur, qui doit faire appel à une réassurance coûteuse. Un assureur qui parvient à identifier des groupes de risque extrême sensiblement différent libère sa compétitivité tarifaire sur les groupes les moins risqués, et développe sa part de marché sans antisélection. La prime pure des graves étant élevée, la discrimination du risque, même faible, peut générer des variations de cotisation importantes. Nous étudions ce risque extrême sur le produit d'assurance professionnelle multirisque de PACIFICA - une filiale de Crédit Agricole Assurances - au cours de la période 2009-2015. Ce produit est commercialisé à travers le réseau de bancassurance du Crédit Agricole. Il est récent, et son portefeuille d'assurés est de taille modeste. Il présente du fait de sa nature multirisque une sinistralité hétérogène, dont nous modélisons la principale source de sinistres graves : les incendies.

Etude des valeurs extrêmes

La théorie des valeurs extrêmes établit en assurance la distinction entre sinistres graves et sinistres attritionnels. Le choix du seuil à partir duquel le montant d'un sinistre conduit à le qualifier de "grave" est issu d'un arbitrage biais-variance : la modélisation des sinistres graves repose sur des propriétés asymptotiques pour lesquelles le seuil doit être suffisamment élevé, mais les sinistres extrêmes étant par nature rares, il convient de veiller à conserver un nombre d'observations suffisant. Nous déterminons ce seuil par l'étude de la distribution des maxima et par deux méthodes d'estimation : la fonction moyenne des excès et l'estimateur de Hill. Ces deux estimateurs nous conduisent à un intervalle centré autour du seuil de 70 000 €, que nous retenons et qui définit 151 incendies graves.

Objectifs et données de l'étude

Notre étude porte sur la fréquence des graves. Le faible volume de données que ces derniers représentent ne permet pas d'envisager une modélisation satisfaisante du coût. Nous modélisons les incendies graves avec trois objectifs. Nous souhaitons obtenir un modèle de prime pure grave dont les critères puissent être inclus dans l'équation tarifaire actuelle ; nous cherchons donc un modèle linéaire performant. Nous voulons d'autre part associer le réseau de distribution spécialiste assurance à notre démarche : il dispose de marges de manoeuvre importantes en termes de réduction tarifaire et d'actions de prévention. Nous cherchons donc un modèle linéaire interprétable, qui puisse apporter une grille de lecture simple aux spécialistes de la distribution, afin de les former au diagnostic du risque grave et à la prévention associée. Nous souhaitons enfin parvenir à identifier des profils de haut risque, d'effectif suffisamment limité pour faire l'objet d'une revue au cas par cas par les caisses régionales. Ce troisième objectif fait appel à l'apprentissage statistique : les arbres de classification et de régression (CART) sont privilégiés pour la souplesse de leurs capacités d'apprentissage et pour

leur interprétabilité.

Nos données font l'objet de divers retraitements : mise en as-if des montants de sinistres, calcul de fréquences d'antécédents de sinistres par types de garantie, et traitement des valeurs manquantes. Nous cherchons d'autre part à mesurer le rôle joué par les variables externes dans la sinistralité grave. À partir de données de l'INSEE ou de ministères, nous récupérons ou construisons une quinzaine de variables externes, disponibles à l'échelle communale ou départementale. Ces variables sont géographiques (occupation des sols), sociales (taux de délits, densité de population), économiques (taux de chômage, taux de défaillance des entreprises) ou d'accidentologie (taux d'intervention des pompiers pour des incendies d'entreprises).

Modèles

Le choix des modèles et algorithmes étudiés est guidé par deux considérations. D'une part, nos données se prêtent spontanément à une modélisation binaire : un assuré ne peut connaître qu'un seul incendie grave au cours d'une année. Ceci nous conduit à utiliser la régression logistique comme premier modèle linéaire généralisé (GLM), ainsi qu'un arbre de classification (CART) et le *strong learner* associé, la random forest. D'autre part, nos données présentent des durées d'exposition très variables. Il est envisageable - sous certaines hypothèses et en changeant de fonction de lien - de modifier la régression logistique pour corriger l'effet de l'exposition, mais ce modèle s'avère inutilisable du fait de problèmes de convergence. Les modèles de fréquence permettent une prise en compte plus adaptée de l'exposition. Nous utilisons un second GLM, la régression de Poisson, ainsi que son algorithme dérivé en machine learning : l'arbre de régression de Poisson, dont le critère de division repose sur la statistique de test d'une loi de Poisson.

Comparer ces modèles exige de trouver une mesure de performance commune : nous retenons la mesure de l'aire sous la courbe ROC (*Receiver Operating Characteristic*), connue sous le terme d'AUC (*Area Under the ROC Curve*), qui n'est pas sensible au déséquilibre des données et mesure la capacité d'un modèle à discriminer le risque en ordonnant les observations par risque croissant. Tout modèle prédisant une probabilité ou une fréquence peut être évalué par l'AUC. Pour mesurer l'AUC, nos données sont échantillonnées en une base d'apprentissage comportant 80% des observations et une base de test contenant les 20% restants. La performance mesurée varie avec l'échantillonnage apprentissage/test, du fait de notre faible nombre de graves. Nous stabilisons la mesure en en prenant la moyenne sur différents échantillonnages apprentissage/test.

Prendre en compte l'exposition s'avère essentiel. Des deux GLMs, nous retenons la régression de Poisson, la régression logistique présentant des difficutés de convergence. L'arbre de régression de Poisson surperforme sensiblement l'arbre de classification, et n'est pas dépassé par la random forest.

Apports du machine learning

Modéliser la survenance des incendies graves se heurte d'emblée à une difficulté : ces sinistres représentent une infime part des observations. Nous disposons de 151 incendies graves pour plus de 600 000 observations, soit moins de 0.03% d'observations sinistrées. La modélisation d'une classe aussi fortement minoritaire est problématique : l'estimation d'un modèle par maximum de vraisemblance n'est pas possible, et la performance prédictive des arbres est limitée. Nous employons pour pallier ce problème une méthode d'apprentissage statistique devenue courante depuis une dizaine d'années dans des champs d'étude tels que la biostatistique : le suréchantillonnage synthétique, ou SMOTE (*Synthetic Minority Over-sampling TEchnique*). Le suréchantillonnage synthétique consiste à générer de nouveaux individus minoritaires, situés aléatoirement sur des segments entre individus voisins de la classe minoritaire, dans l'espace des variables explicatives. Il est proche du suréchantillonnage par réplication dans ses effets sur les modèles linéaires, mais beaucoup plus puissant pour les arbres. La réplication des observations ne permet pas d'améliorer la performance des arbres, qui sur-apprennent en n'identifiant que de petites régions autour des observations minoritaires (Figure 1). La génération d'individus dans l'espace situé entre ces observations le conduit à apprendre une région plus large, et développe sa capacité prédictive (Figure 2).

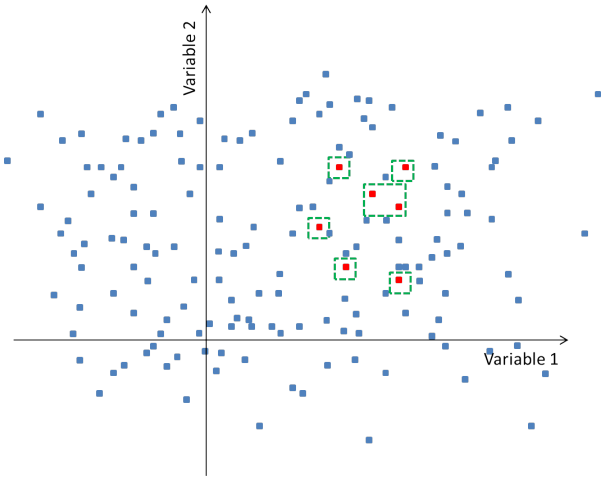


FIGURE 1 – Régions d'apprentissage du CART sur données brutes (ou suréchantillonnées par réplication)

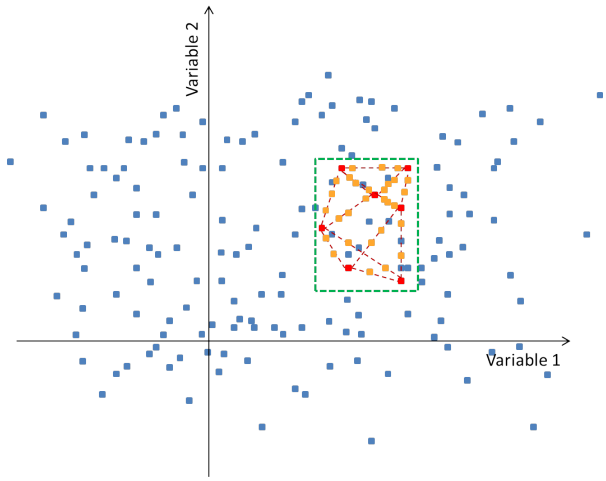


FIGURE 2 – Régions d'apprentissage du CART sur données suréchantillonnées par SMOTE.

Nous étudions la sensibilité des modèles de fréquence (régression de Poisson et arbre de régression de Poisson) au suréchantillonnage synthétique. Nous couplons le suréchantillonnage des graves à un sous-échantillonnage aléatoire des non-graves. Le ré-échantillonnage présente ainsi trois paramètres : le nombre de plus proches voisins utilisé dans l'algorithme de suréchantillonnage synthétique, l'effectif final de la base ré-échantillonnée et son taux de graves. Le suréchantillonnage synthétique peut perturber - par son mode de génération linéaire - des structures de données non-linéaires. L'effet de ces perturbations est fort sur les CART, qui sont sensibles aux structures non-linéaires, et faible sur les modèles linéaires. La performance du modèle linéaire repose sur un taux de graves fortement accru : nous multiplions la fréquence des graves jusqu'à plus de 1500 fois. Sous cette

condition, elle est peu sensible aux paramètres. L'arbre fait, lui, l'objet d'une réelle optimisation.

L'apport du machine learning à notre étude concerne aussi la sélection de variables au sein des modèles linéaires. La procédure stepwise, couramment employée en assurance, est particulièrement coûteuse en temps de calcul et peut présenter des biais statistiques. Notre étude requiert une méthode de complexité informatique moindre, pour concilier l'optimisation du GLM et la fiabilisation de sa performance, lesquelles, conjuguées, nécessitent d'ajuster un grand nombre de modèles. Nous avons recours au LASSO (*Least Absolute Shrinkage and Selection Operator*), qui consiste à introduire une pénalité au sein de l'estimation du modèle, afin de sélectionner les variables et de contrôler l'amplitude des coefficients estimés. Le LASSO présente plusieurs avantages : il est rapide, il possède de bonnes propriétés statistiques et il est autonome, son paramètrage étant réalisé par validation croisée.

La figure 3 résume notre démarche pour développer la capacité prédictive des modèles. Arbres et modèles linéaires n'apparaissent pas seulement complémentaires, mais en interaction. Le modèle linéaire est la meilleure méthode pour retraiter intelligemment les variables catégorielles présentant de nombreuses modalités, qui ne peuvent être utilisées telles quelles par les arbres. Il s'agit dans nos données des variables précisant l'activité principale et le département. Les regroupements permettent d'éviter le surapprentissage des arbres, et de simplifier l'interprétation des GLMs en créant des zoniers. En retour, les arbres offrent une perspective non-linéaire sur les données, qui constitue un bon outil pour améliorer la spécification du modèle. Les non-linéarités et les interactions de variables employées dans la régression de Poisson améliorée sont issues de sélections de variables par le CART et la random forest.

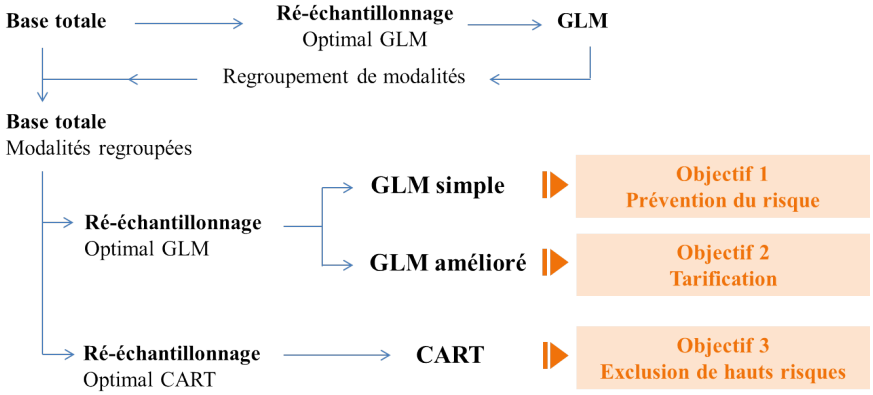


FIGURE 3 – Développement de la capacité d'apprentissage des modèles

Résultats

Notre étude montre un apport majeur des méthodes d'apprentissage statistique à la modélisation des sinistres graves. Confrontés à des données de faible volume, très hétérogènes et extrêmement déséquilibrées, nous parvenons à développer la capacité d'apprentissage de la régression de Poisson jusqu'à un niveau de performance satisfaisant. Nous obtenons un AUC de 67.7% pour la régression

simple, et de 70.0% pour la régression améliorée. L'objectif d'un modèle de prime pure pour les incendies graves est ainsi atteint (en associant un coût moyen au modèle de fréquence). Son interprétation demeure pour autant limitée à une dizaine de variables : le faible volume de nos données réduit la stabilité des sélections de variables, et ne permet pas d'en intégrer davantage avec fiabilité.

L'interprétation de la régression de Poisson attire l'attention sur l'intérêt d'une révision bayésienne du risque des assurés année après année : les antécédents d'incendie et d'autres sinistres sont les deux variables les plus explicatives du modèle linéaire. La surveillance et la ré-évaluation du risque à partir de la sinistralité attritionnelle constituent par conséquent un axe de progrès potentiel dans la prévention des incendies graves. Le modèle indique aussi un effet de la taille : la valeur de marge brute, la valeur vénale du fonds et les capitaux garantis augmentent la fréquence de graves, effets qui peuvent résulter d'une augmentation de la probabilité de survenance d'un incendie, ou d'un effet de seuil. Nos résultats montrent enfin l'intérêt des variables externes, avec un fort effet du taux de chômage. Les variables externes et d'antécédents sont d'autant plus utiles qu'elles sont obtenues en interne et ne nécessitent pas d'être intégrées au dispositif commercial ni aux questionnaires.

L'arbre de régression surperforme le modèle linéaire et atteint un AUC de 74 % après optimisation. Il identifie alors des groupes de haut risque où la fréquence de graves est plus de 20 fois supérieure, et ce en moins d'une dizaine de règles de décision. Nous retrouvons dans ces règles l'importance des antécédents de sinistres et des variables externes. L'arbre est également en mesure de dégager des critères simples et discriminants sur de grands volumes : en seulement deux critères - le premier sur les antécédents d'incendie, le second sur la surface de l'entreprise - il nous permet d'isoler deux grands groupes tarifaires, l'un 7 fois plus risqué que l'autre. Malgré sa performance, tout modèle CART demeure soumis à une certaine instabilité, et ses règles de décision sont susceptibles d'être modifiées avec l'arrivée de nouvelles données. Les règles de décisions initiales demeurent alors pertinentes, mais cessent de constituer l'optimum de l'algorithme.

Executive summary

Predicting extreme risk is a strategical issue : it rarely occurs and its modelling is difficult, but it constitutes a major pricing lever. Extreme damages frequently generate more than the third of the total expenses of an insurance product and are liable to cause doubt on the solvency of the insurance company, which must ask for a costly reinsurance. An insurance company which manages to identify groups of different severe risk levels relieves its tarif competitivity on less risky groups, and develops its market share without adverse selection. The pure premium of severe risk being high, the risk discrimination, even though weak, generates important variations of fees.

We study this extreme risk on the product of comprehensive professional insurance of PACIFICA, a subsidiary of the Credit Agricole, during the period 2009-2015. This product is marketed through the network of bank-insurance of the Credit Agricole. It is recent, and its portfolio is of moderate size. Due to its multirisk character, it has an heterogeneous sinistrality, of which we model the main source of extreme damages : fires.

Extreme value theory

In the insurance sector, the extreme value theory sets up the distinction between severe and attritionnal damages. The choice of the threshold which characterizes a damage as severe comes from a bias-variance arbitrage. The linear model of extreme damage rests on asymptotical properties for which the threshold must be high enough. But severe damages being generaly rare, it is necessary to keep a sufficient number of observations. We determine this threshold through the study of the distribution of maxima and through two estimation methods : the average function of excesses and the Hill estimator. These two estimators lead us to a threshold of 70 k€, which defines 151 severe fires in our data.

Aims and data of the study

Our study deals with the frequency of severe damages. The low volume of data which they represent does not allow to plan an adequate modelling of the cost. We model severe fires with three goals. We wish to obtain a model of severe pure premium which might be integrated into the present pricing equation. We are therefore looking for an efficient linear model. Moreover, we want to associate professionnal insurers in charge of sales to our approach ; they have important room for manoeuvre in terms of price reduction and prevention actions. Thus we are looking for an interpretable linear model which can offer a readable scheme to professionnal insurers, so as to train them to diagnose and prevent severe risk. At last, we want to identify high risk groups, of a limited enough size to allow an individual review of each insured party within them. This third objective requires statistical learning : classification and regression trees are privileged due to the flexibility of their training capacities and to their interpretability.

Our data are subject to various treatments : correcting damages costs as-if in 2015 (through an adapted inflation index), calculating the frequency of previous damages according to the different

types of guarantees and treating missing values. Moreover, we try to measure the part played by external variables within severe losses. With the help of data from the INSEE agency and ministries, we get back or build up about fifteen external variables, available on the scale of local government. These variables are geographical (ground occupation), social (offence rates, density of population), economic (unemployment rate, rate of failure of firms) or of accidentology (rate of intervention of firemen for firms fires).

Models

The choice of models and algorithms studied is guided by two considerations. On the one hand, our data are spontaneously liable to a binary modelling : an insured party can only have one severe fire within a year. This leads us to use logistic regression as first generalized linear model (GLM), as well as a classification tree (CART) and the strong learner associated to it, the random forest. On the other hand, our data present very changing exposure duration. It is possible - under certain hypotheses and through changing the link function - to modify the logistic regression in order to integrate the exposure effect. But this model happens to be unusable owing to problems of convergence. Models of frequency allow better ways of taking exposure into account. We use a second GLM, Poisson regression, and its derived algorithm in machine learning : the Poisson regression tree, which division criterion is set on the statistic test of a Poisson law.

Comparing these models requires finding a common measure of performance. We choose the measure of the Area Under the ROC (Receiver Operating Characteristic) Curve, denoted AUC, which is not sensitive to the unbalance of data and measures the capacity of a model to disciminate through ordering observations by increasing risk. Any model predicting a probability or a frequency can be assessed by the AUC. To measure the AUC, our data are sampled into a training set including 80% of the observations and a testing set containing the 20% left. The performance measured varies with the train/test sampling. Taking into account the exposure appears to be essential. Of these two GLMs, we keep the Poisson regression, the logistic regression presenting difficulties of convergence. The Poisson regression tree clearly overperforms the classification tree and is not beaten by the random forest.

Machine learning contributions

Modelling the occurence of severe fires immediatly sets up a difficulty : indeed these damages represent a minute part of the observations. We have 151 severe fires for more than 600 000 observations, that is, less than 0.03% of observations. The modelling of a class which is so minority is a problem : estimating a model through likekihood maximisation is not possible and the predictive performance of trees is limited. To compensate for this problem we use a method of statistical learning which has become common for about ten years in fields such as biostatistics : the synthetic over-sampling or SMOTE (Synthetic Minority Over-sampling Technique). This method consists in generating new minority individuals, located at random points on segments among individuals of the minority class, in the space of explanatory variables. It is close to over-sampling by replication

in its effects on linear models, but much more powerful for trees. The replication of observations does not permit to improve the performance of trees, which overfit by identifying only small areas around minority observations (Figure 4). The generation of individuals between these observations (Figure 5) leads the tree to learn a wider area and develop its predictive faculties.

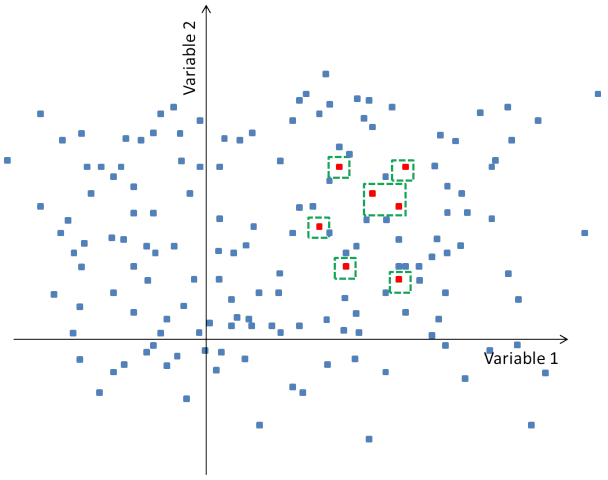


FIGURE 4 – CART learning areas on raw data or data over-sampled through replication.

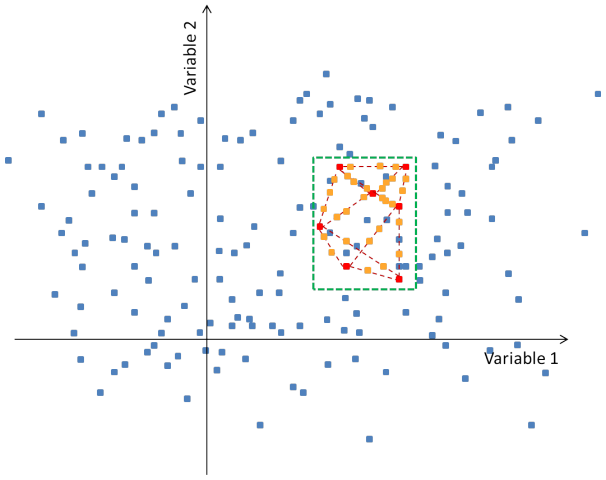


FIGURE 5 – CART learning areas on data over-sampled with SMOTE

We study the sensitivity of the frequency models (Poisson regression and Poisson regression tree) to synthetic over-sampling. We couple the over-sampling of severe damages to a random under-sampling of non-severe damages. Thus the re-sampling presents three parameters : the number of nearest neighbours used in the algorithm of synthetic over-sampling, the final number of observations in the re-sampled basis and its rate of severe damages. The synthetic over-sampling can disturb, through its mode of linear generation, non-linear structures in the data. The result of these disturbances is strong on the trees, which are sensitive to non-linear structures, and weak on linear models. The performance of the linear model is based on a seriously increased rate of severe damages : we multiply the frequency of severe damages as far as more than 1500 times. Under this condition, it is little sensitive to parameters. But the tree is subject to a real optimisation.

The contribution of machine learning to our study also concerns the selection of variables within linear models. The stepwise process, frequently used in insurance, is particularly costly in calculation time and can present statistical biases. Our study requires a method with a lesser computer complexity, to reconcile the optimisation of the GLM and the averaging of its performance for stability, wich put together require the adjustment of a high number of models. We use LASSO (Least Absolute Shrinkage and Selection Operator), which consists in introducing a penalty within the estimation of the model so as to select the variables and control the amplitude of the estimated coefficients. The LASSO offers several advantages : it is fast, possesses good statistical properties and is autonomous, its parameter being selected through cross validation.

Figure 6 sums up our process to develop the predictive faculty of models. Trees and linear models are shown to be not only complementary but in interaction. The linear model is the best method

for an accurate treatment of categorical variables showing numerous categories, which cannot be handled as such by trees. These in our data are the variables precising the main activity and the department. Regrouping allows to avoid the overfitting of trees and to simplify the interpretation of the GLMs. In return, the trees offer a non-linear perspective on the data , which makes up a good tool to improve the specification of the model. The non-linearities and the interactions of the variables used in the improved Poisson regression are originated from the selections of variables by the tree and the random forest.

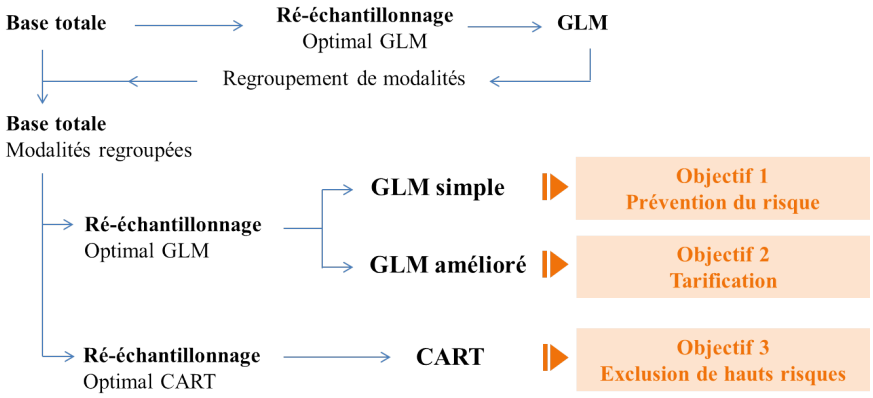


FIGURE 6 – Increasing the models performances

Results

Our study shows a major contribution of statistical learning methods to the modelling of severe damages. Having to deal with data of low volume, very heterogeneous and extremely unbalanced, we manage to develop the learning faculty of the Poisson regression up to satisfactory levels of performance. We obtain a AUC of 67.7% for simple regression and of 70.0% for improved regression. The goal of a pure premium for severe fires is thus reached (by associating an average cost to our frequency model). Yet its interpretation remains limited to ten variables : the low volume of our data reduces the stability of the variables selection and does not allow to integrate more with full reliability.

The interpretation of the Poisson regression draws attention on the interest of a bayesian revision of the risk of the suscriber year after year. Previous cases of attritional fires and various other damages are the two most explanatory variables of the linear model. The supervision and re-evaluation of risk based on attritionnal damage consequently make up an axis of potential progress in the prevention of severe fires. Our results also show the interest of external variables, with a strong effect of the unemployment rate. The external variables and the variables of former attritionnal damages are all the more useful as they are obtained internally and do not require to be integrated into the commercial device and question forms.

The regression tree over-performs the linear model and reaches an AUC of 74% after optimisation. Thus it identifies high-risk groups for which the frequency of severe damages is more than

twenty times higher, and this through less than ten rules of decision. In these rules we notice again the importance of previous damages and of external variables. The tree is also able to set out simple and discriminatory criteria on great volumes : with only two criteria - the first one on previous damage, the second on the building's surface - it allows us to isolate two important pricing groups, one being seven times more risky than the other. In spite of its performance, every CART model remains subject to a certain unstability and its rules of decision are liable to be modified with the coming of new data. The initial rules of decision then remain pertinent, but cease making up the optimum of the algorithm.

Remerciements

Je tiens à remercier mes deux tuteurs de mémoire, Laura Candas et Yann Mercuzot, pour leur encadrement et leur confiance tout au long de ce travail de recherche.

Je souhaite aussi remercier Guillaume Rosolek, directeur du marché des professionnels, pour son temps et ses précieux conseils.

Je remercie Isaac Sabban Cohen, stagiaire - et aujourd'hui en thèse - à PACIFICA, pour les riches échanges qui ont nourri ce mémoire, illustrant la maxime "Seul, on va plus vite. Ensemble, on va plus loin.".

Je remercie mes parents pour leur soutien et leur aide tout au long de mes études.

Je remercie enfin Estelle Blumereau pour ces deux belles années passées ensemble à l'ENSAE.

Nous allongeons le cou pelé par l'ignorance.
Toujours quelques nuages au moment d'y voir clair...
Nous n'en restons pas moins dans notre vigilance,
Espérant en connaître un peu plus long demain.

Jules Supervielle, *Bonne garde*

Table des matières

Introduction

La prévision du risque extrême constitue un enjeu primordial pour les compagnies d'assurance non–vie, par son coût et par la spécificité de sa distribution statistique. La théorie des valeurs extrêmes définit un seuil, qui établit en assurance la distinction entre sinistres graves, de montants supérieurs à ce seuil, et sinistres attritionnels. Les sinistres graves sont par définition rares, et les historiques issus de portefeuilles jeunes ou en évolution apportent peu de données. La discrimination tarifaire qui résulte de la modélisation est souvent faible, voire inexistante : la tarification repose essentiellement sur la prime pure attritionnelle.

Pour autant les sinistres graves constituent une part considérable de la charge totale des sinistres, susceptible de grèver lourdement le résultat de l'assureur, voire de mettre en cause sa solvabilité. Les compagnies d'assurance font de fait appel aux réassureurs, auxquels ils cèdent une part des cotisations acquises.

Prédire les graves est donc stratégique : un assureur qui parvient seul à identifier des profils d'assurés de forte sinistralité grave possède un avantage concurrentiel majeur. Même faible, la discrimination appliquée à une charge de sinistre élevée constitue un levier tarifaire. L'assureur est en mesure de réduire l'exposition de son portefeuille aux graves et d'abaisser la cotisation des profils de ses clients peu risqués. Il développe ainsi sa part de marché tout en améliorant son ratio de sinistres/cotisations et le coût relatif de sa réassurance.

C'est dans cette perspective que modélisons la sinistralité grave du portefeuille d'assurance multirisque professionnelle de PACIFICA. Ce portefeuille est jeune, de taille modeste et présente par sa nature multirisque une diversité de sinistres graves. Nous commencons par déterminer le seuil de définition d'un grave à l'aide de la théorie des valeurs extrêmes. Les sinistres graves étant majoritairement des incendies, nous nous appliquons à prédire ce type de sinistre.

Nous modélisons les incendies graves avec trois objectifs. Nous voulons tout d'abord élaborer un modèle de prime pure grave qui puisse être intégré à l'équation tarifaire actuelle. Deuxièmement, nous voulons associer les assureurs professionnels responsables de la vente à notre démarche : ils disposent de marges de manoeuvre importantes en termes de réduction tarifaire et d'actions de prévention. Nous souhaitons leur apporter une grille de lecture des éléments générateurs d'incendies graves pour leur permettre un diagnostic plus fin lors de la souscription. Enfin, nous souhaitons parvenir, dans la mesure du possible, à identifier une sous-population de hauts risques d'effectif suffisamment limité pour qu'elle puisse être ciblée par des actions de prévention ou par un non-renouvellement du contrat.

Notre premier objectif fait appel à des modèles linéaires généralisés, outil tarifaire par excellence, et notre deuxième objectif également, pour leur stabilité et leur interprétabilité. Notre troisième objectif requiert une grande souplesse dans l'apprentissage des données ; nous explorerons par conséquent l'apport d'algorithmes de machine learning : les arbres de classification et de régression ainsi que le *strong learner* qui en découle, la *random forest* (ou forêt aléatoire). Loins d'être concurrents, modèles linéaires et algorithmes d'apprentissage statistique s'avèrent complémentaires, et nous utiliserons chacun pour améliorer les performances de l'autre.

Modéliser la survenance des incendies graves se heurte d'emblée à une difficulté : ces sinistres représentent une infime part des observations. Nous disposons de 151 incendies graves pour plus de 600 000 observations, soit moins de 0.03%. Les modèles linéaires généralisés et les arbres sont inopérants sur des données marquées par un tel déséquilibre. Nous aurons par conséquent recours à une méthode d'apprentissage statistique devenue courante depuis une dizaine d'années dans des champs d'étude tels que la biostatistique : le suréchantillonnage synthétique, ou SMOTE (Synthetic Minority Over-sampling TEchnique), qui rééquilibre les données.

Notre recherche de performance prédictive s'appuiera par ailleurs sur l'élargissement du spectre des variables explicatives. Nous intégrons à notre modèle un ensemble de variables externes : données économiques, sociales, pénales et territoriales issues de l'INSEE et de ministères, disponibles à l'échelle de la commune ou du département. Et nous construisons des variables décrivant précisément les antécédents de sinistre d'un assuré depuis son entrée en portefeuille, afin de mesurer l'opportunité d'une surveillance du portefeuille.

Chapitre 1

Contexte, objectifs et données de l'étude

1.1 L'assurance multirisque professionnelle de PACIFICA

PACIFICA est la filiale assurance dommages du groupe Crédit Agricole Assurance. L'entreprise a été créée en 1990 afin de diversifier les activités du groupe, qui a pris dans les années 80 le virage de la bancassurance. Le rôle de PACIFICA est de protéger les clients du groupe Crédit Agricole face aux aléas de la vie ou aux risques qui pèsent sur leur activité professionnelle. En 2015, PACIFICA a généré 3 milliards d'euros de chiffre d'affaires contre 2,8 milliards en 2014, pour 10,4 millions de contrats en portefeuille (+5%). Symbole de l'avancée continue des bancassureurs, PACIFICA affichait une progression de 8% de ses affaires nouvelles en 2014. Son réseau de distribution regroupe les 39 caisses régionales du groupe Crédit Agricole, ce qui représente en 2015 7000 agences avec 26 500 professionnels de l'assurance, 480 assureurs "Pro" dédiés aux marchés des professionnels et des agriculteurs, ainsi que 2 077 agences LCL (6 000 professionnels de l'assurance). PACIFICA bénéficie du fort ancrage du groupe sur le territoire français et reste loin d'avoir saturé le portefeuille bancaire. En 2015, 32% des clients particuliers des caisses régionales et 19% des clients de LCL étaient équipés en assurance dommages chez PACIFICA.

L'assurance multirisque professionnelle est commercialisée depuis 2006 par PACIFICA, et consolidée depuis 2009. Elle propose différents types de garanties :

— Les dommages aux biens : incendies, dégâts des eaux, événements climatiques, terrorisme, vol, bris de machine et dommages électriques principalement
— La responsabilité civile : professionnelle, d'exploitation, du fait des locaux ou de propriétaire d'immeuble.
— La protection financière : frais fixe, perte d'exploitation, perte d'exploitation suite à sinistre, suite à maladie ou accident ou suite à carence du fournisseur, et la perte de valeur vénale du fonds.
— La protection juridique : assistance juridique et financière en cas de litige survenu dans le cadre de l'activité professionnelle.

En 2015, les cotisations acquises au titre de l'assurance multirisque professionnelle sont supérieures à 35 M€.

1.2 Objectifs : tarification, pré-sélection du risque et mesures préventives

Nous avons identifié trois axes principaux pour réduire l'exposition du portefeuille d'assurance professionnelle aux sinistres graves, auxquels correspondent trois modélisations distinctes :

- Développer un modèle linéaire pour intégrer une prime pure grave à la tarification (en complément du modèle de prime pure attritionnelle), et instaurer ainsi un premier niveau de discrimination. A la différence de la prime pure attritionnelle, la prime pure grave s'appuiera sur un modèle de fréquence seulement, la modélisation du coût n'étant pas entreprise dans le cadre de ce mémoire.

- Disposer d'éléments de discrimination du risque à partir d'un modèle linéaire simple (sans spécification de non-linéarités ou d'interactions) pour une communication à destination des assureurs professionnels. Ces derniers sont en effet des acteurs-clés de la pré-sélection du risque. La souscription d'une assurance multi-risque professionnelle débute par un devis, réalisé en agence auprès de conseillers professionnels, ou par téléphone grâce à des plateformes dédiées. Les informations obtenues servent à calculer une note de risque, qui détermine deux cas de figure. Si la note est inférieure à un seuil, le devis est maintenu en l'état. Si la note dépasse le seuil, une visite est réalisée sur le site du risque par un assureur professionnel qui établit un diagnostic et majore ou minore la cotisation en fonction de la note obtenue à ce diagnostic. Bien qu'issue d'une formule de calcul, cette note dépend d'un certain nombre d'appréciations de la part de l'assureur professionnel, et est très dépendante de la perception que l'assureur professionnel a du risque.
  Outre cette note de diagnostic, les assureurs professionnels disposent d'un certain nombre de leviers. Ils peuvent en particulier conditionner la souscription à des mesures telles que :

  — Une visite annuelle des installations électriques, ou un diagnostic par thermographie infrarouge.
  — La mise en place de télésurveillance pour prévenir les actes de vol ou de vandalisme, qui constituent une cause connue de sinistres graves.
  — L'éloignement des poubelles du bâtiment afin de réduire les risques d'incendie.

  Informer les assureurs professionnels sur le risque grave afin de leur permettre d'affiner leur perception est donc essentiel. Ils peuvent alors auditer certains risques avec une vigilance accrue, et auditer des profils dont la note de risque n'impose pas de visite s'ils le jugent utile.

- Identifier des profils de risque représentant une population de faible effectif et de marge for-

tement négative du fait des sinistres graves. Ces profils relèvent d'une modélisation complexe - arbres de décision, modèle linéaire amélioré - et sont ciblés au sein du service Actuariat. L'information sur les dossiers concernés est ensuite transmise aux Caisses Régionales, qui prennent les décisions qu'elles jugent adéquates (nouvelle visite de risque, mesures préventives, non-renouvellement du contrat).

1.3 Données

Nous étudions l'ensemble des contrats et des sinistres du portefeuille multirisque professionnelle, de janvier 2009 à décembre 2015. Ces données sont étalées dans le temps, et certaines sont récentes. Il est donc nécessaire de procéder à certaines correction de nature statistique - prise en compte des boni/mali - et économiques - prise en compte de l'inflation et de la dérive des coûts.

1.3.1 Prise en compte des boni/mali

Un certain nombre de sinistres ne sont pas encore "clôturés" dans notre portefeuille : leur charge finale n'est connue que par une estimation. On parle de boni ou de mali selon que la charge finale du sinistre s'avère inférieure ou supérieure à son montant estimé. Si l'estimation du montant des sinistres graves présente un biais, i.e. une tendance globale au boni ou au mali, il est utile de le prendre en compte. Le produit d'assurance multi-risque professionnelle présente une tendance au boni sur les sinistres de coût initialement estimés à moins de 50 k€ (dits "sous-crête"), tandis que les sinistres de montant supérieur à 50 k€ (dits "sur-crête") présentent une tendance au mali de 20%. Ces biais sont-ils à même d'affecter notre estimation de la fréquence des sinistres graves ? Il est ici nécessaire d'anticiper sur les résultats de la détermination du seuil des sinistres graves : nous serons amenés dans ce qui suit à définir les sinistres graves comme les sinistres de montant supérieur à 70 k€. La tendance au boni sur la sous-crête n'affecte pas la fréquence des graves : les montants de ces sinistres étant inférieurs au seuil, leur diminution est sans effet. La tendance au mali de la sur-crête est également sans effet, hormis pour les sinistres de montant estimé compris entre 50 et 70 k€. Une tendance au mali sur cette tranche conduirait à sous-estimer la fréquence des graves. Mais l'analyse plus fine de l'évolution de la charge sur-crête montre que la tendance au mali est principalement portée par des sinistres de montant supérieurs à 70 k€ - des sinistres en responsabilité civile notamment. La tendance au mali sur la tranche 50-70 k€ est ainsi nettement inférieure à 20%. Il apparaît donc peu important de corriger les montants de ces sinistres, dont le faible volume rendrait d'ailleurs toute estimation de tendance hasardeuse. Nous choisissons par conséquent de nous en tenir au même stade de liquidation pour toutes les années, en prenant chaque année à vision juin N+1. Nous évitons aussi par là même le biais lié aux tardifs de 2016, qui interviendrait si nous incluions la sinistralité janvier-juin 2016 à vision juin 2016.

1.3.2 Mise en as-if

Du fait des variations économiques, le coût d'un sinistre en 2009 n'est pas le même que celui d'un sinistre identique qui surviendrait en 2017. Nous corrigeons les montants de sinistre en fonction de leur année de survenance afin de les rendre comparables : cette transformation s'appelle communément une mise en "as–if".

L'analyse des coûts des sinistres sous-crête fournit un indice de dérive des coûts de 3.5% sur le long terme, qui est analogue à un indice d'inflation. Pour les graves, les garanties touchées sont la garantie dommage, la garantie RC, et les garanties financières (perte d'exploitation et de valeur vénale du fonds). La garantie dommage a été touchée par une augmentation de la TVA des professionnels en 2013 (+1.5%) qui impacte les coûts de réparation. Pour autant nous n'observons pas de véritable saut dans l'évolution des coûts moyens sous–crête qui permette d'évaluer l'impact réel de ce changement. Il est possible que les professionnels n'aient répercuté la hausse de TVA sur leurs tarifs que progressivement.

Avant 2015, la garantie FA (Frais Annexes) prenait en compte les frais liés au désamiantage. Les composants amiantés étaient souvent enterrés tels quels avec les déblais. Mais la réglementation a changé : il faut désormais faire appel à une compagnie de désamiantage, ce qui peut conduire à une forte augmentation des coûts. Pour répondre à cette évolution, une garantie spécifique FAM (Frais de désamiantage) a été mise en place en 2015. Par souci d'homogénéité nous rattachons la garantie FAM à la garantie FA (dans les variables indicatrices de garantie). Lorsqu'un volume suffisant de données sera disponible, il conviendra de mesurer l'impact des nouveaux coûts de désamiantage, à même de générer une nouvelle typologie de graves, sachant que le coût d'un sinistre peut être multiplié par 10 du fait du désamiantage.

1.3.3 Analyse de la sinistralité par montant

Dans cette partie, nous anticipons par commodité la définition du sinistre grave à laquelle nous conduira l'analyse des valeurs extrêmes. Pour notre portefeuille, un sinistre est dit "grave" lorsqu'il génère une charge totale as-if supérieure à 70 000 €.

Nous répartissons les sinistres selon la tranche de montant total dans laquelle ils se situent. Les sinistres de montant inférieur à 10k€ représentent plus de 95% des 40 000 sinistres de notre base, pour environ 35% de la charge totale. Les figures 1.1 et 1.2 donnent le nombre de sinistres par tranche pour les sinistres de montant supérieur à 10k€ et le poids de chaque tranche dans la charge totale.

On constate que si les sinistres de montant supérieur à 70 k€ représentent moins de 1% des sinistres en fréquence, ils représentent plus de 40% de la charge totale.

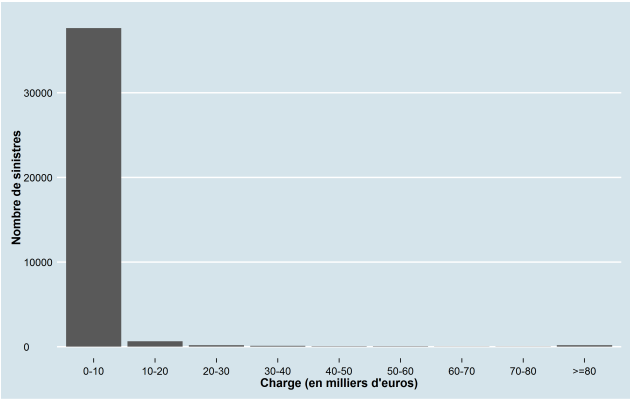


FIGURE 1.1 – Nombre de sinistres par tranches de coût

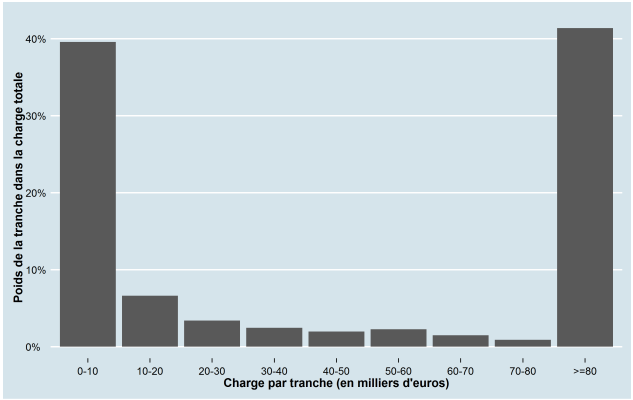


FIGURE 1.2 – Poids des tranches dans la charge totale

FIGURE 1.3 – Principales garanties sinistres de la multirisque professionnelle

ASS	Assistance	PE	Perte d'exploitation
ATT	Attentats	PEMA	Perte d'exploitation suite à maladie
AVAC	Avance corporelle	PJ	Protection juridique
BDG	Bris de glace	PMP	Perte de marchandises périssables
BDM	Bris de machine	PVVF	Perte de valeur vénale du fonds
DD	Dommages divers	RCAT	Responsabilité Civile (RC) après travaux
DDE	Dégâts des eaux	RCNC	RC non-auto corporelle
DR	Sauvegarde de vos droits	RCNI	RC dommages immatériels non consécutifs
ELEC	Dommages électriques	RCNM	RC non–auto immatérielle
FA	Frais annexes	RCPC	RC professionnelle corporelle
FAM	Frais amiante	RCPI	RC professionnelle dommages immatériels
FFIX	Frais fixes	RCPL	RC produits livrés
GEL	Gel	RCPM	RC professionnelle matérielle
INC	Incendie, foudre, explosion	VAN	Vandalisme
MT	Marchandises transportées	VOL	Vol

1.3.4 Analyse de la sinistralité par garantie

Il existe deux niveaux de décomposition des garanties : un niveau contractuel, qui relève du Code des Assurances et qui est celui utilisé pour le calcul des ratios Sinistres / Cotisations, et un niveau sinistre, plus fin, qui est utilisé pour l'analyse interne. La garantie contractuelle Incendie se décompose ainsi en plusieurs garanties sinistres : incendie pur, dommages divers, frais annnexes, etc.L'assurance multirisque professionnelle de PACIFICA comporte 18 garanties contractuelles, et 31 garanties sinistres. Les sinistres graves survenus du fait d'événements climatiques font l'objet de modélisations spécifiques et sont hors du champ de notre étude. Nous présentons dans le tableau 1.3 les garanties principales en termes de coût.

Afin de comprendre l'implication des différentes garanties dans la sinistralité grave, nous décomposons la charge selon les garanties au sein de chaque tranche (figure 1.4). Ce graphe permet d'observer que les différents processus générateurs de sinistres interviennent avec des fréquences sensiblement différentes selon que l'on considère la sinistralité de coût faible ou de coût élevé. Ainsi, si la sinistralité de coût 10-20 k€ provient pour près de 70% de vols et de dégâts des eaux, celle de coût 70-80 k€ est issue à plus de 85% d'incendies et de responsabilités civiles corporelles.

Le changement le plus important dans la composition des garanties est la forte croissance du poids des incendies avec la tranche de coût considérée.

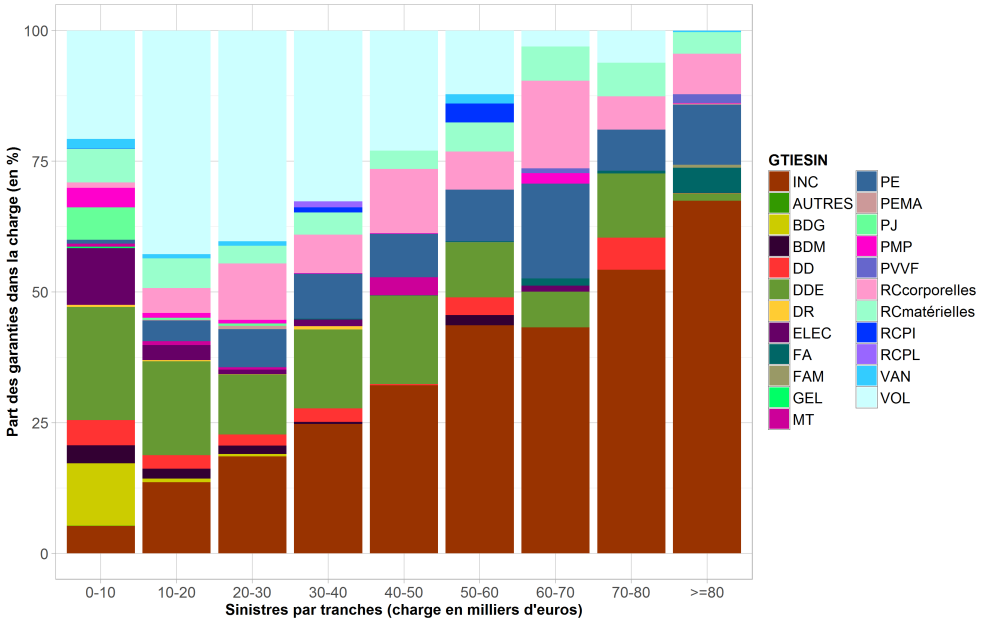


FIGURE 1.4 – Part des garanties dans la charge

Pour un examen plus approfondi des relations entre les différentes garanties, nous décomposons le montant de chaque sinistre de montant supérieur à 45 k€ selon les différentes garanties activées (figure 1.5). Les incendies (INC - 151 sinistres graves) activent de façon quasi-systématique une ou plusieurs autres garanties :

— La garantie perte d'exploitation (PE) génère fréquemment une augmentation de la charge, qui peut aller de 5 à 200%, du fait de l'interruption temporaire d'activité.

— La garantie Perte de Valeur Vénale du Fonds (PVVF) est plus rarement activée mais peut également accroître la charge de 200%. Cette garantie intervient lorsque la valeur patrimoniale d'un fonds de commerce dépend de la configuration de ses locaux et de son lieu d'implantation, ce qui est souvent le cas des artisans et commerçants. La perte peut être partielle si la réduction de surface des locaux ou leur relocalisation entraînent une désaffectation définitive d'une partie de la clientèle. Elle est totale si cela conduit à une perte définitive de l'intégralité de la clientèle.

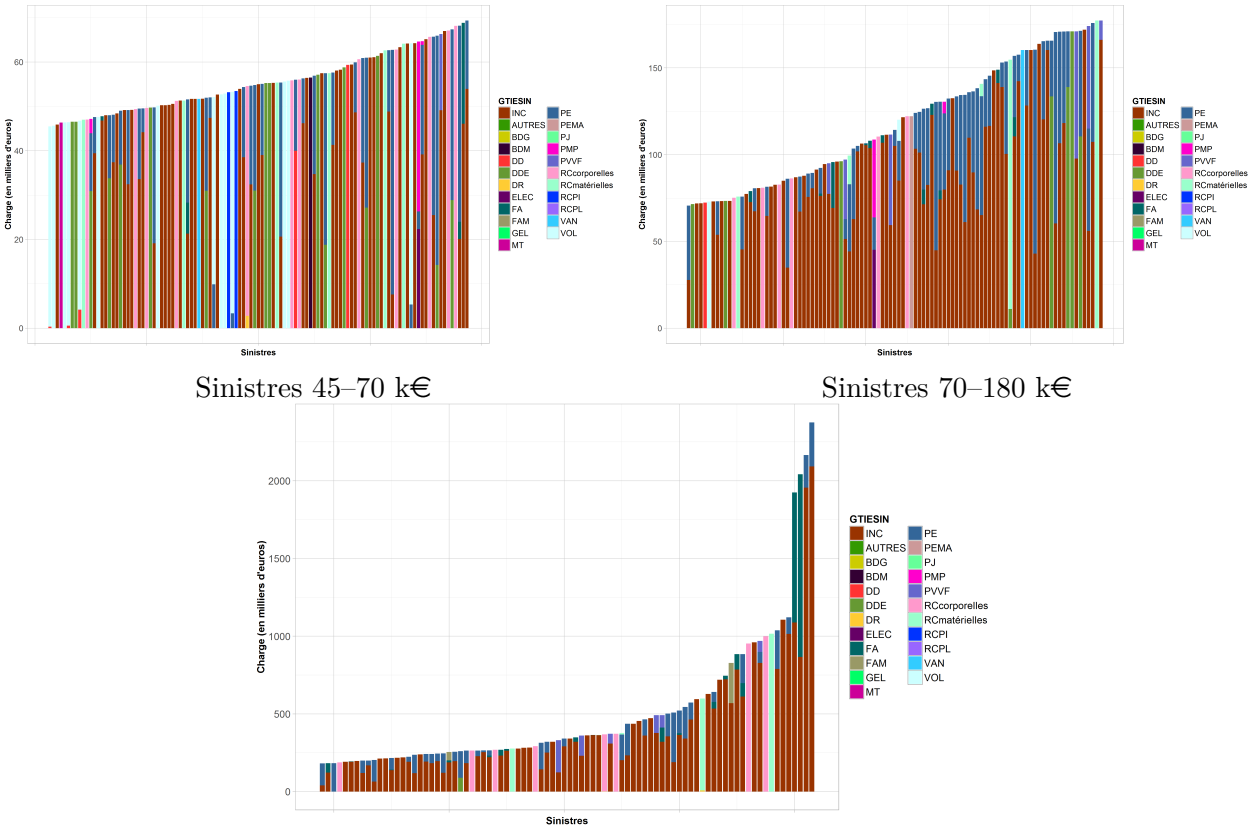
— Les garanties Frais Annexes (FA) et Frais de désamiantage (FAM – incluse dans FA avant 2015) génèrent lorsqu'elles sont activées une augmentation de charge inférieure à 30%, sauf dans le cas des deux des sinistres les plus importants du portefeuille (> 1.8 M€) où la garantie FA accroît la charge de 80 et 140%.

— Nous observons enfin un cas d'aggravation du fait d'un vol et deux cas du fait de responsabilités civiles matérielles, de montants faibles.

Les Dégâts Des Eaux (DDE) sont à l'origine de 9 sinistres graves, avec une charge maximale de 250 k€ et un coût moyen de 125 k€, souvent porté par des pertes d'exploitation élevées, et dans un cas par une RC matérielle. Les garanties responsabilités civiles corporelles (14 graves) et matérielles

(5 graves) n'activent aucune autre garantie. Nous observons enfin un grave pour vandalisme, un pour vol, un autre pour dommages divers et un dernier pour perte d'exploitation suite à maladie, tous inférieurs à 180 k€.

FIGURE 1.5 – Montants de sinistres décomposés par garanties



Sinistres 45–70 k€ Sinistres 70–180 k€

Sinistres > 180 k€

Cette analyse par garantie fait apparaître que la typologie des sinistres graves est diverse. Sur les 183 sinistres graves du portefeuille (185 moins deux sinistres retirés suite à des tests de cohérence des données), on compte quatre types principaux d'événements générateurs de sinistre : les incendies (151 sinistres), les responsabilités civiles corporelles (14), les dégâts des eaux (9) et les responsabilités civiles matérielles (5 sinistres). Une modélisation conjointe de l'ensemble de ces sinistres est peu opportune : un incendie d'origine électrique, la chute d'une personne du fait d'un encombrant laissé au sol dans des locaux de vente, et un dégât des eaux causé par un joint défectueux sont trois événements résultant de causalités bien distinctes. Celles-ci peuvent certes avoir en commun un manque d'entretien des locaux, du fait d'une situation économique difficile ou d'une gestion à risque de la part du souscripteur. Mais on peut raisonnablement penser que modéliser ces sinistres ensemble conduira à un modèle prédisant correctement les incendies (qui sont le type principal) et prédisant mal les autres sinistres, lesquels nuiront en outre à la qualité de la prédiction des incendies. Pour ces raisons, nous choisissons de ne conserver que les 151 sinistres graves déclenchés par des incendies, seul type de sinistre présentant un volume suffisant pour entreprendre une modélisation.

1.4 Retraitement des données de contrat

Nos données contractuelles consistent en un ensemble d'observations, correspondant chacune à l'image donnée d'un contrat au sein d'une année donnée. Notre base comporte plus de 600 000 observations, issues de 130 000 contrats environ (soit près de cinq observations par contrat en moyenne), d'une exposition moyenne de 200 jours. La distinction de deux observations d'un même contrat peut résulter de trois événements : la modification du contrat, le renouvellement de ce contrat - ces deux événements générants un changement d'image - ou simplement le changement d'année. Notre base comporte donc des lots - de très faibles effectifs - d'observations non-indépendantes : celles correspondant au même contrat. Certaines sont en outre identiques du point de vue des variables tarifaires, lorsque le changement d'image résulte du renouvellement du contrat ou du passage d'une année à une autre.

Nous pourrions souhaiter d'un point de vue statistique agréger ces observations identiques en termes de variables contractuelles. Nous ne le faisons pas, pour deux raisons. Premièrement, nous souhaitons conserver la possibilité d'évaluer la stabilité du modèle dans le temps, en appliquant par exemple un modèle calibré sur la période 2009-2014 aux données de 2015. Deuxièmement, le découpage d'un contrat par périodes courtes permet d'évaluer l'opportunité d'une surveillance du portefeuille, grâce à des variables de fréquence d'antécédents de sinistre régulièrement actualisées au cours de la durée de vie d'un contrat.

Nous créons en effet des variables indiquant la sinistralité antérieure d'une observation dans le portefeuille. Nous générons des variables de charge et fréquence par année d'assurance, ainsi qu'une décomposition de la fréquence par aggrégats de garanties -Vol, RC, Dommage électrique, Incendies, et Autres garanties - et ce à deux échelles de temps : sur le court terme (sur l'année précédente) ou sur le long terme (sur l'ensemble des observations depuis la souscription).

1.5 Statistiques descriptives des données

Pour des raisons de confidentialité, nous ne présentons pas dans ce mémoire les travaux de statistique descriptive réalisés. Évoquons simplement le point principal que cette étude a fait ressortir : la disparité des comportements de vente. Nous avons étudié la façon dont chaque vendeur utilisait les marges de manoeuvre tarifaire à sa disposition. Les résultats indiquent de forts biais individuels, et confirment l'importance d'une communication auprès des vendeurs pour le diagnostic du risque grave.

1.6 Intégration de données externes

Quelle est la part des facteurs externes dans la sinistralité grave ? La situation économique et sociale, la criminalité, la géographie du lieu sont-ils des éléments explicatifs du risque ? Ces questions nous ont conduit à une recherche approfondie de données externes.

Nous récupérons une quinzaine de bases de données nationales issues de l'INSEE ou disponibles sur le site data.gouv. Connaissant le code géographique INSEE du contrat (i.e. sa commune d'appartenance), nous pouvons récupérer ou construire un certain nombre de variables, à l'échelle de la commune, ou du département selon la granularité disponible. Les variables ainsi obtenues sont :

- Taux de défaillance des entreprises (par département et par année) de 2009 à 2015, construit à partir des données du nombre de défaillances par an croisées avec les estimations du nombre d'entreprises par département (données INSEE). Cette variable souffre d'une différence de périmètre sectoriel entre le nombre de défaillance calculé et le nombre d'entreprises estimées, mais constitue le meilleur proxy disponible pour capter l'aléa moral - l'incendie de leurs locaux par certains entrepreneurs en cas de difficultés économiques étant un cas avéré.
- Densité de population par commune (nombre d'habitants $/km^2$ en 2013).
- Médiane du revenu disponible par Unité de Consommation en 2012 (en euros).
- Taux de pauvreté en 2012 par commune (ou par département lorsqu'indisponible à l'échelle communale).
- Taux d'inactivité chez les 15-64 ans (hors élèves, étudiants, stagiaires, pré-retraités et retraités).
- Taux de chômage des personnes actives de 15 à 64 ans en 2013.
- Nombre de policiers par habitant en 2014 (commune, ou département lorsqu'indisponible à la commune).
- Taux de délits par habitant en 2014, que nous avons regroupés en 6 catégories :
  - Incendies volontaires de biens publics
  - Incendies volontaires de biens privés
  - Violences
  - Vols chez les professionnels
  - Autres vols
  - Autres délits
- Taux d'intervention des pompiers pour des incendies d'entreprises en 2014, calculé comme le nombre de feux de locaux industriels ou artisanaux rapporté au nombre d'entreprises, par département.
- Occupation des sols, à partir des données Corinne Land Cover – base de données européenne d'occupation des sols – à la commune. Nous aggrégeons les nombreuses catégories du Corinne Land Cover en cinq classes :
  - Tissu industriel
  - Tissu urbain
  - Tissu routier
  - Tissu artificialisé d'autre nature
  - Tissu non–artificialisé

1.7 Traitement des valeurs manquantes

Les données contractuelles présentent fréquemment des valeurs manquantes en assurance : certaines questions dépendent de la grantie souscrite, certaines sont conditionnelles à la réponse à une première question, d'autres enfin relèvent d'un diagnostic plus avancé du risque. Comment traiter ce données manquantes ?

1.7.1 Typologie des données manquantes

Commencons par rappeler les trois types de valeurs manquantes, et la façon pertinente de les traiter.

Missing Completely At Random - MCAR Les données de type MCAR désignent les valeurs manquantes d'origine complètement aléatoire : lorsque la probabilité d'absence de données pour une variable Y est indépendante de la variable Y elle-même ou des autres variables X de la base. Prenons l'exemple d'une enquête avec deux variables, l'âge et le revenu des sondés. Les données de revenu ne sont pas MCAR dans le cas où les sondés qui ne répondent pas à la question sont en moyenne plus jeunes que ceux y ayant répondu. Dans ces conditions, l'absence de réponse à l'enquête dépend de l'âge et la probabilité d'absence de réponse n'est pas la même pour tous les sondés. Supposons à l'inverse que chaque sondé décide de répondre à la question du revenu en lançant au préalable un dé non truqué : si la face 1 apparaît, alors le participant ne donne pas de réponse. Dans le cas inverse, il donne son revenu. Il s'agit ici d'une situation MCAR, la probabilité d'absence de réponse est la même pour tous les sondés. En présence de données MCAR, on peut retirer les observations manquantes sans générer de biais.

Missing At Random - MAR Les données sont MAR lorsque la probabilité d'absence dépend d'une ou plusieurs variables que l'on connait, sans pour autant dépendre de la variable à expliquer Y. Dans l'exemple précédent l'hypothèse MAR est satisfaite si la probabilité qu'une donnée de revenu soit manquante dépend uniquement de l'âge de l'individu. En présence de données MAR, on ne peut pas retirer les observations présentant des données manquantes. On doit alors avoir recours à une méthode d'imputation à partir des autres variables explicatives (imputation gaussienne multivariée par exemple).

Non Missing At Random - NMAR Dans le cas NMAR, la probabilité d'absence dépend de variables inobservées ou de la variable elle–même. Les données sont dans notre exemple NMAR si les personnes avec un revenu important refusent de le dévoiler. Les données NMAR induisent donc un biais non-maîtrisable sur la variable considérée.

1.7.2 Traitement des données contractuelles manquantes

Sur les 150 variables contractuelles disponibles, 81 présentent des valeurs manquantes - toutes des variables catégorielles. Nous pouvons d'emblée écarter l'hypothèse MCAR pour un grand nombre d'entre elles. L'absence ou la présence de la variable "durée de garantie perte d'exploitation souscrite" dépend ainsi directement de la variable "Garantie Perte d'Exploitation" qui indique si cette garantie a été souscrite. En outre nous pouvons penser que certaines variables sont NMAR : c'est le cas de la réponse quant au contrôle de l'installation électrique par un organisme agréé, dans la mesure où un assuré sera plus enclin à répondre s'il est agréé que s'il ne l'est pas.

On voit à travers ces deux exemples qu'une imputation des données manquantes n'aurait dans de nombreux cas aucun sens (indiquer un montant pour une garantie non-souscrite) ou générerait un biais (données NMAR). Nous préférons donc conserver les données manquantes telles quelles en considérant la non-réponse comme un type de réponse : nous recodons une variable X à deux réponses *oui* et *non* comme une variable à trois réponses : *oui*, *non* et *NR* (Non-Renseignée).

1.7.3 Traitement des variables d'antécédent

Le fait de générer des variables d'antécédents à partir de l'historique du portefeuille pose la question du traitement des observations correspondant à une entrée en portefeuille. Les observations issues d'une entrée en portefeuille ne présentent pas d'historique : contrairement à d'autres types d'assurances - assurance auto avec l'Agira - les assurances multirisques professionnelles ne font pas l'objet d'un historique partagé entre les assureurs (hors requête concernant un cas particulier, et sous réserve d'acceptation de l'assureur précédent). Les informations dont nous disposons sur la sinistralité antérieure des entrées en portefeuille sont partielles et déclaratives. Leur analyse indique qu'elles ne sont pas fiables, la fréquence et la charge déclarées étant bien inférieures à celles observées en portefeuille par la suite.

Pour autant on ne saurait exclure ces observations : elles représentent plus de 15% du portefeuille et plus de 10% des graves. En outre, si l'absence de ces variables n'est pas un processus en tant que tel - une entrée en portefeuille génère automatiquement une donnée manquante - l'entrée en portefeuille qui conduit à la générer constitue, elle, un processus. Les éléments motivants le choix de s'assurer ou de changer d'assureur ne sont a priori pas complètement aléatoires ; nous ne sommes pas dans le cas de données *Missing Completely At Random*.

Pouvons-nous corriger totalement le biais que représentent ces observations ? Il peut être *Non Missing At Random* dans certains cas : un professionnel peut avoir connu une sur-sinistralité le contraignant à changer d'assureur - celui-ci n'ayant pas renouvelé le contrat ou ayant augmenté son tarif en conséquence. L'occurence de la valeur manquante dépend alors de la variable elle-même. Nous disposons d'une variable indiquant la résiliation par la compagnie d'assurance précédente, mais qui n'apporte qu'une correction partielle, étant donnée qu'elle présente elle-même des valeurs manquantes et ne capte pas le cas où la compagnie précédente a augmenté son tarif. Nous ne pouvons donc pas corriger complètement les cas NMAR.

Nous pouvons en revanche corriger les cas de données *Missing At Random* : dans un certain nombre de situations, les variables conduisant à souscrire sont aussi des variables explicatives du modèle. Par exemple, l'évolution des variables tarifaires peut conduire à une première souscription d'assurance (augmentation du risque avec l'augmentation de la valeur de contenu par exemple) ou à un changement d'assureur du fait d'un positionnement tarifaire différent des compagnies sur les variables ayant évolué.

Nous choisissons donc de procéder à une imputation par arbre de régression. Pour chaque variable descriptive de la sinistralité antérieure :

— Nous ajustons un arbre de régression sur les observations renseignées (i.e. disposant d'un historique en portefeuille), prédisant la fréquence (ou la charge d'antécédent).

— Nous attribuons aux observations non-renseignées la valeur prédite par l'arbre à cette variable.

Notons qu'il demeure un biais général sur les fréquences de sinistralité antérieure depuis l'entrée en portefeuille, lié au fait que nous calculons des fréquences sur des anciennetés différentes. La fréquence calculée sur des assurés anciens est plus stable que celle des assurés récents (hors première année que nous imputons). La fréquence globale de sinistres attritionnels du produit multirisque étant inférieure à 15%, la plupart des assurés récents auront une fréquence calculée nulle, et ceux sinistrés une fréquence particulièrement élevée. Ce biais des fréquences de long terme, à même d'impacter les CART (mais sans effet sensible sur les modèles linéaires à priori), peut pour autant difficilement être corrigé.

Chapitre 2

Détermination du seuil des sinistres graves

La théorie des valeurs extrêmes est essentielle en assurance. Les montants de sinistre les plus élevés présentent en effet une distribution spécifique, qui indique (par contre-apposée) qu'ils relèvent de causalités distinctes des autres sinistres.

L'estimation des queues de distribution s'appuie sur cette théorie, qui fournit des procédures rationnelles et scientifiques. Elle permet d'isoler les valeurs extrêmes et de les modéliser séparément du reste de la distribution, afin d'améliorer leur prédiction tout comme celle des valeurs ordinaires (dont l'analyse n'est plus perturbée par des "outliers"). La distinction suppose de définir un seuil à partir duquel une valeur est considérée comme extrême. Le choix de ce seuil repose sur un arbitrage biais-variance : il doit être choisi suffisamment élevé pour que les distributions asymptotiques soient valables, mais le plus faible possible - sous cette contrainte de biais - afin de disposer d'un nombre d'observations correct, i.e. qui permette d'obtenir un estimateur de variance modérée pour la distribution des valeurs extrêmes. Notre étude de fréquence doit être envisagée comme faisant partie d'un modèle de prime pure, qui comporte également une modélisation de la distribution de ces coûts extrêmes. Aussi, si nous ne nous étendons pas sur la modélisation des extrêmes, c'est dans la perspective de cette dernière que nous choisissons le seuil, selon l'arbitrage biais-variance.

Le seuil issu de ce travail déterminera notre définition d'un sinistre "grave", par opposition à la sinistralité dite "attritionnelle" qui répond à des problématiques de modélisation différentes. Les quelques théorèmes, propriétés et définitions rappelés sont issus du cours de C.Y. Robert [13].

Si nous ne modéliserons par la suite que les sinistres graves provoqués par des incendies, nous menons l'étude des valeurs extrêmes conjointement sur l'ensemble des sinistres graves et sur les incendies seuls.

2.1 Estimation du paramètre de queue

2.1.1 Domaines d'attraction des valeurs extrêmes

Quelle est la fonction de répartition du maximum d'une suite de variables aléatoires $X_1, ..., X_n$ indépendantes et identiquement distribuées (IID), de fonction de distribution F ?
Soit $M_n = max(X_1, ..., X_n)$. Alors

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, ..., X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x)...\mathbb{P}(X_n \leq x) \\ &= [F(x)]^n\end{aligned}$$

Cette formule qui explicite la loi du maximum est peu utile en pratique : nous ne connaissons pas F, et son estimation à partir de l'ensemble des données s'appuierait principalement sur le coeur de la distribution. Toute inférence sur les queues de distribution par cette méthode serait par conséquent hasardeuse.

Notons x^F le point extrêmal de F, défini par :

$$x^F = sup\{x \,|F(x) < 1\}$$

Le support de F peut être borné ($x^F < \infty$: lois uniforme, beta...) ou infini ($x^F = \infty$: lois normale, exponentielle, Pareto, Gamma...). $M_n \xrightarrow[n\to\infty]{\mathbb{P}} x^F$, autrement dit la distribution asymptotique de M_n est dégénérée : une normalisation de la suite est donc nécessaire.

Le premier théorème fondamental de la théorie des valeurs extrêmes précise les lois asymptotiques que peut suivre le maximum normalisé d'une suite de variables aléatoires IID.

Théorème 1 (Fisher - Tippett)**.** *S'il existe des suites de réels a_n et b_n telles que quand $n \to \infty$,*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(xa_n + b_n)]^n \to G(x)$$

pour une distribution non–dégénérée G, alors G est du même type que l'une des trois distributions suivantes

$$\text{Fréchet } (\alpha > 0) : \quad \phi_\alpha(x) = \begin{cases} 0 & \text{si x} \leq 0, \\ exp(-x^{-\alpha}) & \text{si x} > 0 \end{cases}$$

$$\text{Weibull } (\alpha > 0) : \quad \psi_\alpha(x) = \begin{cases} 0 & \text{si x} \leq 0, \\ exp(-(-x)^{\alpha}) & \text{si x} > 0 \end{cases}$$

$$\text{Gumbel :} \qquad \Lambda_\alpha(x) = \quad exp(-e^{-x}) \quad x \in \mathbb{R}$$

Pour résoudre des problèmes statistiques, Von Mises (1954) et Jenkins (1955) ont proposé la loi d'extremum généralisée (notée GEV pour *Generalized Extreme Value*) qui unifie les distributions de Fréchet, Weibull et Gumbel :

Définition 1 (Loi d'extremum généralisée)**.** *La loi d'extremum généralisée $GEV(\mu, \sigma, \xi)$ est définie par la fonction de répartition*

$$G(x) = \begin{cases} exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right) & \text{si } \xi \neq 0 \\ exp\left(-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right) & \text{si } \xi = 0 \end{cases}$$

ξ est le paramètre de forme (également appelé paramètre de queue), μ le paramètre de position, σ le paramètre d'échelle.

Définition 2 (Domaine d'attraction)**.** *F appartient au domaine d'attraction de G ($F \in D(G)$) s'il existe deux suites (a_n) et (b_n) telles que la convergence précédente ait lieu.*

La correspondance entre la GEV et les lois limites est déterminée par le paramètre de queue :

- La distribution de Gumbel (queue fine) correspond à $\xi = 0 : GEV(0, 1, 0) =$ Gumbel
- La distribution de Fréchet (queue lourde) correspond à $\xi > 0 : GEV(1, \alpha^{-1}, \alpha^{-1}) =$ Fréchet(α)
- La distribution de Weibull (support borné) correspond à $\xi < 0 : GEV(-1, \alpha^{-1}, -\alpha^{-1}) =$ Weibull(α)

2.1.2 Estimation de la loi d'extremum généralisée

Bien que la modélisation GEV ne nous intéresse pas en tant que telle - puisque nous ne cherchons pas à modéliser les montants de sinistres - il est nécessaire d'estimer son paramètre de forme avant d'appliquer les méthodes de détermination du seuil basées sur la fonction moyenne des excès ou l'estimateur de Hill. Si nos données peuvent exceptionnellement présenter un sinistre sériel (lorsqu'un incendie touche deux assurés voisins), l'hypothèse d'observations indépendantes est globalement respectée sur nos données. Celle d'observations identiquement distribuées est davantage sujette à caution. Nous pouvons raisonnablement considérer qu'elle est vérifiée pour les montants de sinistres dûs aux incendies, et qu'elle l'est moins lorsqu'on considère l'ensemble des sinistres : incendie et dégât des eaux sont ainsi issus de processus à priori différents. Par conséquent nous mènerons en parallèle de l'estimation basée sur l'ensemble des sinistres une estimation basée uniquement sur les incendies, afin de vérifier que les résultats issus de ce sous-ensemble ne divergent pas trop des résultats globaux.

L'estimation de la GEV repose sur l'hypothèse que la loi limite est adaptée pour modéliser celle à distance finie ; autrement dit que nous avons l'approximation suivante :

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx GEV(\mu, \sigma, \xi)$$

Notons que la validité de l'hypothèse repose sur :

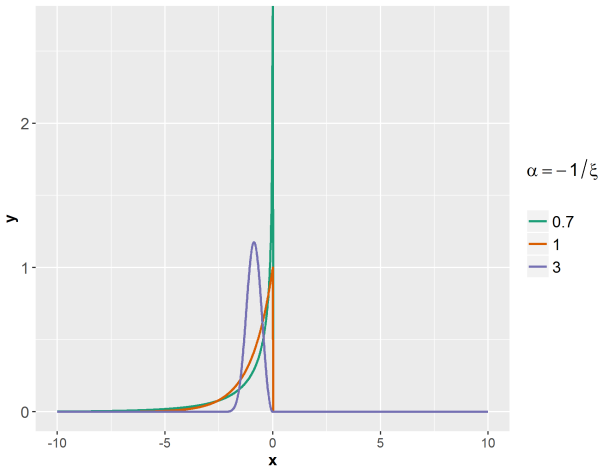


FIGURE 2.1 – Densité de la loi de Weibull

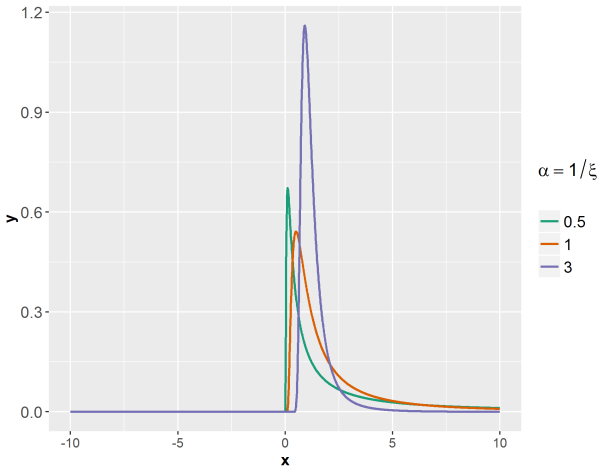


FIGURE 2.2 – Densité de la loi de Frechet

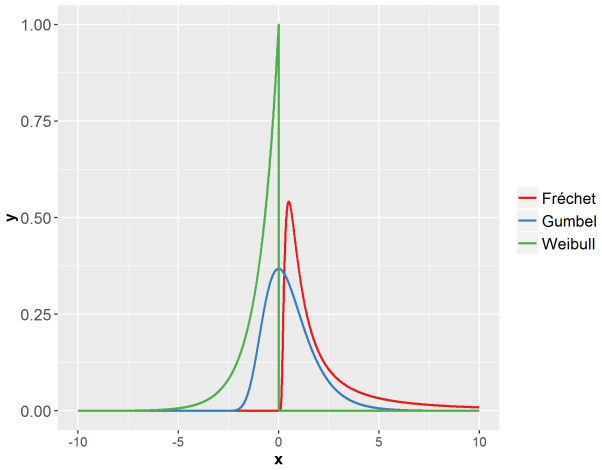


FIGURE 2.3 – Densités des différentes lois extrêmes

Densités des lois de Weibull et de Fréchet pour différentes valeurs du paramètre de queue, et comparaison des densités des trois lois : Weibull, Fréchet et Gumbel.

— une valeur de n qui soit suffisamment élevée.
— le fait que l'inverse de la fonction de hasard $h(x) = \frac{1-F(x)}{f(x)}$ soit de dérivée h' quasi-constante, ce qui dépend de la distribution F.

Les suites a_n et b_n dépendent de F, mais pour n fixé ce sont de simples constantes de normalisation. Le lot des cinq paramètres a_n, b_n, μ, σ et ξ ne présente qu'une complexité apparente, puisqu'il se résume aux trois paramètres de la GEV. Ce point n'étant précisé dans aucune des différentes sources bibliographiques que nous avons consultées, nous démontrons ici la correspondance des paramètres :

- Cas $\xi = 0$
  Supposons la relation asymptotique valable ; nous avons : $\mathbb{P}(M_n \leq a_n x + b_n) = e^{-e^{-x}}$.
  Nous posons $y = a_n x + b_n$, car ignorant a_n et b_n nous cherchons à estimer $\mathbb{P}(M_n \leq y)$.

Nous réécrivons donc la relation : $\mathbb{P}(M_n \leq y) = e^{-e^{\frac{y-b_n}{a_n}}}$, que nous allons estimer par la $GEV(\mu, \sigma, 0) : \mathbb{P}(M_n \leq y) = e^{-e^{\frac{y-\mu}{\sigma}}}$.
Nous avons ici une correspondance immédiate : $\mu = b_n$, $\sigma = a_n$

- Cas $\xi > 0$
  Nous estimons ici :

$$\mathbb{P}(M_n \leq y) = \begin{cases} 0 & \text{si } \frac{y-b_n}{a_n} \leq 0, \\ exp(-\left(\frac{y-b_n}{a_n}\right)^{-\alpha}) & \text{si } \frac{y-b_n}{a_n} > 0 \end{cases}$$

  par

$$\mathbb{P}(M_n \leq y) = exp\left(-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right)$$

  L'équivalence entre ces deux formules définit un système d'équation qui présente une unique solution :

$$\alpha = \frac{1}{\xi}, a_n = \frac{\sigma}{\xi} \text{ et } b_n = -\frac{\sigma}{\xi}$$

  Nous obtenons le même résultat dans le cas $\xi < 0$.

Ainsi, pour modéliser la loi limite du maximum il suffit de l'estimer directement par une loi $GEV(\mu, \sigma, \xi)$.

Pour estimer les paramètres de la $GEV(\mu, \sigma, \xi)$, nous utilisons une méthode dite des **"Maxima par bloc"**. Nous subdivisons l'échantillon de n observations en K sous-échantillons de taille n/K ; nous disposons alors d'un échantillon de K maxima.

Nous notons $\theta = \begin{pmatrix} \xi \\ \mu \\ \sigma \end{pmatrix}$ et $m_1, ..., m_K$ les maxima issus des sous-échantillons. La log-vraisemblance prend une expression différente selon la valeur du paramètre de queue :

— Pour $\xi \neq 0$:

$$l(\theta, m_1, ..., m_K) = -Kln(\sigma) - \left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{K} ln\left(1+\xi\left(\frac{m_i-\mu}{\sigma}\right)\right) - \sum_{i=1}^{K}\left(1+\xi\left(\frac{m_i-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}$$

— Pour $\xi = 0$:

$$l(\theta, m_1, ..., m_K) = -Kln(\sigma) - \sum_{i=1}^{K}\left(\frac{m_i-\mu}{\sigma}\right) - \sum_{i=1}^{K} e^{-\frac{m_i-\mu}{\sigma}}$$

Aucune solution analytique aux équations annulant les dérivées n'existe ; il faut donc avoir recours à une résolution numérique. Le calcul du maximum de vraisemblance ne pose pas de difficultés (Embrechts, Klüppelberg et Mikosch, [8]), et l'estimateur conserve les propriétés classiques de l'Estimateur du Maximum de Vraisemblance (EMV) tant que $\xi > -1/2$ (Smith, 1985 [14]). Nous sommes dans ce cas favorable, la distribution de sinistres que nous considérons n'étant à l'évidence pas une distribution de Weibull.

La GEV étant une loi limite pour le maximum, il est important que les blocs utilisés pour modéliser la loi soient de taille suffisamment grande. La largeur des blocs réduit cependant l'information disponible : à titre d'exemple, si l'échantillon k contient trois valeurs supérieures au maxima de l'échantillon $k-1$, on perd deux maxima potentiels. Ce phénomène de « superposition de maxima » s'accroît avec la taille des blocs. Nous pourrions discuter ici du fait que nos données proviennent vraisemblablement d'un mélange de lois (l'évolution des garanties concernées selon la charge du sinistre le suggère). Ceci pourrait justifier de tronquer les données à un seuil s. Supposons que l'on ait une loi D_1 à queue fine de quantile 99% égal à 10 000 et une loi D_2 à queue plus lourde de quantile 99% égal à 50 000. Retirer les observations inférieures à un seuil s=10 000 avant de procéder à l'échantillonnage par bloc maxima permet de se concentrer sur les observations de la loi D_2, laquelle génère les extrêmes que nous cherchons à modéliser. Un tel seuil présente une utilité si la baisse du nombre d'observations réduit les superpositions de maxima évoquées ci-avant (à nombre de blocs constant). Nous pouvons ainsi espérer limiter la perte d'information associée sans pour autant générer de biais - car les valeurs de D_2 retirées ont une très faible probabilité de pouvoir générer un maximum. Nous nous contentons d'évoquer cette réflexion sans la mener à son terme, qui nécessite d'expliciter par dénombrement l'espérance du nombre de superpositions de maxima en fonction du nombre de blocs et du nombre d'observations. Nous pourrions également envisager, afin de concilier la double exigence d'un nombre et d'une taille d'échantillons suffisamment élevés, d'accroître le nombre d'échantillons par bootstrap.

La GEV étant une loi discontinue en 0, il n'est pas possible de déterminer si nos données appartiennent au domaine d'attraction de la loi de Gumbel ou de celle de Fréchet à partir de l'estimation par maximum de vraisemblance de chacune des deux lois : les vraisemblances obtenues ne sont pas comparables du fait de la discontinuité entre ces deux lois. L'expérience actuarielle suggère qu'une distribution comme la nôtre appartient au domaine d'attraction de la loi de Fréchet, mais son paramètre de queue reste à estimer. Or il est nécessaire de connaître la distribution des maxima pour choisir un estimateur adéquat du seuil.

Anticipons sur les hypothèses nécessaires aux estimateurs :

— L'estimation par la fonction moyenne des excès nécessite de vérifier que $\xi < 1$. Ceci est possible à partir de l'estimation de la GEV pour $\xi \neq 0$.

— L'estimateur de Hill nécessite de vérifier que $\xi > 0$, ce qui n'est pas possible à partir de l'estimation de la GEV.

Nous estimons donc le cas $\xi \neq 0$, afin de vérifier que $\xi < 1$, ce qui permettra une première estimation du seuil par la fonction moyenne des excès.

Nous présentons ci-dessous les résultats de l'estimation par maximum de vraisemblance, réalisée sur l'ensemble des montants de sinistres - 39 000 observations, avec 400 observations par bloc - et sur les montants supérieurs à 5000 - 3000 observations, 30 observations par bloc (ce qui permet de conserver le même nombre de blocs dans les deux modèles). Les résultats obtenus étant susceptibles de varier avec l'échantillonnage des blocs, nous fournissons la valeur moyenne de $\hat{\xi}$ obtenue par ajustement de la GEV sur trois échantillonnages différents.

Paramètre de queue	Seuil d'écrêtement	$\hat{\xi}$ moyen
$\xi \neq 0$	0	0.631
	5000	0.624

TABLE 2.1 – Estimation de la GEV par maximum de vraisemblance

Nous obtenons une valeur moyenne de $\hat{\xi}$ similaire dans les deux cas - avec ou sans écrêtement. Ces valeurs sont significativement inférieures à 1 au vu des écart-types estimés lors de ces ajustements. Il est important de noter que la valeur de ξ estimée dépend fortement du nombre d'observations par bloc. Ainsi, si nous réduisons le nombre d'observations par bloc, $\hat{\xi}$ augmente, jusqu'à dépasser 1 - pour un nombre d'observations par bloc inférieur à 100 sur les données non-écrêtées. Mais le choix de blocs de taille aussi faible rend le fruit de ces estimations peu crédible : un certain nombre de maxima obtenus relèvent de la sinistralité attritionnelle et sont faibles, conduisant à une forte volatilité des maxima et à un $\hat{\xi}$ artificiellement haut. L'estimation de la loi de Pareto généralisée nous permettra de vérifier ce point.

Pour les sinistres ayant activé la garantie incendie (2000 sinistres), nous modélisons la GEV avec 20 observations par bloc. Les résultats mènent aux mêmes conclusions que ceux obtenus ci-dessus sur l'ensemble des sinistres.

2.2 Première estimation du seuil par la fonction moyenne des excès

2.2.1 Loi de Pareto généralisée

Définition 3 (Loi des excès)**.** *Pour un seuil u fixé, nous notons F_u la fonction de répartition de la loi des excès, définie par*

$$F_u(x) = \mathbb{P}[X - u \leq x | X > u]$$

et nous avons

$$\begin{aligned} F_u(x) &= 1 - \mathbb{P}[X - u > x | X > u] = 1 - \tfrac{\mathbb{P}[X > x+u]}{\mathbb{P}[X > u]} \\ &= \tfrac{F(u+x) - F(u)}{1 - F(u)} \end{aligned}$$

Le second théorème fondamental de la théorie des valeurs extrêmes - ci-dessous - indique que la convergence du maximum de variables aléatoires IID correctement normalisées est équivalente à la convergence de la distribution des excès vers une loi de Pareto généralisée (notée GPD pour *Generalized Pareto Distribution*).

Définition 4 (Loi de Pareto généralisée)**.** *La loi de Pareto généralisée $GPD(\beta, \xi)$ est définie par la fonction de répartition*

$$G^p_{\beta,\xi}(x) = \begin{cases} 1 - (1 + \xi \frac{x}{\beta})^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - e^{-\frac{x}{\beta}} & \text{si } \xi = 0 \end{cases}$$

où

$$\begin{array}{ll} x \geq 0 & \text{si } \xi \geq 0 \\ 0 \leq x \leq -\beta/\xi & \text{si } \xi < 0 \end{array}$$

Théorème 2 (Pickands, Balkema, de Haan)**.** *Soit $\xi \in \mathbb{R}$, les propositions suivantes sont équivalentes :*

— *Il existe deux suites a_n et b_n telles que $[F(a_n x + b_n)]^n \to G_\xi(x)$*

— *Il existe une fonction positive $\beta(.)$ telle que*

$$\lim_{u \to x^F} \sup_{0 < x < x^F - u} \left| F_u(x) - G^p_{\xi,\beta(u)}(x) \right| = 0$$

2.2.2 Fonction moyenne des excès

Lorsqu'une variable aléatoire suit une loi de Pareto généralisée, sa fonction moyenne des excès (notée MEF pour *Mean Excess Function*) présente une propriété que nous allons utiliser pour déterminer le seuil des valeurs extrêmes.

Définition 5. *(Fonction moyenne des excès) Soit X une variable aléatoire, on définit la fonction moyenne des excès, notée $e(u)$, par*

$$e(u) = \mathbb{E}(X - u | X > u)$$

Propriété 1. *Soit $X \sim GPD(\beta, \xi)$. $\mathbb{E}(X) < \infty$ si et seulement si $\xi < 1$.*
En outre si $\xi < 1$

$$\mathbb{E}(X - u | X > u) = \frac{\xi}{1 - \xi} u + \frac{\sigma}{1 - \xi}$$

Cette espérance conditionnelle est définie pour $u \in [0; \infty[$ si $\xi \in [0; 1[$, et pour $u \in [0; -\sigma/\xi[$ si $\xi < 0$
.

L'estimateur empirique de u est donné par

$$\hat{e}(u) = \frac{1}{\sum_{i=1}^n \mathbb{1}_{\{x_i > u\}}} \sum_{i=1}^n (x_i - u) \mathbb{1}_{\{x_i > u\}}$$

Le théorème précédent nous indique que lorsque $\xi < 0$, à partir du seuil u^* pour lequel l'équation $\lim_{u \to x^F} \sup_{0<x<x^F-u} \left| F_u(x) - G^p_{\xi,\beta(u)}(x) \right| = 0$ se traduit par une approximation $F_u(x) \approx G^p_{\xi,\beta(u)}(x)$ correcte, la fonction moyenne des excès est une fonction linéaire de u.

Cette propriété nous permet une détermination graphique du seuil u^*, à partir du graphe des estimateurs $\hat{e}(u)$ calculés pour tous les seuils (i.e. en pratique pour toutes les valeurs ordonnées distinctes de x, qui définissent chacune un nouveau seuil), figures 2.4 et 2.5. Nous cherchons la valeur u^* à partir de laquelle la courbe a une forme linéaire. Nous ne prenons pas en compte les estimateurs correspondant aux valeurs de seuil les plus élevées, car le faible nombre d'observations sur lesquels ils sont basés les rend très volatils. Embrechts, Klüppelberg et Mikosch (1999,[8]) rappellent que dans la mesure où la représentation graphique s'appuie sur un estimateur, il faut chercher une linéarité approximative, ce qui peut rendre l'interprétation difficile. Nous bénéficions ici d'un résultat assez heureux en comparaison d'autres cas observés en pratique. La fonction moyenne des excès semble raisonnablement linéaire à partir de $u^* = 70$ k€, avec une incertitude de $\pm$ 5 k€. Le graphique obtenu spécifiquement sur les sinistres incendies est très proche. Nous retenons ainsi une première estimation du seuil à 70 k€.

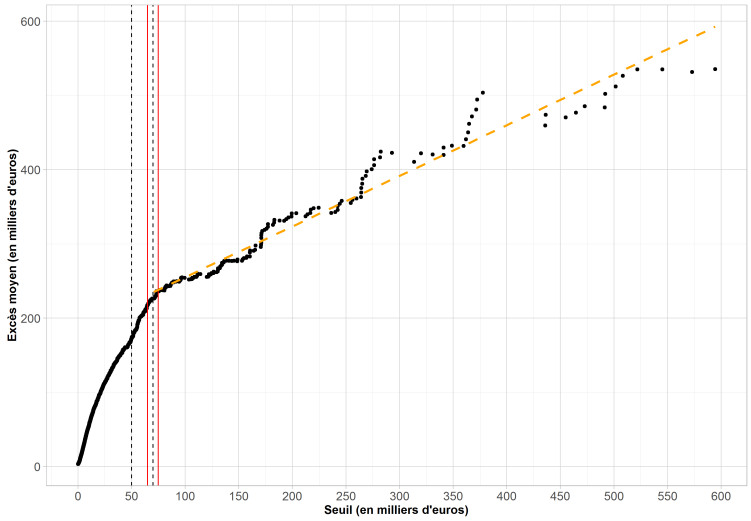


FIGURE 2.4 – Fonction moyenne des excès

Trait orange pointillé : résultat de la régression linéaire sur les valeurs de fonction moyenne des excès pour des seuils compris entre 75 000 et 600 000. Trait noir vertical : indique le seuil u = 70 000.*

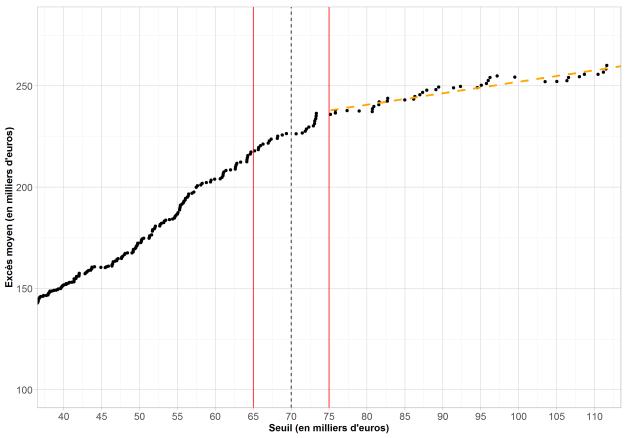


FIGURE 2.5 – Fonction moyenne des excès - Zoom autour du seuil 70 000

2.3 Ré-estimation du paramètre de queue

Ayant défini une première estimation du seuil, nous allons estimer par d'autres méthodes le paramètre de queue afin de vérifier que nous avons bien $\xi < 1$ (sans quoi l'emploi de la fonction moyenne des excès n'est pas cohérent) et de déterminer si $\xi > 0$, ce qui permettrait d'utiliser l'estimateur de Hill.

2.3.1 Graphique quantile-quantile

Nous nous appuyons en premier lieu sur une méthode graphique introduite par Embrechts, Klüppelberg et Mikosch (1999, [8]).

Propriété 2. *Soit une variable X appartenant au domaine d'attraction de la loi de Gumbel. Alors la loi des excès de X au-delà d'un seuil u converge vers une loi exponentielle lorsque $u \to \infty$.*

Un graphique quantile-quantile est une méthode graphique permettant de mesurer l'adéquation d'une variable observée à une loi théorique de fonction de répartition F continue. Lorsque X est à fonction de répartition F continue, $F(X) \sim U[0;1]$. En notant $X_{(1)} \geq ... \geq X_{(n)}$ la statistique d'ordre associée à une variable, on a donc

$$\Big(F(X_{(i)})\Big)_{i=1,...,n} = \Big(U_{(i)}\Big)_{i=1,...,n}$$

soit de façon équivalente :

$$\Big(X_{(i)}\Big)_{i=1,...,n} = \Big(F^{-1}(U_{(i)})\Big)_{i=1,...,n}$$

Le graphique quantile-quantile est donné par :

$$\Big\{X_{(i)}, F^{-1}(1-i/n) : i = 1, ..., n\Big\}$$

L'adéquation des observations à la loi de fonction de répartition F se traduit par la linéarité du nuage de points obtenu. Soit dans notre cas, avec la loi exponentielle standard :

$$\Big\{X_{(i)}, -ln(i/n) : i = 1, ..., n\Big\}$$

Méthode d'interprétation graphique – Embrechts, Klüppelberg et Mikosch [8]

Supposons u suffisamment grand ($u = u^*$ dans notre cas). Si la loi dont sont issues les observations appartient au domaine d'attraction maximum de la loi de Gumbel, alors les points du graphique sont approximativement alignés. Sinon, on observe une forme concave (domaine d'attraction de la loi de Fréchet) ou convexe (domaine d'attraction de la loi de Weibull).
Notre graphique quantile-quantile des excès au-delà de $u^* = 70000$ (figure 2.6) présente une légère concavité, qui traduit l'appartenance au domaine d'attraction de la loi de Fréchet ($\xi > 0$). Nous allons le vérifier en estimant la loi de Pareto généralisée.

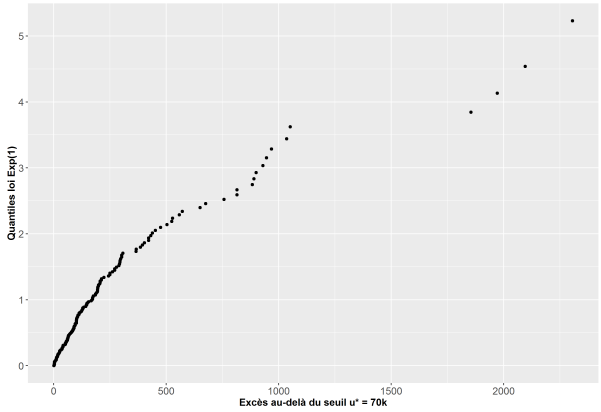


FIGURE 2.6 – Graphique quantile–quantile : loi exponentielle standard vs excès de sinistres au-delà de $u^* = 70000€$

2.3.2 Estimation de la loi de Pareto généralisée

L'estimation de la loi de Pareto généralisée au seuil $u^* = 70000$ fournit un $\hat{\xi}$ égal à 0.466, avec un écart-type de 0.089. Sur les données spécifiques aux sinistres incendies nous obtenons $\hat{\xi} = 0.449$, avec un écart-type de 0.096. Ces résultats confirment que la distribution étudiée appartient bien au domaine d'attraction de la loi de Fréchet - nous pouvons donc réaliser une seconde estimation du seuil avec l'estimateur de Hill - et que $\xi < 1$, ce qui confirme la validité de l'hypothèse nécessaire à notre première estimation par la fonction moyenne des excès.

2.4 Seconde estimation du seuil par l'estimateur de Hill

L'estimateur de Hill est le plus fréquemment utilisé en théorie des valeurs extrêmes lorsque $\xi > 0$. Il assure un bon équilibre biais-variance (Dress, de Haan et Resnick, 1998 [1]). En notant $X_{(1)} \geq ... \geq X_{(n)}$ la statistique d'ordre, n le nombre d'observations, et k un entier inférieur ou égal

à n, l'estimateur de Hill s'écrit :

$$\xi_{k,n}^{Hill} = \frac{1}{k}\sum_{i=1}^{k} ln\left(X_{(i)}\right) - ln\left(X_{(k)}\right)$$

Si $k \to \infty$ (soit $k/n \to 0$ lorsque $n \to \infty$), l'estimateur est faiblement convergent. Sous des hypothèses supplémentaires sur k et sur la fonction de répartition, il est de plus asymptotiquement gaussien, ce qui conduit à tracer un intervalle de confiance sur le graphe de l'estimateur de Hill (en pointillés rouges figure 2.7) :

$$IC_{95\%}(\xi) = \left[\hat{\xi}^{Hill} - 1.96 \cdot \frac{\hat{\xi}^{Hill}}{\sqrt{k}} \; ; \; \hat{\xi}^{Hill} + 1.96 \cdot \frac{\hat{\xi}^{Hill}}{\sqrt{k}}\right]$$

Le graphe de l'estimateur de Hill (figure 2.7) consiste à représenter la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, soit l'estimateur construit sur les observations supérieures ou égales au seuil $X_{(k)}$. En haut du graphe sont indiqués les seuils $X_{(k)}$. L'estimateur de Hill est volatil lorsque k est faible, puis se stabilise : le seuil des graves est déterminé comme celui à partir duquel l'estimateur se stabilise. Sur le graphe obtenu, l'estimateur de Hill croît puis se stabilise en effet : nous identifions le début de la zone de stabilité autour de 70 000. Les variations qui suivent sont d'amplitude sensiblement inférieure aux intervalles de confiance, aussi semble-t-il raisonnable de définir la zone de stabilité à partir de la fin de la tendance croissante. Un agrandissement de cette région (figure 2.8) conduit à un seuil entre 67 000 et 75 000.

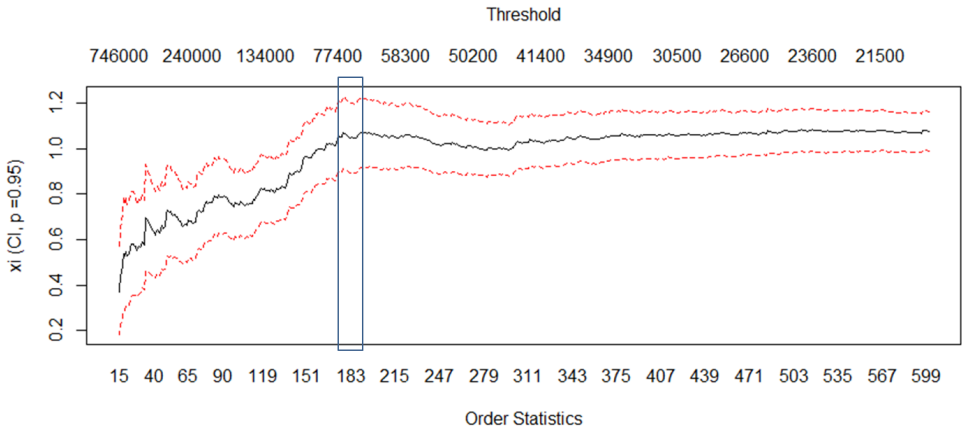


FIGURE 2.7 – Graphe de l'estimateur de Hill

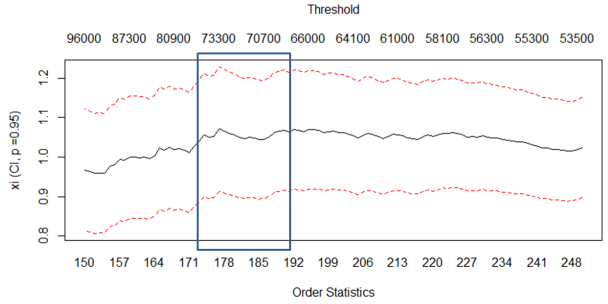


FIGURE 2.8 – Graphe de l'estimateur de Hill - Zoom autour du seuil 70 000

Le résultat de l'estimateur de Hill est en cohérence avec celui de la mean excess function : nous obtenons une estimation du seuil autour de 70 000 €. Nous conservons donc le résultat de la première estimation : nous définirons les sinistres graves comme des sinistres de montant as-if total supérieur à 70 000 €. Notons cependant que la valeur de ξ correspondant à ce seuil est proche de 1 sur le graphe de l'estimateur de Hill (en ordonnée sur le graphe), ce qui peut remettre en cause la première méthode d'estimation.

Chapitre 3

Modèles de prédiction du risque grave

Nos données nous placent spontanément dans un cadre binaire : dans notre portefeuille, aucun assuré n'a eu plus d'un sinistre grave au cours d'une année d'assurance. En outre, si la sinistralité multiple reste une éventualité - en responsabilité civile par exemple, deux accidents la même année sont possibles - elle est très improbable pour les sinistres incendie : étant donnés les délais d'expertise et de remise en état à la suite d'un incendie grave, l'activité reprend dans des délais de plusieurs mois. Nous nous orientons par conséquent vers un modèle de propension :

$$Y = \begin{cases} 1 & \text{si l'assuré a (au moins) un sinistre grave au cours de l'année d'assurance} \\ 0 & \text{sinon} \end{cases}$$

Pour autant, notre choix de modèle est aussi guidé par l'importance de l'exposition dans nos données : les individus étudiés présentent en effet des durées d'observation très variables, ce que montre la figure 3.1. La durée d'exposition est en moyenne de 200 jours, et varie de 1 jour à une année entière (la structure particulière de l'histogramme vient du fait que de nombreux contrats commencent au début d'un mois).

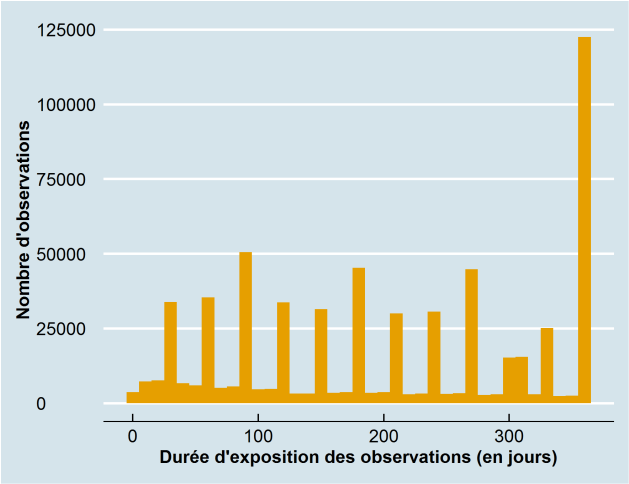


FIGURE 3.1 – Distribution de l'exposition des observations

Ne pas prendre en compte l'exposition pose deux types de problèmes :

— Il est assez fréquent - dans une situation d'exposition inférieure ou égale à un an - de considérer que tous les individus sont exposés une année entière. Cette hypothèse conduit à sous-estimer la prime annuelle. Pour le montrer, notons π la probabilité journalière de survenance

d'un grave, et E la durée d'exposition en jours. En supposant le processus de survenance des graves sans auto-corrélation, nous avons :

$$\mathbb{P}(Y = 0) = (1 - \pi)^E$$

Prenons le cas d'une personne ayant été assurée 100 jours ; nous prédisons

$$\mathbb{P}(Y = 0) = (1 - \pi)^{100} \quad \text{par} \quad \widehat{\mathbb{P}(Y = 0)} = (1 - \hat{\pi})^{365}$$

et obtenons $\hat{\pi} < \pi$. Nous pouvons certes revenir à un équilibre financier par une transformation linéaire faisant correspondre les primes pures globales estimée et observée, mais nous n'obtenons pas la correspondance au niveau individuel (certaines primes pures sont sous-estimées, d'autres surestimées).

— Les variations d'exposition entre observations peuvent réduire la performance prédictive du modèle : entre deux observations sinistrées, celle de plus faible exposition témoigne d'un risque plus grand. Or le modèle les traite de façon identique.

L'importance de l'exposition dans nos données nous conduit au-delà du cadre binaire, vers les modèles de fréquence, qui prennent mieux en compte l'exposition que les modèles de propension. Le cadre binaire restera cependant notre référence, pour évaluer les modèles avec l'AUC. C'est après avoir transformé les modèles de fréquence en classifieurs binaires que nous les évaluerons, afin de pouvoir les comparer. Le produit des modèles sera traité comme un vecteur de probabilités : à chaque observation prédite est associée une valeur comprise entre 0 et 1, correspondant à la probabilité de survenance d'un grave au cours d'une année ($Y_i = 1$).

Nous explorons une variété de modèles à même de répondre à nos objectifs : modèles linéaires généralisés et algorithmes d'apprentissage statistique - arbres de classification et de régression, et random forests.

3.1 Mesure de la performance prédictive des modèles

3.1.1 Evaluation de la performance d'un classifieur binaire

Comment mesurer la performance d'un modèle de classification binaire ? Afin de confronter le vecteur de probabilités prédites aux valeurs Y_i observées, il est nécessaire de définir un seuil s : pour une probabilité supérieure à ce seuil, on prédit $\hat{Y}_i = 1$, et pour une probabilité inférieure on prédit $\hat{Y}_i = 0$. La performance du classifieur pour ce seuil est ensuite mesurée à partir de sa matrice de confusion.

	Prédits Négatifs	Prédits Positifs
Négatifs	TN	FP
Positifs	FN	TP

FIGURE 3.2 – Matrice de confusion

TN (True Negatives) est le nombre d'observations négatives ($Y = 0$) correctement prédites, FP (False Positives) le nombre d'observations négatives incorrectement prédites. TP (True positives) est le nombre d'observations positives ($Y = 1$) correctement prédites, FN (False Negatives) le nombre d'observations positives incorrectement prédites.

Différentes mesures peuvent être construites à partir de la matrice de confusion. La plus globale est l'exactitude de prédiction (Accuracy) :

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Et le taux d'erreur est défini comme $1-$Accuracy. Les taux de prédiction correcte des observations positives (TPR) et des observations négatives (TNR) sont définis par :

$$\begin{aligned} \text{True Positive Rate} &= \frac{TP}{TP+FN} \\ \text{True Negative Rate} &= \frac{TN}{TN+FP} \end{aligned}$$

On définit également la précision - taux de prédictions positives correctes - et le rappel (Recall) - taux de prédiction correct des positifs, égal au TPR (permettant de construire la F-measure).

Dans une situation de données déséquilibrées, les effectifs issus des observations négatives, TN et FP, dominent largement ceux issus des observations positives, TP et FN. L'exactitude de prédiction (Accuracy) est alors inadaptée : le meilleur classifieur possible est en général celui qui consiste à prédire toutes les observations négatives, et un tel classifieur est sans intérêt. Les taux de vrais positifs et vrais négatifs sont en revanche insensibles à ce déséquilibre et peuvent être utilisés.

La courbe Receiver Operating Characteristic (ROC) est définie par l'ensemble des couples (FPR, TPR) obtenus en faisant varier le seuil s. On peut l'interpréter comme la frontière des meilleures décisions possibles lorsque le coût relatif des faux positifs et des faux négatifs varie. La droite $y = x$ représente la performance d'un classifieur aléatoire, tandis qu'un classifieur parfait serait situé dans le coin supérieur gauche du graphe (TPR = 100%, FPR = 0%).

Afin de comparer deux modèles, on superpose leurs courbes ROC. Si l'un des modèles présente une courbe ROC supérieure en tout point à la courbe de l'autre, alors il est à privilégier. Si les courbes se croisent, la décision n'est pas immédiate. Une mesure de performance permettant de comparer les deux modèles est alors l'aire sous la courbe ROC (AUC – Area Under the receiver operating characteristic Curve) : le modèle avec l'AUC la plus élevée est celui qui performe globalement le mieux. En pratique, le choix du modèle dépend du taux de faux positifs que l'on choisit : on retient le modèle qui fournit le meilleur taux de vrais positifs à TFP donné.

Il est essentiel de noter que la courbe ROC et l'AUC indiquent si la façon dont on a ordonné les observations (par probabilité croissante) est correcte, mais n'indiquent pas si ces probabilités sont justes. Pour un ordre donné, les probabilités prédites par le modèle peuvent être distribuées de façons très différentes. La seule contrainte sur cette distribution est que la moyenne des probabilités prédites sur la base d'apprentissage soit égale à la proportion de positifs observée. Les probabilités

prédites par un classifieur ne sont donc pas nécessairement justes, fût-il performant au sens de l'AUC, et l'on observe couramment en pratique des probabilités prédites très différentes entre deux modèles de même AUC [6].

3.1.2 Bases d'apprentissage et de test

L'évaluation d'un modèle implique de réserver une partie des données, qui n'est pas utilisée lors de la modélisation, et sert à tester le modèle. On s'affranchit ainsi des effets de surapprentissage qui peuvent survenir lorsque le modèle est évalué sur les données mêmes qui ont servi à son ajustement, donnant une performance artificiellement bonne.

La pénalisation LASSO et l'élagage des arbres de régression - que nous verrons par la suite - nécessitent chacun le choix d'un paramètre : paramètre de pénalisation pour le LASSO, paramètre de complexité pour l'arbre. Ces paramètres influencent fortement la performance du modèle, par l'arbitrage entre sous- et surapprentissage qu'ils déterminent. Comment choisir ces paramètres ? Si l'on optimise ces paramètres sur la base d'apprentissage, c'est-à-dire si l'on maximise la performance de prédictions faites sur une base qui a servi à l'ajustement du modèle, les valeurs sélectionnées conduiront à un surapprentissage maximal. Nous pouvons ainsi aboutir à ce que tous les résidus soient nuls dans le cas d'un arbre : le modèle prédit exactement les valeurs observées. Il est fort probable qu'un tel modèle soit peu performant sur la base de test. Si nous optimisons ces paramètres sur la base de test, nous ne surapprenons plus les données de la base d'apprentissage, mais risquons de surapprendre la base de test : en choisissant le paramètre qui conduise à la meilleure performance sur ces données, nous sommes là encore susceptibles de générer une performance qui ne soit pas généralisable à de nouvelles données. On a donc souvent recours à une troisième base.

On divise ainsi classiquement les données en trois groupes : une base d'apprentissage, sur laquelle le modèle est ajusté, une base de calibrage, qui permet si le modèle dépend d'un paramètre de choisir la valeur optimale en termes de performance, et une base de test sur laquelle on mesure la performance du modèle calibré. Leur répartition est généralement de l'ordre de 70% apprentissage - 20% calibrage - 10% test.

Nous avons retenu pour calibrer les modèles une autre méthode que celle de la base de calibrage. Nous avons eu recours à la validation croisée, pour 3 raisons :

— La validation croisée est une méthode plus robuste qu'une base de calibrage. Ainsi, pour une validation croisée 10-fold, on calcule dix estimateurs de l'erreur du modèle, ce qui permet d'estimer en même temps la variance de cet estimateur.
— Nos données étant rares, nous devons préserver la taille des bases d'apprentissage et de test afin d'estimer la performance finale du modèle de façon fiable. Cette contrainte est peu compatible avec la constitution d'une base de calibrage.
— La validation croisée est implémentée dans les packages R que nous utiliserons, et peut être menée en parallélisant les calculs, ce qui rend sa durée raisonnable.

Nous pouvons ainsi nous affranchir d'une base de calibrage pour n'extraire des données qu'une base de test. Le choix du volume relatif de la base de test relève d'un arbitrage entre stabilité de l'ajustement et stabilité de la performance mesurée. Une base de test de volume trop faible génère une mesure peu fiable de la performance : la loi des grands nombres ne s'y appliquant pas, la performance mesurée est très variable avec l'échantillonnage des observations. Une base d'apprentissage de volume insuffisant conduit, elle, à des ajustements instables, du fait de la variabilité du nuage de points issus de l'échantillonnage. Nous mesurons alors de façon stable (la base de test étant plus grande) la performance d'un modèle qui lui ne l'est pas, et qui présente en outre une performance dégradée du fait d'un manque d'observations.

Le faible volume de nos données conduit à penser que ces deux problèmes seront rencontrés simultanément. Mais de ces deux écueils, l'un peut être contourné : en mesurant la performance pour différents échantillonnages apprentissage/test et en en prenant la moyenne, on fiabilise la mesure issue de la base de test même si celle-ci est de faible volume. Nous choisissons donc une base de test de 20%, laissant ainsi 80% des données à la base d'apprentissage, et nous veillons à ce que la répartition des sinistres graves suive également ce ratio (120 graves dans la base d'apprentissage, 31 dans la base de test), tout en demeurant aléatoire. La base de test ne sera jamais modifiée (hormis dans sa composition, du fait des différents échantillonnages), puisqu'elle a pour objectif de déterminer la faculté de généralisation des modèles ajustés sur la base d'apprentissage.

3.2 Modèles linéaires généralisés

Nous envisageons trois modèles linéaires généralisés. En premier lieu une régression logistique, modèle qui s'inscrit naturellement dans notre cadre binaire. Puis deux modèles prenant en compte l'exposition : une régression logistique dont la fonction de lien est modifiée, et une régression de Poisson dont les fréquences prédites seront normalisées en probabilités de survenance.

Bien que plusieurs observations puissent correspondre à un même contrat, nous faisons l'hypothèse qu'elles sont indépendantes, hypothèse raisonnable au vu des éléments que nous avons discuté partie 1.4.

3.2.1 Régression logistique

La modélisation probabiliste d'une variable binaire repose sur la loi de Bernoulli. Rappelons cette loi :

Définition 6. *La loi de Bernoulli $\mathcal{B}(\pi)$ est une distribution discrète de probabilité qui prend la valeur 1 avec la probabilité π et la valeur 0 avec la probabilité $1-\pi$. Autrement dit,*

$$\mathbb{P}(Y=y) = \begin{cases} \pi & \text{si } y=1 \\ 1-\pi & \text{si } y=0 \\ 0 & \text{sinon} \end{cases}$$

Rappelons également les deux premiers moments de la loi de Bernoulli : $\mathbb{E}(Y) = \pi$ et $\mathbb{V}(Y) = \pi[1-\pi]$. Estimer le paramètre π revient par conséquent à estimer l'espérance de Y.

Nous disposons d'une population de n individus dont nous connaissons les $Y_{i,\, i\in[[1,n]]}$. Nous supposons que π est fonction de p variables explicatives X_i, i.e. que $\pi_i = \mathbb{E}(Y_i|X_i)$. L'approche linéaire classique est inadaptée au cas binaire. En effet, elle repose sur la relation $\mathbb{E}(Y_i|X_i) = X_i'\beta$, alors que $\pi_i \in [0,1]$ et $X_i'\beta \in \mathbb{R}$. Une transformation du lien entre l'espérance conditionnelle et le prédicteur linéaire est nécessaire, ce qui nous conduit au cadre linéaire généralisé.

Bien qu'une modélisation par régression logistique puisse s'affranchir d'une présentation du cadre général des modèles linéaires généralisés (ou GLM pour *Generalized Linear Model*), nous rappelons dans l'encadré ci-dessous le socle théorique des GLM. Celui-ci nous sera utile par la suite pour appréhender la façon dont on peut traduire l'effet de la durée d'exposition des observations.

Rappel du cadre théorique des modèles linéaires généralisés

Définition 7. *Une loi de paramètres θ et ϕ appartient à la famille exponentielle si sa densité par rapport à la mesure dominante adéquate (mesure de comptage sur $\mathbb{N}$ ou mesure de Lebesgue sur $\mathbb{R}$) peut s'écrire :*

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

où $a()$, $b()$ et $c()$ sont des fonctions, et où θ est le paramètre d'intérêt (propre à chaque individu) tandis que ϕ est un paramètre de nuisance indépendant de l'individu concerné.

La loi de Bernoulli $\mathcal{B}(\pi)$ correspond au cas $\theta = log\left(\frac{\pi}{1-\pi}\right)$, $a(\phi) = 1$, $b(\theta) = log\left(1 + e^{\theta}\right)$ et $c(y,\phi) = 0$ (le paramètre ϕ est donc sans importance dans notre cas).

Propriété 3. *Pour une variable aléatoire Y dont la densité est de la forme exponentielle, alors $\mathbb{E}(Y) = b'(\theta)$ et $\mathbb{V} = b''(\theta)\phi$.*

Chacune des lois de la famille exponentielle possède une fonction de lien canonique permettant de relier l'espérance au paramètre θ. En notant $\mu = \mathbb{E}(Y)$, la fonction de lien canonique est définie telle que $g_*(\mu) = \theta$. Or $\mu = b'(\theta)$, donc $g_*(.) = b'(.)^{-1}$. Pour la loi de Bernoulli $\pi = \mu$; nous avons donc trivialement $\theta = g_*(\mu) = log\left(\frac{\mu}{1-\mu}\right)$.

Définition 8. *Un modèle linéaire généralisé est composé de trois éléments, à savoir :*

1. *de variables à expliquer $Y_{i,\ i\in[[1,n]]}$ dont la loi est dans la famille exponentielle*
2. *d'un ensemble de paramètres $\beta = (\beta_1, ..., \beta_p)' \in \mathbb{R}^p$ et de variables explicatives $X_1, ..X_p$. la matrice $(n \times p)$ X est supposée être de rang p, (i.e. la matrice carrée $X'X$ est supposée inversible).*
3. *d'une fonction de lien g telle que $g(\mu_i) = x_i'\beta$ où $\mu_i = \mathbb{E}[Y_i]$ qui lie le prédicteur linéaire à l'espérance de Y_i.*

On suppose que conditionnellement aux variables explicatives X, les variables Y sont indépendantes et identiquement distribuées. On considère un modèle de la forme :

$$f(y_i|\theta_i,\phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right),$$

où l'on suppose que $g(\mu_i) = X_i'\beta$.

La log–vraisemblance d'un modèle exponentiel s'écrit :

$$log\mathcal{L}(\theta_1, ..., \theta_n, \phi, y_1, ..., y_n) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right]$$

Ce qui correspond pour le cas Bernoulli à :

$$log\mathcal{L}(\pi_1, ..., \pi_n, y_1, ..., y_n) = \sum_{i=1}^{n} y_i log(\pi_i) + (1 - y_i)log(1 - \pi_i)$$

$$log\mathcal{L}(x_1, ..., x_n, y_1, ..., y_n) = \sum_{i=1}^{n} y_i X_i'\beta - log\left(1 + e^{X_i'\beta}\right)$$

Les conditions du premier ordre n'apportent pas de solution analytique ; il est nécessaire d'avoir recours à des méthodes numériques de descente de gradient comme l'algorithme de Newton-Raphson.

3.2.2 Régression logistique modifiée - Prise en compte de l'exposition

A la différence du modèle de fréquence poissonnien avec fonction de lien logarithmique, où la prise en compte de l'exposition se fait de façon directe, le modèle logistique est plus complexe, ce qui constitue un problème fréquemment laissé de côté dans les travaux consultés, au profit d'un hypothèse d'exposition uniforme des observations.

Dans un article, A. Charpentier [5] propose de choisir une fonction de lien logarithmique et de placer le logarithme de l'exposition en offset (un offset est une variable dont le coefficient est contraint à la valeur 1). Il montre que si l'une des deux probabilités - $\mathbb{P}(Y=0)$ ou $\mathbb{P}(Y=1)$ - est proche de 0, la formule de calcul peut se simplifier par développement limité, l'exposition n'étant plus en puissance mais en facteur.

Considérons la variable Y indicatrice d'un sinistre grave (au moins) sur une période de durée E (E étant l'exposition, exprimée en nombre de jours). Notons $Y = \mathbb{1}_{\sum_{j=1}^{E} N_j \geq 1}$, où N_j est la variable indicatrice d'un sinistre grave le jour j. Supposons que le processus de survenance d'un grave est Poissonien (absence de mémoire), i.e. que les N_j sont indépendants, et qu'il n'y a pas de saisonnalité. L'hypothèse d'indépendance est fausse dans la mesure où la survenance d'un grave empêche la survenance d'un autre dans les mois qui suivent (absence d'activité). Mais la probabilité d'un grave étant très faible, celle de deux graves rapprochés est quasiment nulle, et l'hypothèse a peu d'importance. Ainsi les N_j sont IID de loi de Bernoulli $\mathcal{B}(\pi)$ où π est la probabilité quotidienne d'un grave. Nous avons alors :

$$\begin{aligned}
\mathbb{P}(Y=0) &= \mathbb{P}\left(\mathbb{1}_{\sum_{j=1}^{E} N_j \geq 1} = 0\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{E} N_j = 0\right) \\
&= \mathbb{P}\left(\cap_{j=1}^{E} N_j = 0\right) \\
&= \prod_{j=1}^{E} \left(\mathbb{P}(N_j = 0)\right) \\
&= \mathbb{P}(N=0)^E = (1-\pi)^E
\end{aligned}$$

Supposons que $\pi \approx 0$: c'est le cas de nos données, la fréquence des graves étant très faible. Par développement limité, $\mathbb{P}(Y=0) \approx 1 - \pi E$, ce qui donne $\mathbb{P}(Y=1) \approx \pi E$.

Conditionnellement à X, nous avons donc :

$$\mathbb{P}(Y=1|X,E) = E.\mathbb{P}(N=1|X)$$

ou encore

$$\mathbb{E}(Y|X,E) = E.\mathbb{E}(N|X)$$

La contribution de de l'article d'A. Charpentier [5], unique ressource bibliographique que nous ayons trouvée sur le sujet, se limite à ce point. Il n'y est pas démontré que la nouvelle spécification du modèle permette le calcul d'une vraisemblance adaptée à la formule de probabilité simplifiée. Nous avons développé les deux expressions de vraisemblance - celle issue de la formule de probabilité, et celle estimée par le modèle - afin d'en vérifier la correspondance.

La vraisemblance du modèle s'écrit :

$$\begin{aligned}\mathcal{L}(y_i, x_i, i = 1, ..., n) &= \prod_{i=1}^{n} y_i \, \mathbb{P}(Y_i = 1) + (1 - y_i)\mathbb{P}(Y_i = 0) \\ &= \prod_{i=1}^{n} y_i \, E_i \, \mathbb{P}(N_i = 1) + (1 - y_i)(1 - E_i \, \mathbb{P}(N_i = 1)) \\ &= \prod_{i=1}^{n} y_i \, E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} + (1 - y_i) \left(1 - E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)\end{aligned}$$

Et sa log-vraisemblance :

$$\begin{aligned}log(\mathcal{L}) &= \sum_{i=1}^{n} log \left(y_i \, E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} + (1 - y_i) \, \left(1 - E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)\right) \\ &= \sum_{i=1}^{n} y_i \, log \left(E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) + (1 - y_i) \, log \left(1 - E_i \, \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)\end{aligned} \tag{3.1}$$

Développons à présent l'expression de la vraisemblance estimée par la régression logistique modifiée : un modèle linéaire généralisé avec loi de Bernoulli, fonction de lien logarithmique (et non plus logit) et logarithme de l'exposition en offset.

Nous avons vu que la vraisemblance associée à la loi de Bernoulli vaut :

$$\log(\mathcal{L}) = \sum_{i=1}^{n} y_i \, log(\pi_i) + (1 - y_i) \, log(1 - \pi_i)$$

Et que la fonction de lien g définit la relation $g(\pi) = X'\beta$, soit avec une fonction de lien logarithmique $\pi = e^{X'\beta}$.

$$log(\mathcal{L}) = \sum_{i=1}^{n} y_i \, x_i'\beta + (1 - y_i) \, log \left(1 - e^{x_i'\beta}\right)$$

En ajoutant le logarithme de l'exposition en offset dans le prédicteur linéaire, nous obtenons :

$$\sum_{i=1}^{n} y_i \, (log(E_i) + x_i'\beta) + (1 - y_i) \, log \left(1 - E_i \, e^{x_i'\beta}\right) \tag{3.2}$$

L'équivalence entre les vraisemblances (3.1) et (3.2) repose sur l'approximation :

$$\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \approx e^{x_i'\beta}$$

C'est-à-dire $\left(e^{x_i'\beta}\right)^2 \approx 0$. Comme $\hat{\pi} = e^{x_i'\beta}$ est faible (l'estimé est proche de l'observé π, lui-même proche de 0), l'approximation est valable.

Ainsi la régression logistique avec une fonction de lien logarithmique et la log-Exposition en offset permet d'intégrer l'exposition à l'ajustement du modèle, sous réserve que l'hypothèse $\pi \approx 0$ soit maintenue.

3.2.3 Régression de Poisson

Il est plus aisé d'intégrer l'effet de l'exposition dans un modèle de fréquence que dans un modèle binaire. Nos données ne présentent pas de surdispersion (le rapport de l'espérance et de la variance empiriques est très proche de 1), aussi nous retenons une régression de Poisson. La fonction de lien canonique associée à la loi de Poisson est le logarithme, qui permet de passer l'exposition en offset dans le prédicteur linéaire. Nous avons :

$$Y_i \sim \mathcal{P}(\lambda_i \cdot E_i) \quad \text{avec} \quad \lambda_i = exp[X_i'\beta]$$

que l'on peut écrire

$$Y_i \sim \mathcal{P}(\tilde{\lambda}_i) \quad \text{avec} \quad \tilde{\lambda}_i = E_i \cdot exp[X_i'\beta] = exp[X_i'\beta + log(E_i)]$$

À la différence de la régression logistique, il n'est ici besoin d'aucune hypothèse, et la fonction de lien n'est pas modifiée. Du fait de sa simplicité, ce modèle est couramment utilisé.

Le modèle de Poisson est un modèle de fréquence, qui ne répond pas directement à notre recherche de classifieur binaire : il produit un vecteur de nombre de sinistres prédits, et non de probabilités de survenance d'un sinistre. Pour autant, nous avons vu que les probabilités prédites avaient peu de sens en elles-mêmes, et que l'important réside dans l'ordonnancement des observations. Nous proposons ainsi d'utiliser la régression de Poisson pour construire un vecteur de probabilités, en normalisant le vecteur des fréquences prédites entre 0 et 1, transformation linéaire qui en préserve l'ordre.

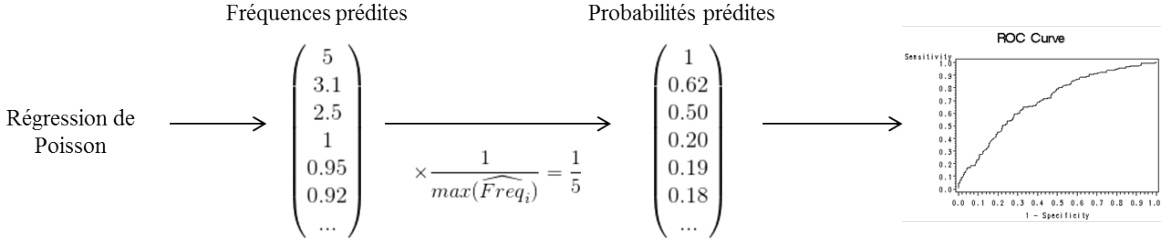


FIGURE 3.3 – Normalisation des fréquences prédites par la régression de Poisson - Obtention d'un classifieur binaire

3.3 Arbres de classification et de régression

Les arbres de classification et de régression présentent plusieurs avantages par rapport aux modèles linéaires :

— Ils modélisent spontanément les non-linéarités et les interactions, ce qui n'est possible dans un GLM qu'à condition d'améliorer le prédicteur linéaire.

— Ils sont susceptibles d'isoler des groupes de sursinistralité significative à partir d'un nombre restreint de règles de décision, ce qui présente un grand intérêt opérationnel : on peut alors envisager de filtrer le portefeuille en ciblant une population de faible effectif et de niveau de risque élevé, que la tarification issue du modèle linéaire ne discrimine pas suffisamment.

Ces avantages vont de pair avec certains inconvénients bien connus, tels que la faible stabilité du modèle et un biais vers la sélection des variables catégorielles lorsqu'elles présentent un nombre de modalités élevé.

3.3.1 Arbre de classification

Nous ne rappelons pas ici le cadre théorique des modèles CART (Classification And Regression Trees), qui est abondamment documenté, et renvoyons à titre d'exemple vers l'article de Therneau et Atkinson [15], régulièrement mis à jour. Développons en revanche un point essentiel, bien que souvent laissé de côté en pratique : le contrôle du surapprentissage au sein des arbres.

La taille d'un CART est contrôlée par son paramètre de complexité (noté cp pour *complexity parameter*) : pour être réalisé (lors de la construction)/conservé (lors de l'élagage), un noeud doit permettre une réduction de l'indice d'impureté supérieur à la valeur du cp. Il est largement admis qu'une bonne modélisation par arbre s'effectue en deux temps : on commence par construire un arbre profond, dont les feuilles contiennent un faible nombre d'observations (voire une seule), puis on élague cet arbre en retirant les noeuds de réduction d'impureté inférieure au paramètre de complexité. Cette approche est préférable à une construction de l'arbre directement limitée par le cp : il peut arriver qu'un noeud de faible réduction d'impureté conduise à un noeud inférieur de forte réduction d'impureté. L'élagage permet alors de préserver ce noeud.

Comment choisir le paramètre de complexité ? Trop faible, il conserve un arbre profond qui sur-apprend. Trop élevé, il élague lourdement l'arbre et n'exploite pas suffisamment le potentiel des données. Ce paramètre est donc calibré par validation croisée, méthode d'arbitrage biais-variance courante en machine learning :

— on définit une grille de valeurs pour le paramètre de complexité.

— pour chaque valeur de cp, on mesure la performance des arbres élagués avec ce cp par validation croisée.

— on trace l'évolution de la performance prédictive des arbres avec le cp.

On peut alors choisir de retenir la valeur de cp minimisant l'erreur de validation croisée. Pour autant, la mesure de cette erreur commet elle-même des erreurs, et présente des variations selon

l'échantillon considéré dans la validation croisée. Cette variabilité est représentée par les barres verticales à chaque observation, dont la longueur est égale à deux fois l'écart-type de la mesure de l'erreur de validation croisée.

Afin de s'affranchir avec quasi-certitude du surapprentissage, une règle consiste à retenir la valeur de cp la plus élevée dont l'erreur de validation croisée ne dépasse pas le plafond défini par l'ajout d'un écart-type à l'erreur de validation croisée minimale observée. Cette règle est appelée "first-standard error rule" et notée 1SE.

De même que les modèles linéaires, l'arbre de classification fournit un vecteur de probabilités, une pour chaque observation, correspondant à la proportion de positifs au sein de la feuille contenant cette observation. En ce sens, la probabilité fournie par un arbre est plus interprétable que celle d'un GLM. Notons que l'on rencontre dans certaines publications des sorties d'arbres n'indiquant pas une probabilité, mais une valeur 0 ou 1 : ceci résulte d'une simplification du vecteur de probabilités avec un seuil (s = 0.5 par défaut dans rpart).

3.3.2 Random forest

Les random forests - ou forêts aléatoires - sont des agrégats d'arbres construit sur des échantillons de la base obtenus par bootstrap (tirage avec remise d'observations dans la base). Cette méthode est connue sous le nom plus général de *bagging*, pour *bootstrap aggregating*.

Les random forests sont susceptibles d'améliorer sensiblement la performance de l'arbre sous-jacent : en agrégeant un aléa imposé aux observations (bootstrap) et aux variables - un nombre limité de variables est choisi aléatoirement à chacun des noeuds de chaque arbre - elles rendent les prédictions plus stables, et ne surapprennent pas : lorsqu'on augmente le nombre d'arbres de la forêt, celle-ci converge vers une valeur limite de l'erreur de prédiction [3].

Nous incluons la random forest parmi les modèles étudiés, tant pour son rôle de benchmark quant à la performance prédictive accessible que pour son intérêt dans l'amélioration du modèle linéaire par interactions et non linéarités (Partie 6.7).

3.3.3 Arbre de régression de Poisson

Therneau et Atkinson [15] ont développé une adaptation de l'algorithme CART pour les modèles de fréquence, qui repose sur la régression de Poisson. Cet algorithme, nommé arbre de régression de Poisson, permet d'intégrer l'exposition à la construction de l'arbre.

Pour développer cet arbre, Therneau et Atkinson s'appuient sur le fait que les arbres de régression standards sont performants malgré le fait que leur critère de séparation soit la réduction de la somme des carrés, une mesure à robustesse limitée. Ils retiennent de même la méthode valide la plus simple possible. Notons c_i le nombre d'événements de l'observation i, et e_i sa durée d'exposition.

La variable Y du programme est une matrice à deux colonnes. L'algorithme est construit sur les règles suivantes :

— On définit au sein de chaque noeud la fréquence observée des individus de ce noeud :

$$\hat{\lambda} = \frac{Nombre \quad d'événements}{Exposition \quad totale} = \frac{\sum_i c_i}{\sum_i e_i}$$

— L'erreur d'un noeud est mesurée par la déviance au sein de ce noeud :

$$D = \sum \left[c_i \cdot log\left(\frac{c_i}{\hat{\lambda} \cdot e_i}\right) - (c_i - \hat{\lambda} \cdot e_i) \right]$$

— Le critère de division du noeud est la statistique de test pour deux groupes d'une loi de Poisson :

$$D_{\text{noeud parent}} - (D_{\text{noeud fils gauche}} - D_{\text{noeud fils droit}})$$

— L'erreur de prédiction sur une nouvelle observation est la contribution de cette observation à la déviance, en conservant la fréquence prédite $\hat{\lambda}$.

Cette simplicité souffre d'un défaut : elle peut rencontrer un problème numérique lors des validations croisées. Supposons qu'un noeud contienne n observations, dont une seulement présente un (ou plusieurs) événements. Ce cas est fréquent même en présence de données équilibrées : l'arbre purifie les groupes, conduisant à certains noeuds de très faible fréquence observée. L'estimateur de l'erreur (i.e. la déviance) de validation croisée présentera un échantillon d'apprentissage avec $\hat{\lambda} = 0$ lorsque l'observation avec un événement se trouve dans l'échantillon de test. La contribution de l'événement à la déviance est alors $c_i \cdot log(c_i/0)$, infinie puisque $c_i > 0$. L'erreur de prédiction du modèle - qui est la moyenne de l'ensemble des contributions à la déviance mesurée au cours de la validation croisée - est infinie à son tour : le modèle devient "infiniment mauvais".

Pour contourner ce problème, Therneau et Atkinson retiennent un estimateur réduit (*shrinkage estimator*) correspondant à l'estimateur de Bayes d'une distribution Gamma :

$$\hat{\lambda}_k = \frac{\alpha + \sum c_i}{\beta + \sum e_i} \quad \text{où} \quad \alpha = \frac{1}{k^2} \quad \text{et} \quad \beta = \frac{\alpha}{\hat{\lambda}}$$

La valeur par défaut du paramètre k dans rpart est de 1.

De même que le GLM Poisson, l'arbre prédit ici une fréquence pour chaque nouvelle observation, égale à la fréquence observée au sein de la feuille qui lui est assignée. Nous procédons par conséquent à une normalisation du vecteur de fréquence en vecteur de probabilités, à l'identique de ce que nous avons présenté dans le cas du GLM Poisson, pour mesurer sa performance par l'AUC.

Chapitre 4

Sélection et retraitement de variables

4.1 Sélection de variables

Réduire le nombre de coefficients non-nuls du prédicteur linéaire dans un modèle linéaire généralisé est utile à plusieurs titres. Un premier est la qualité de prédiction : réduire le nombre de coefficients limite la variance des valeurs prédites, ce qui est susceptible d'augmenter le pouvoir prédictif du modèle. Un deuxième est l'interprétabilité des résultats : nous cherchons en général à identifier un petit nombre de variables ayant un fort pouvoir prédictif. Un dernier est de limiter le temps de calcul : l'ajout d'une contrainte sur les coefficients accélère la convergence de l'algorithme de maximisation de vraisemblance.

4.1.1 Procédure stepwise et p-hacking

La procédure stepwise est une pratique de sélection de variables très courante pour les modèles généralisés. La méthode forward stepwise (principale variante) consiste, partant d'un modèle sans variable, à l'augmenter en ajoutant itérativement la variable qui améliore le plus le modèle, sous réserve qu'elle valide un test basé sur un critère de sélection (tests de Student, de Fisher, coefficient de détermination ajusté, ou critères d'information d'Akaike, de Bayes). Elle souffre pour autant d'importants défauts, montrés par différents travaux de recherche (Harrell, 1996[10]) :

— Les hypothèses de distribution sur lesquels s'appuient les tests de significativité (test de Fisher et du Chi-deux) entre deux itérations ne sont pas valides.
— Les p-values du modèle final n'ont par conséquent plus la même signification, et il est difficile de les corriger.
— Les intervalles de confiance obtenus pour les coefficients et le valeurs prédites sont artificiellement étroits : la procédure "s'auto–valide".
— Les coefficients des variables conservées sont biaisés à la hausse.

Pour ces différentes raisons, connues sous le nom de "p-hacking", et d'autres que nous développons ci-dessous, nous expérimentons et privilégions la méthode de régularisation LASSO. Les résultats du LASSO seront in fine comparés à ceux du stepwise lors de la mise en oeuvre opérationnelle sur

l'outil Emblem.

4.1.2 Régression pénalisée - LASSO

Le LASSO (Least Absolute Shrinkage and Selection Operator) est une méthode de réduction des coefficients de régression développée par Robert Tibshirani (1996,[16]). Appliquée aux modèles linéaires généralisés, elle consiste à introduire dans la fonction de vraisemblance une pénalisation proportionnelle à la norme L_1 du vecteur β des coefficients (i.e. à la somme des valeurs absolues des coefficients).

Dans le cas du modèle logistique, on maximise alors la vraisemblance :

$$\frac{1}{N}\sum_{i=1}^{N}\left\{y_i(\beta_0+\beta' x_i) - log\left(1+e^{\beta_0+\beta' x_i}\right)\right\} - \lambda\|\beta\|_1$$

La pénalisation des coefficients par leur norme requiert une transformation préalable : ceux-ci doivent être à la même échelle. En effet une variable comprise entre 0 et 1000 a un coefficient plus faible que cette même variable normalisée entre 0 et 1. La première contribue moins au $\|\beta\|_1$ que la seconde, et la pénalisation par la norme du coefficient conduit à privilégier une variable de même valeur informative mais d'échelle plus grande. Pour cette raison toutes nos variables discrètes et continues sont normalisées entre 0 et 1. On peut également choisir de centrer-réduire les variables (Hastie, 2015, [11]), choix équivalent en termes de résultats.

Le paramètre λ de la fonction de vraisemblance contrôle la complexité du modèle : de faibles valeurs de λ "libèrent" davantage les coefficients et permettent au modèle de s'ajuster plus finement aux données, tandis qu'un λ élevé conduit à un modèle plus parcimonieux, moins ajusté et plus simple d'interprétation. Ce paramètre détermine la faculté de généralisation du modèle : d'un modèle trop simple au surapprentissage, il est nécessaire de trouver le juste équilibre, à partir d'une validation croisée.

Nous utilisons ici les fonctions du package R *glmnet*. Ce package permet d'appliquer le LASSO aux différentes lois des modèles linéaires généralisés, mais pas de modifier leur fonction de lien (qui demeure la fonction de lien canonique). À notre connaissance, il n'existe de fonction pré-implémentée permettant cette modification ni sous R, ni sous Python. La sélection de variable par LASSO dans la régression logistique modifiée ne sera donc pas possible.

Présentons la méthode d'analyse usuelle d'une régression pénalisée LASSO à partir d'un jeu de données disponible dans le package R *glmnet*. On représente (Figure 4.1) le tracé des valeurs de chaque coefficient pour différentes valeurs de λ, en fonction de la norme L_1 du vecteur des coefficients, $\|\beta\|_1$. Afin de développer la partie gauche du graphe, on peut aussi tracer les valeurs des coefficients en fonction de la part de déviance (Figure 4.2) expliquée plutôt qu'en fonction de λ. La part de déviance expliquée par les données (d'apprentissage) est :

$$D^2_\lambda = \frac{Dev_{null} - Dev_\lambda}{Dev_{null}}$$

Ici la déviance Dev_λ est définie comme moins deux fois la différence de log-vraisemblance entre un modèle ajusté avec le paramètre λ et le modèle "saturé" (i.e. avec $\hat{y}_i = y_i$). Dev_{null} est la déviance nulle, calculée sur un modèle constant ($\hat{y}_i = \hat{y}$).

Représenter les coefficients en fonction de la part de déviance expliquée permet de développer la partie du graphe correspondant aux faibles valeurs de $\|\beta\|_1$, la plus intéressante puisque c'est celle où s'opère la sélection des variables prépondérantes. En effet on observe que la fraction de déviance est une fonction concave de $\|\beta\|_1$, ce qui vient du fait qu'elle est calculée à partir de la log-vraisemblance (et non de la vraisemblance).

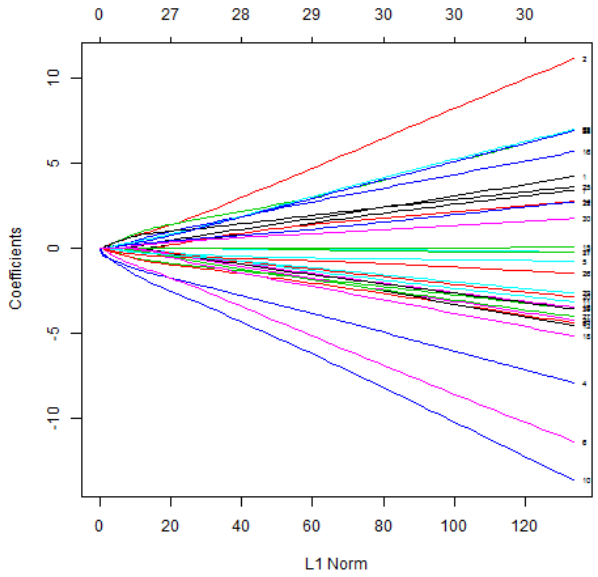


FIGURE 4.1 – Valeurs des coefficients en fonction de la norme $\|\beta\|_1$

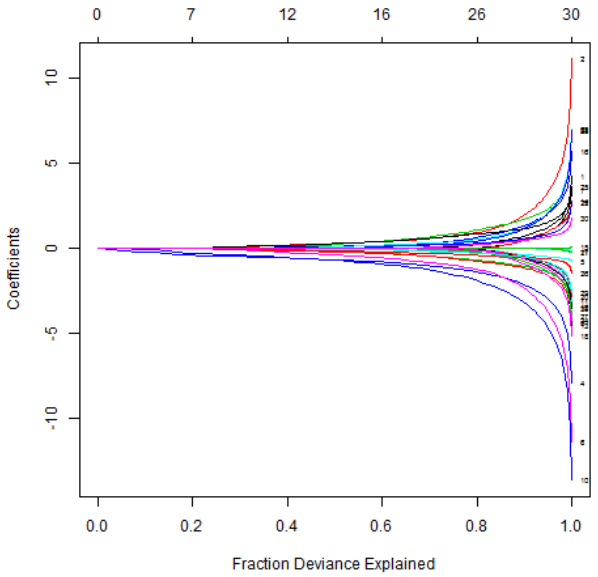


FIGURE 4.2 – Valeurs des coefficients en fonction de la fraction de déviance

La figure 4.3 montre la courbe de validation croisée pour les différentes valeurs de λ. Nous représentons l'erreur quadratique moyenne estimée (MSE, *mean-standard error*) et une barre d'erreur d'amplitude égale à deux fois l'erreur standard des estimateurs de validation croisée. La première barre verticale - intersectant la courbe à son minimum - indique la valeur de λ pour laquelle la MSE est la plus faible. La seconde barre (à droite) indique la valeur de λ correspondant à la *one standard error rule* : il s'agit de la plus grande valeur de λ pour laquelle la MSE ne dépasse pas sa valeur minimale de plus d'un écart-type. Cette valeur de λ correspond au plus haut niveau de réduction des variables que nous puissions nous permettre sans réduire significativement notre capacité prédictive.

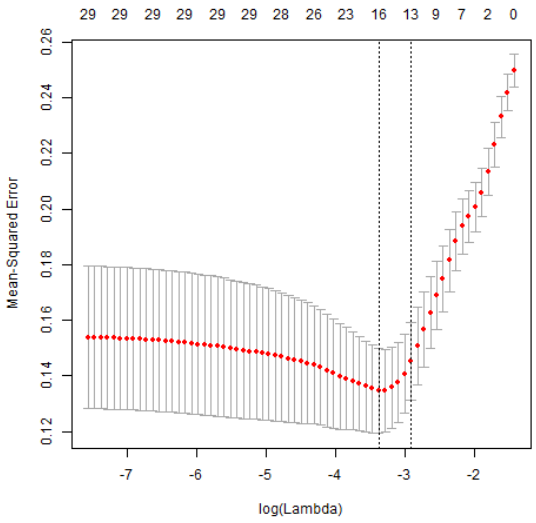


FIGURE 4.3 – Erreur quadratique moyenne de validation croisée en fonction de $log(\lambda)$

Nous utiliserons dans cette étude la *one standard error rule* : elle permet de construire des modèles parcimonieux, ce qui correspond aux modélisations que nous souhaitons obtenir in fine. Nous disposons ainsi d'une estimation plus juste de la performance finale, tout en conservant un volume assez important de variables (généralement une quarantaine).

4.1.3 Avantages du LASSO

Le LASSO possède deux avantages statistiques principaux :

— Une sélection parcimonieuse : le LASSO permet de sélectionner un sous-ensemble restreint de variables.

— Une sélection consistante : lorsque seul un sous-ensemble de variables est véritablement utilisé pour la prédiction, le LASSO est en mesure de sélectionner ces variables d'intérêt avant toute autre variable (Zhao & Yu, 2006 [17]). Ceci est vrai en l'absence de fortes corrélations : si des variables d'intérêt sont fortement corrélées avec d'autres, la consistance du LASSO n'est plus assurée. D'autre part, si des variables d'intérêt sont fortement corrélées entre elles, le LASSO en privilégiera une au détriment des autres. Sur nos données, ce problème ne se présentera que pour les variables externes et avec un impact mineur en termes opérationnels : les variables externes n'étant par nature pas contrôlables par l'assureur, elles sont avant tout utilisées pour leur pouvoir prédictif.

Le LASSO constitue aussi un outil pratique dans notre approche. Partant de données très déséquilibrées et de faible volume, notre travail consiste avant tout à développer la performance prédictive de modèles linéaires au départ très mauvais. Nous procédons à de nombreux ajustements de modèles, tant pour l'optimisation des paramètres que pour la fiabilisation des mesures

de performance : nous souhaitons par conséquent un outil de sélection de variables automatisé, rapide (la pénalisation de la régression accélère la convergence) et orienté vers la maximisation de la performance prédictive. Le LASSO remplit ces trois conditions.

4.2 Retraitement des variables catégorielles

Parmi les variables catégorielles à notre disposition, deux ne peuvent être employées telles quelles dans les modèles, du fait de leur trop grand nombre de modalités. Il s'agit des variables indiquant l'activité principale (231 modalités, issues des nombreux champs professionnels couverts par le produit multirisque) et le département (94 modalités). Les variables à grand nombre de modalités posent problème à plusieurs titres :

- Elles ne sont pas souhaitables pour les CART. Il y a $2^{m-1}-1$ façons de diviser une variable à m modalités en deux groupes ; le nombre de choix augmente exponentiellement avec m. Il existe des moyens d'optimiser la recherche de la meilleure partition (Breiman et al, 1984 [4]) pour limiter le temps de calcul, mais le vrai problème est statistique : plus il y a de partitions, plus il est probable que l'une d'elles permette un bon ajustement des données, qui soit en réalité un surapprentissage. Ce surapprentissage ne peut être corrigé en élaguant, la variable étant généralement utilisée dès les premiers noeuds de l'arbre.

- Elles sont transformées en variables indicatrices au sein des modèles linéaires, et une variable à m modalités génère $m-1$ indicatrices (la première modalité étant posée en référence). Ceci alourdit considérablement la lecture du modèle, et rend son interprétation fastidieuse. Notons par ailleurs que sur les CART, transformer les variables catégorielles en $m-1$ indicatrices réduit sensiblement la performance : un essai réalisé sur les random forests conduit à des pertes de plus de 6 points d'AUC.

- Elles présentent souvent des modalités "orphelines" de faible effectif, susceptibles lors de l'échantillonnage d'être intégrées à la base de test uniquement. Ce phénomène rend impossible la prédiction des observations concernées à partir du modèle ajusté sur la base d'apprentissage, et génère une erreur de programmation.

- Certains algorithmes implémentés sous R (le SMOTE notamment) ne sont paramétrés que pour des variables à nombre de modalités restreint - couramment une cinquantaine de modalités.

Qu'il s'agisse d'améliorer l'apprentissage des arbres ou l'interprétabilité des modèles linéaires généralisés, il est nécessaire de passer à une granularité moins fine sur ces variables - en procédant à des regroupements - ou de les retirer. Retirer ces variables semble peu pertinent, dans la mesure où elles sont susceptibles d'être fortement corrélées (au sens de la statistique du X^2) à des variables non présentes dans le modèle, et d'être de ce fait explicatives. Supposons que quelques activités

emploient depuis quelques années de nouveaux produits chimiques, hautement inflammables, dont la présence n'est pas encore renseignée à la souscription. Il est utile de pouvoir grouper ces activités et mesurer l'effet pur de l'appartenance à ce groupe.

Comment alors regrouper de façon pertinente les modalités ? Lorsqu'il existe un lien entre la variable à regrouper et les autres variables explicatives, regrouper directement les modalités par fréquences observées n'est pas judicieux. Prenons le cas des activités : elles diffèrent sur certaines variables explicatives de la sinistralité (surface, valeur de contenu etc.), et regrouper par fréquences observées conduit à réunir les activités structurellement plus risquées ensemble. On maintient ainsi le lien entre la variable activité et les autres variables explicatives. Ceci ne permettra pas de capter l'effet de variables non-observées, et risque en outre de fournir un résultat peu utile au sein des modèles linéaires : quelle est l'utilité d'un groupe dont le coefficient est porté par sa forte corrélation aux autres variables explicatives ? L'objectif n'est pas de pénaliser une activité globalement risquée avec un tarif général élevé, mais de comprendre et discriminer le risque au sein de cette activité.

La méthode la plus utile pour regrouper les modalités semble donc être de les regrouper par proximité d'effet pur, afin d'en extraire l'information cachée, plutôt que d'en faire des agrégats de variables déjà présentes dans le modèle.

4.2.1 Regroupement de modalités par arbre

La simplification des variables catégorielles par arbre semble en développement dans l'enseignement et la pratique actuarielle, sous la forme suivante : on explique la variable à prédire par la variable à regrouper, et l'on récupère les groupes ainsi obtenus. Quelle est la validité statistique de cette pratique, dans l'objectif de former des groupes de même effet pur ?

Etudions-la à partir d'un exemple simple, présenté figure 4.4 : supposons que les deux principales variables expliquant les graves soient la surface, et une variable inobservée, la présence de fours. Considérons un portefeuille avec 10 boulangeries, 40 pharmacies et 50 cordonneries. Les boulangeries et les pharmacies ont une surface moyenne de 100 m^2, les cordonneries une surface moyenne de 30 m^2. Boulangerie et pharmacie ont des fréquences élevées, celle des boulangeries étant la plus grande du fait de la présence de fours. L'effet pur de l'activité, corrigé de l'effet de la variable explicative connue (la surface) est déterminé par la présence de fours : il est positif pour les boulangeries, et négatif pour les pharmacies et les cordonneries. Nous souhaitons par conséquent disposer d'une méthode qui sache regrouper pharmacies et cordonneries et isoler les boulangeries. Nous supposons que les fréquences varient de façon raisonnable et selon une distribution standard (gaussienne par exemple), de sorte qu'il n'y ait pas d'effets inattendus dans les sommes de carrés des écarts à la moyenne, et que l'on puisse raisonner directement à partir des moyennes des groupes.

Examinons le résultat obtenu à l'aide d'un arbre de régression. Nous savons que construire un arbre de régression sur des fréquences n'est pas pleinement correct lorsque l'exposition est variable, mais nous retenons cet algorithme pour la simplicité de sa fonction de séparation des noeuds (somme des carrés) à vertu illustrative, et supposons l'exposition constante dans cet exemple. Les résultats

Activité	Nombre de contrats	Surface moyenne	Variable inobservée – Présence de fours	Fréquence moyenne de graves
Boulangerie	10	100 m²	Oui	0.5
Pharmacie	40	100 m²	Non	0.4
Cordonnerie	50	30 m²	Non	0.1

FIGURE 4.4 – Données illustratives pour les regroupements par arbre - Première version

que nous montrons sont selon toute vraisemblance transposables au cas de l'arbre classification, ou à celui de l'arbre de régression de Poisson.

Il apparaît immédiatement (Figure 4.5) que la méthode de regroupement par arbre simple n'aboutit pas au but recherché : pharmacies et boulangeries sont groupées ensemble du fait de leur relation (au sens du X^2) à la variable surface.

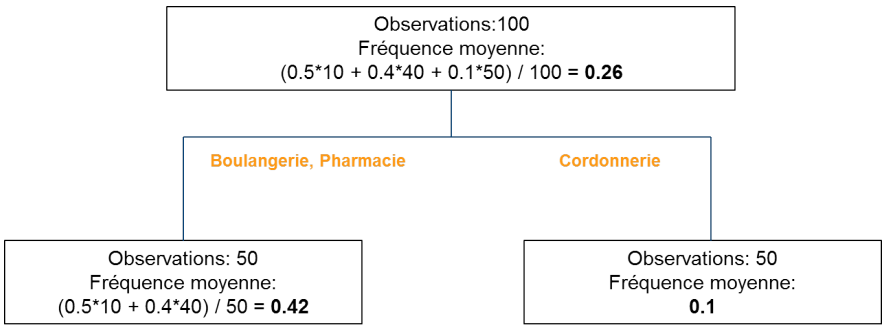


FIGURE 4.5 – Regroupement de modalités par arbre de régression sur la fréquence de sinistres

4.2.2 Regroupement de modalités par arbre - Méthode améliorée

Essayons de corriger ce problème par une méthode courante en apprentissage statistique, qui consiste dans un premier temps à modéliser la variable à prédire par la (les) variable(s) explicative(s) dont on veut retirer l'effet, puis à modéliser les résidus obtenus par la variable dont on veut capter l'effet pur. Ceci peut permettre de contourner le fait que les algorithmes tels que le CART ne permettent pas directement de mesurer les effets purs, à la façon d'un GLM.

Notons Z la variable à regrouper, la méthode consiste donc à :

— Construire un modèle prédisant des sinistres graves à partir d'un ensemble de variables explicatives et corrélées à Z.

— Récupérer les résidus de cet arbre : $fréquenceobservée - fréquenceprédite$

— Construire un arbre de régression des résidus à partir de la variable Z.

Nous supposons par commodité que les variations de surface au sein de chaque activité sont limitées, de sorte que la séparation des groupes sur la variable surface puisse pleinement isoler les activités les unes des autres (sans "fausse route" d'une observation de surface atypique).

Le premier arbre (Figure 4.6) regroupe pharmacies et boulangeries, en choisissant un seuil entre 30 et 100 m^2. On obtient alors des résidus (Figure 4.7) positifs élevés pour les boulangeries, et faiblement négatifs pour les pharmacies. Le second arbre (Figure 4.8) regroupe alors les pharmacies avec les cordonneries, de résidus nuls en moyenne. La méthode améliorée permet un regroupement adapté sur ces données.

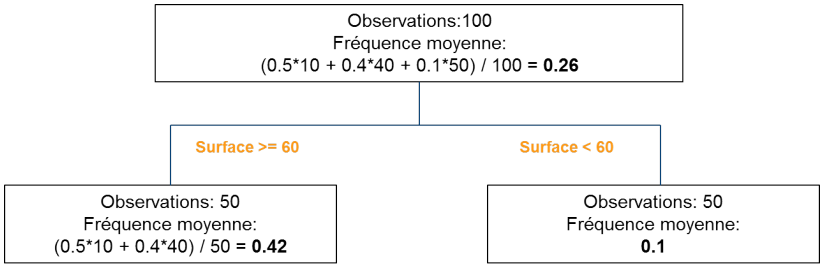


FIGURE 4.6 – Premier arbre de régression - Prédiction de la fréquence de sinistres à partir de la surface

Activité	Résidus Fréquence observée – Fréquence prédite
Boulangerie	0.50 – 0.42 = 0.08
Pharmacie	0.40 – 0.42 = - 0.02
Cordonnerie	0.10 – 0.10 = 0

FIGURE 4.7 – Résidus du premier arbre de régression

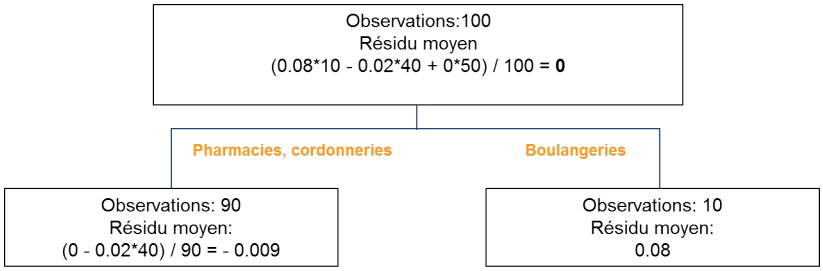


FIGURE 4.8 – Second arbre de régression - Prédiction des résidus du premier arbre à partir de l'activité

Application de la méthode à un second jeu de données Nous allons pour autant montrer que le succès de la méthode améliorée sur les données précédentes est fortuit, et ce en modifiant simplement les effectifs des groupes : nous considérons à présent (Figure 4.9) une base avec 40 boulangeries et 10 pharmacies, le nombre de cordonneries étant inchangé.

Le premier arbre (Figure 4.10) regroupe toujours pharmacies et boulangeries. Mais cette fois, les boulangeries étant en majorité dans la feuille, ce sont elles qui ont des résidus (positif) proche de 0, tandis que ceux des pharmacies s'en éloignent. Le second arbre (Figure 4.12) regroupe alors les

Activité	Nombre de contrats	Surface moyenne	Variable inobservée – Présence de fours	Fréquence moyenne de graves
Boulangerie	40	100 m²	Oui	0.5
Pharmacie	10	100 m²	Non	0.4
Cordonnerie	50	30 m²	Non	0.1

FIGURE 4.9 – Données illustratives pour les regroupements par arbre - Seconde version

boulangeries avec les cordonneries, de résidus nuls en moyenne, ce qui constitue un regroupement inadapté et ne permet pas de capter l'effet de la variable inobservée - la présence de fours.

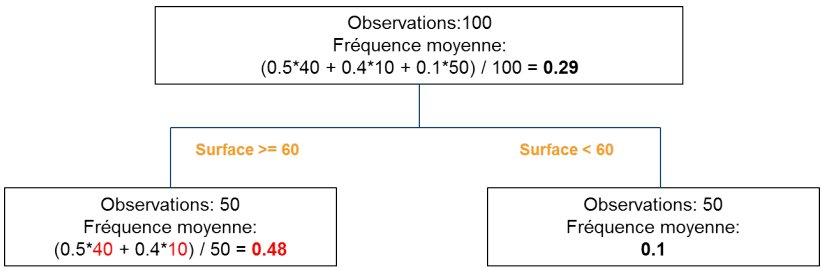


FIGURE 4.10 – Premier arbre de régression - Prédiction de la fréquence de sinistres à partir de la surface

Activité	Résidus Fréquence observée – Fréquence prédite
Boulangerie	0.50 – 0.48 = 0.02
Pharmacie	0.40 – 0.48 = - 0.08
Cordonnerie	0.10 – 0.10 = 0

FIGURE 4.11 – Résidus du premier arbre de régression - Seconde version

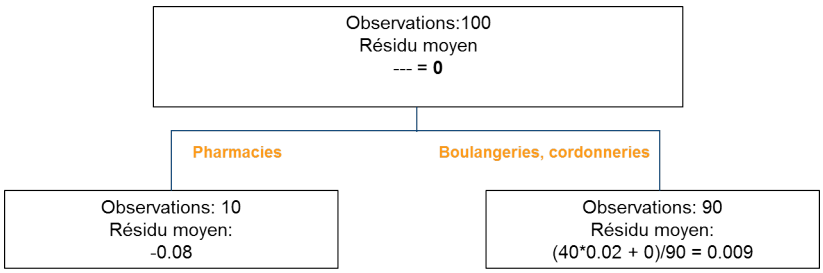


FIGURE 4.12 – Second arbre de régression - Prédiction des résidus du premier arbre à partir de l'activité

Les méthodes de regroupement par arbres nous paraissent donc inadaptées, du fait de leur difficulté à s'affranchir des corrélations pour capter l'effet pur d'une variable. Les regroupements par arbre sont d'autant moins souhaitables qu'ils sont souvent réalisés sur la base entière, avant l'échantillonnage d'une base de test, ce qui génère un surapprentissage pernicieux : les regroupements ayant été en partie déterminés sur l'échantillon de test, ils intègrent de l'information de cet

échantillon et sont artificiellement explicatifs.

4.2.3 Regroupement de modalités par GLM

Les regroupements par arbre n'étant pas satisfaisants, nous revenons aux modèles linéaires, plus adaptés pour capter l'effet pur des modalités d'une variable. Rappelons que, dans un modèle linéaire, une variable catégorielle à m modalités est transformée en $m-1$ variables indicatrices, l'une d'elle étant posée en référence - de coefficient nul par définition. En intégrant l'ensemble des variables explicatives au modèle, on peut interpéter le coefficient associé à chaque modalité comme l'effet pur de cette modalité, soit l'effet intrinsèque des variables inobservées sur la sous-population de cette modalité. On regroupe alors les modalités par "tranches" de coefficients de valeurs proches. Nous choisissons pour le regroupement des modalités un GLM qui prenne en compte l'exposition : la régression de Poisson (dont nous verrons par la suite qu'elle est le meilleur modèle linéaire pour nos données). En effet, du fait de la rareté des graves, de nombreuses modalités ne présentent aucun sinistre, et auront un coefficient de même valeur si l'on ne prend pas en compte l'exposition. Or entre deux modalités n'ayant aucun sinistre en historique, une modalité d'exposition cumulée dix fois supérieure à l'autre est significativement moins risquée. La modalité de faible exposition cumulée est, elle, "acquittée faute de preuve". Un modèle qui prend en compte l'exposition permet d'ordonner les coefficients des modalités selon leur exposition cumulée en portefeuille. Nous conservons en outre la pénalisation LASSO : tout en prévenant le surapprentissage, elle permet d'identifier directement un groupe de modalités non-significativement différentes de la modalité de référence, celles de coefficient nul.

Prenons l'exemple de la variable activité (Figure 4.13), avec une modalité de référence (REF) de même risque pur que les pharmacies. Les pharmacies ont un coefficient nul, de même que les cordonneries, de même effet pur. Les boulangeries ont un effet pur supérieur, et ressortent avec un coefficient positif.

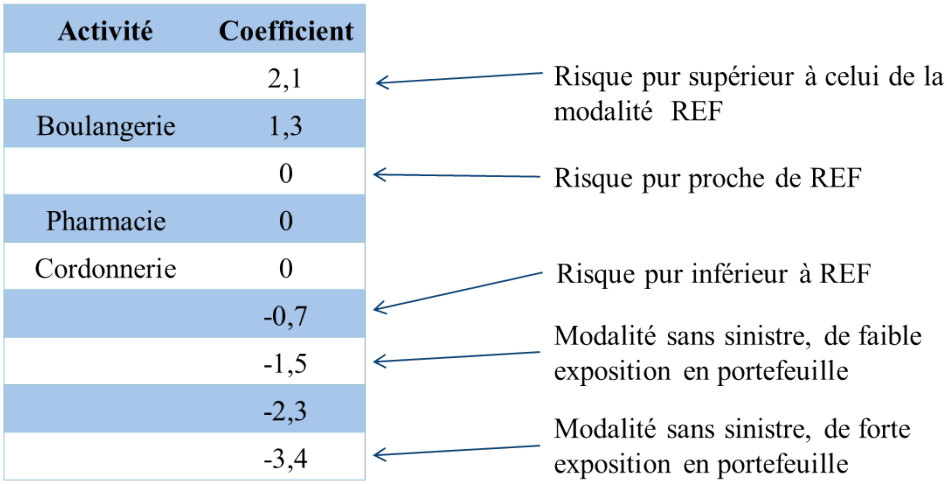


FIGURE 4.13 – Coefficient des différentes modalités de la variable Activité - GLM Poisson avec pénalisation LASSO

Notons que d'autres variables catégorielles nécessitent des regroupements mineurs, du fait de modalités à très faibles effectifs. Lors d'une validation croisée, ces modalités sont susceptibles d'être présentes dans la base d'apprentissage et non dans la base de test, générant une erreur. Nous nous contentons pour ces cas marginaux d'agréger les modalités d'effectif inférieur à 100 observations à la modalité la plus proche en termes de sinistralité grave.

4.2.4 Regroupement de modalités et données de test

Il est un élément important à noter pour l'évaluation des performances des modèles : les données de la base de test ne doivent en aucun cas participer à l'ajustement du modèle qui détermine le regroupement des modalités. En effet, si l'on regroupe à partir d'un modèle ajusté sur l'ensemble des données, les regroupements contiennent de l'information issue des données de la future base de test, et l'on obtient une performance artificiellement bonne. Pour mesurer cet effet, nous avons effectué des regroupements par arbre (méthode simple) sur l'ensemble des données, puis échantillonné une base de test (20%) sur laquelle nous avons mesuré la performance des modèles GLMs. L'AUC des modèles se trouve augmenté de près de 10 points en moyenne par rapport au cas ou les données de test sont isolées préalablement aux regroupements. Cet effet est probablement moindre dans le cas des regroupements par modèle linéaire - qui en ne captant que les effets purs (et non les effets de structure) dans ses regroupements récupère moins d'information des données - mais ne peut être ignoré pour autant. La figure 4.14 illustre la méthode à suivre pour s'assurer de la fiabilité des mesures réalisées sur la base de test.

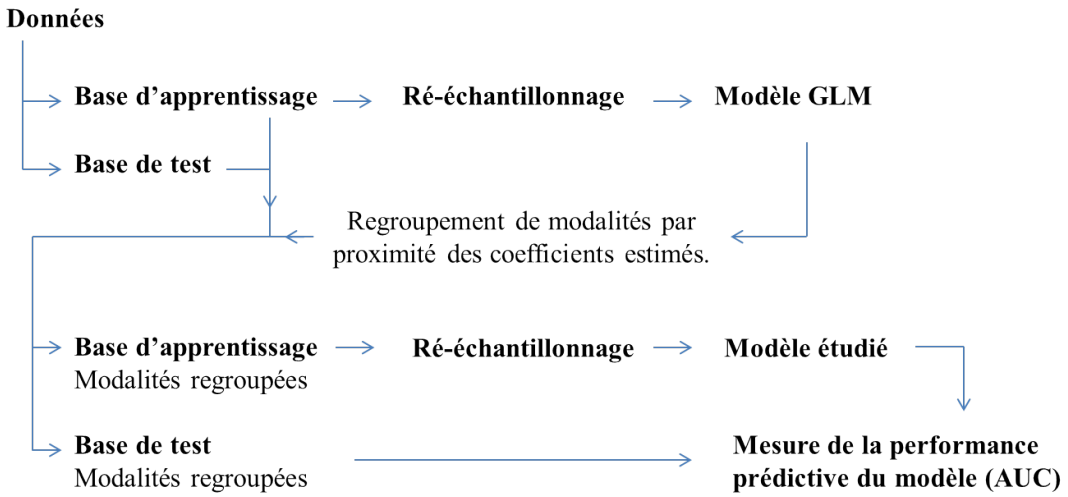


FIGURE 4.14 – Isolement des données de test préalablement au regroupement des modalités par GLM

Chapitre 5

Rééquilibrage des données

Avec 151 graves pour plus de 600 000 observations, nos données sont très déséquilibrées. La modélisation d'une classe aussi fortement minoritaire est problématique : l'estimation d'un modèle par maximum de vraisemblance n'est pas possible, et la performance prédictive des arbres est limitée. Trois types de méthodes sont envisageables afin de remédier aux difficultés de modélisation des classes minoritaires :

— Le rééquilibrage des données, par sous-échantillonnage de la classe majoritaire et/ou sur-échantillonnage de la classe minoritaire.
— Le boosting du classifieur, qui accroît progressivement la pondération des observations mal prédites.
— Les méthodes d'apprentissage sensibles aux coûts (*cost sensitive learning*), qui associent une matrice de coûts à la matrice de confusion.

Le rééquilibrage des données conjugue retrait et création d'individus : le sous-échantillonnage retire des individus de la classe majoritaire, tandis que le suréchantillonnage en ajoute, par bootstrap (tirage avec remplacement) au sein des observations minoritaires.

Le boosting repose sur la matrice de confusion : les observations mal prédites voient leur poids s'accroître, tandis que celui des observations bien prédites décroît légèrement. Le modèle est réajusté avec cette nouvelle pondération ; la nouvelle matrice de confusion est calculée et les poids sont à nouveau corrigés. Ce processus se poursuit tant que la repondération permet d'améliorer le modèle. Le choix d'un seuil est nécessaire, dans la mesure où à chaque itération une matrice de confusion est calculée : ce seuil constitue a priori un paramètre important du modèle. Dans ses premières itérations, le boosting a un effet proche de celui du suréchantillonnage : il accroît le poids des observations de la classe minoritaire. Or il est équivalent pour la maximisation de vraisemblance de tripler le poids d'une observation ou de générer deux nouvelles observations identiques à l'originale. L'effet du boosting sur le rééquilibrage des classes après les premières itérations est complexe, et dépend du seuil de la matrice de confusion : plus ce seuil est élevé, plus la classe minoritaire sera surpondérée par rapport à la classe majoritaire (les individus minoritaires étant plus fréquemment mal classés).

Les méthodes d'apprentissage sensibles au coût sont relativement difficiles à mettre en oeuvre dans notre cadre : nous y consacrons la partie 5.5.

Les méthodes de rééchantillonnage des données ont fait l'objet d'un certain nombre de travaux. Japkowicz (2000, [12]) considère plusieurs approches sur des données artificielles à une dimension. Elle procède à un suréchantillonnage aléatoire de la classe minoritaire, et à un suréchantillonnage concentré dans lequel seules les observations proches de la frontière entre les classes minoritaires et majoritaires sont suréchantillonnées. Elle procède également à un sous-échantillonnage aléatoire de la classe majoritaire, et à un sous-échantillonnage concentré dans lequel les observations retirées sont celles les plus éloignées de la frontière. Japkowicz observe que ces deux approches sont efficaces, et que les méthodes d'échantillonnage avancées ne sont pas significativement meilleures. En outre ces méthodes avancées reposent sur la détermination d'une frontière, ce qui est simple sur des données à une dimension mais plus délicat dans le cas multi-dimensionnel.

Nous allons ici utiliser la méthode développée par Chawla et al (2002, [7]), une technique de suréchantillonnage synthétique de la classe minoritaire (SMOTE, *Synthetic Minority Oversampling Technique*) aujourd'hui fréquemment utilisée pour le rééquilibrage de données dans certains champs d'étude tels que la biologie. L'objectif du suréchantillonnage synthétique est d'éviter le surapprentissage lié à la simple réplication des observations. Chawla et al. mettent en évidence ce surapprentissage dans le cas d'un arbre de décision : à mesure que l'on réplique, on identifie des régions très spécifiques de l'espace autour des points de la classe minoritaire, qui se prêtent mal à la généralisation sur de nouvelles données.

5.1 Suréchantillonnage synthétique

L'algorithme SMOTE consiste à générer de nouveaux individus minoritaires, situés aléatoirement sur des segments entre individus de la classe minoritaire, dans l'espace des variables explicatives. Nous en donnons l'algorithme détaillé en Annexe (Algorithme 1). Considérons, figure 5.1, un exemple dans un espace à deux variables explicatives. Les carrés rouges correspondent aux observations minoritaires, les carrés bleus aux observations majoritaires. On identifie les k plus proches voisins (que nous appelerons également par leur abréviation anglaise, k-NN, pour *k-Nearest Neighbors*) d'un individu au sein des individus minoritaires (Figure 5.2, avec k=3). On choisit ensuite aléatoirement l'un des k voisins de l'individu et un nombre α entre 0 et 1, qui donne une position sur le segment formé par l'individu et le voisin choisi : cette position détermine la valeur des variables du nouvel individu minoritaire généré. Pour un taux de SMOTE de 100%, on crée un individu synthétique à partir de chaque individu original et de ses k plus proches voisins (Figure 5.3).

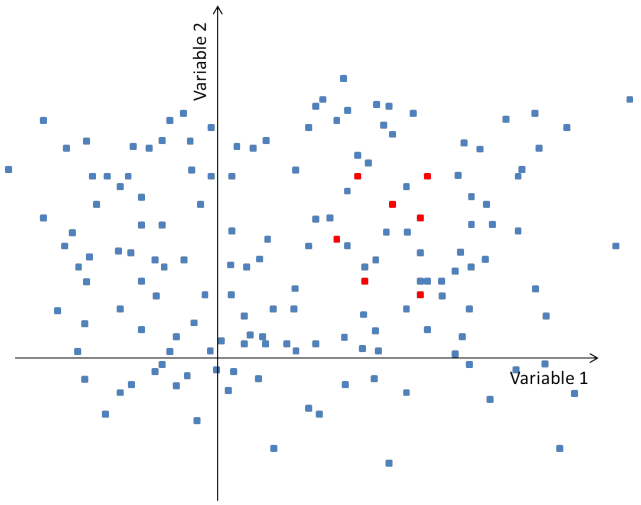


FIGURE 5.1 – Exemple de données dans un espace à deux variables explicatives

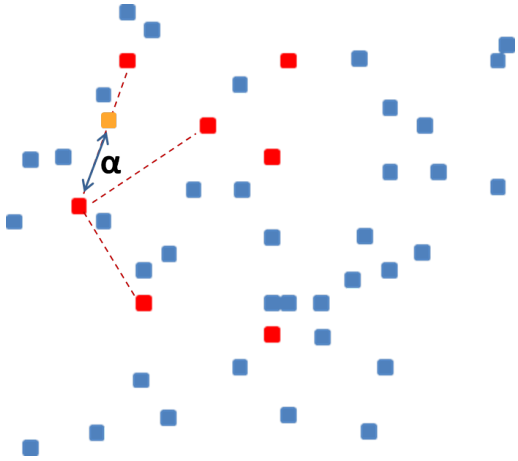


FIGURE 5.2 – Sélection des k plus proches voisins ($k = 3$) et génération d'un individu synthétique

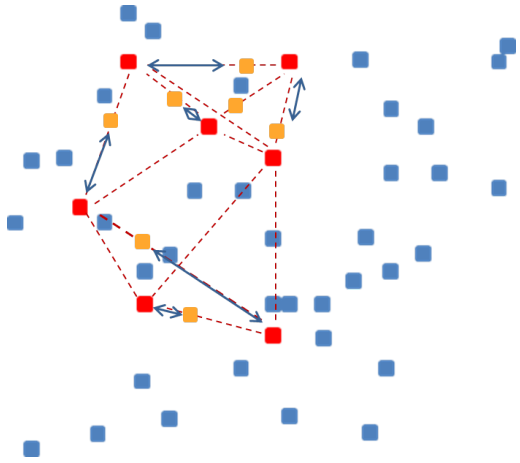


FIGURE 5.3 – Génération de n individus synthétiques

Nous utilisons la fonction SMOTE du package R *DMwR*. Cette fonction a trois principaux arguments :

— *perc.over* : le nombre d'individus synthétiques générés pour chaque individu de la classe minoritaire (détermine le taux de suréchantillonnage).

— *perc.under* : le nombre d'individus de la classe majoritaire conservés par individu synthétique créé (détermine le taux de sous-échantillonnage).

— k : le nombre de plus proches voisins utilisés pour la création des individus synthétiques.

Il est plus lisible de discuter les paramètres d'effectif final et de part de classe minoritaire que les taux de sur et sous-échantillonnage. Explicitons la correspondance entre ces deux couples de paramètres. On note :

— o : nombre de suréchantillonnage

— u : nombre de sous-échantillonnage

— n : effectif original de la classe minoritaire (151 graves dans notre cas)
— x : part de la classe minoritaire dans l'échantillon souhaité
— N : nombre total d'individus souhaités dans la base obtenue

Pour un individu de la classe minoritaire, on génère $o + 1$ individus minoritaires synthétiques et on sélectionne $u \cdot o$ individus de la classe majoritaire. On obtient donc une part de la classe minoritaire

$$x = \frac{o+1}{o+1+u \cdot o}$$

Et un nombre total d'observations $N = n(o + 1 + u \cdot o)$. En résolvant ce système on obtient

$$o = \frac{Nx}{n} - 1 \quad \text{et} \quad u = \frac{1-x}{x - \frac{n}{N}}$$

et on paramètre l'algorithme en fonction de x et N. Notons que la fonction SMOTE procède aux arrondis nécessaires et qu'il n'est en général pas possible d'atteindre sa cible exacte pour x et N à la fois.

Le suréchantillonnage synthétique requiert une normalisation des variables. Les variables quantitatives sont exprimées entre 0 et 1, afin d'éviter que les variables exprimées sur une grande échelle aient un poids plus grand que les autres dans le calcul des distances entre individus. Le traitement des variables qualitatives n'est précisé dans aucune documentation que nous ayons trouvée, aussi avons-nous examiné le code source du package *DMwR*. Les variables qualitatives sont traitées de la façons suivante :

— Pour les k plus proches voisins (k-NN), la variable est recodée par rapport à la modalité prise par l'individu dont on cherche les voisins : les individus partageant cette modalité prennent la valeur 1, les autres la valeur 0.
— Une fois que les k plus proches voisins ont été identifiés et l'un d'eux choisi aléatoirement, la modalité prise par l'individu synthétique est choisie aléatoirement parmi les deux modalités - celle de l'individu et celle du voisin choisi.

Chawla et al (2002, [7]) ont développé une seconde méthode plus générale, nommée *Synthetic Minority Over-sampling TEchnique-Nominal Continuous* (SMOTE-NC), mais qui semble n'être pas implémentée sous R.

Notons que l'on ne peut se contenter de binariser en $m - 1$ fonctions indicatrices une variable catégorielle à m modalités : cela reviendrait à fréquemment doubler le poids de la variable dans le k-NN. Considérons deux individus prenant des modalités différentes, et différentes de la modalité de référence : elles présentent chacune une indicatrice égale à 1 (les autres étant égales à 0). Le calcul de la distance euclidienne pour les indicatrices de cette variable conduit donc à $(0-1)^2 + (1-0)^2 = 2$, ce qui est le double de la contribution maximale qu'une variable quantitative puisse apporter à la distance euclidienne.

5.2 Effet du suréchantillonnage sur les CART

En travaillant sur des modèles CART, Chawla et al [7] montrent que le suréchantillonnage par réplication ne permet que l'apprentissage de régions très spécifiques aux environs des individus observés et n'améliore pas significativement la capacité à prédire les observations minoritaires. Le suréchantillonnage synthétique permet en revanche l'identification de zones plus larges au sein de la région des observations minoritaires. Pour illustrer ce phénomène, considérons figures 5.4 et 5.5 le résultat d'un CART sous sa forme simplifiée, dans laquelle le vecteur des probabilités est transformé en un vecteur binaire à partir d'un seuil (nous nous trouvons donc sur un unique point de la courbe ROC). Nous identifions par des rectangles les régions de l'espace (un espace à 2 variables explicatives) au sein desquelles l'arbre prédit des observations positives ($\hat{Y} = 1$). La figure 5.4 illustre le résultat obtenu avec un suréchantillonnage par réplication tandis que la figure 5.5 illustre le résultat obtenu avec un suréchantillonnage synthétique. L'arbre construit sur les données de suréchantillonnage synthétique génère bien plus de vrais positifs sur des données de test, du fait qu'il a véritablement appris la région occupée par la classe minoritaire. Ceci n'est possible qu'au prix d'une augmentation des faux positifs, mais dont l'effet est marginal lorsqu'elle est rapportée au nombre total d'observations négatives : la performance du modèle augmente.

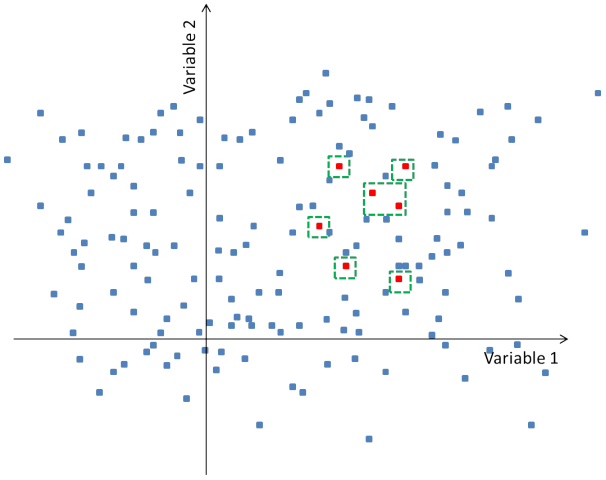


FIGURE 5.4 – Régions d'apprentissage du CART sur données brutes ou suréchantillonnées par réplication

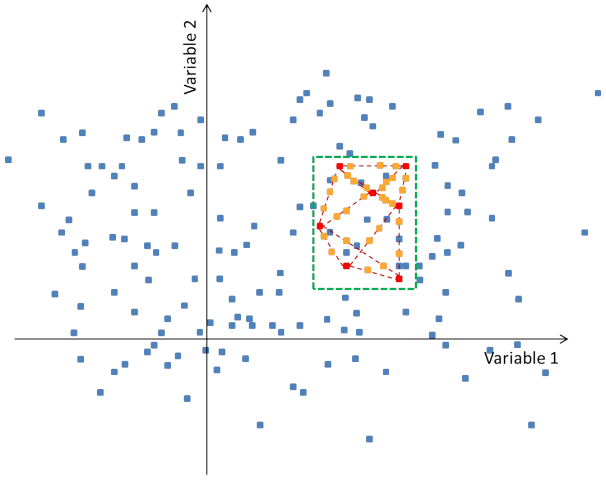


FIGURE 5.5 – Régions d'apprentissage du CART sur données suréchantillonnées par SMOTE

5.3 Effet du suréchantillonnage sur les modèles linéaires

Quel est l'effet du suréchantillonnage sur les performances d'un modèle linéaire ? La figure 5.6 rappelle la surface des prédictions d'un modèle logistique appliqué à deux variables explicatives. Étant soumis à une spécification linéaire, un GLM n'est pas touché par l'écueil qui consiste à ap-

prendre des régions très spécifiques de l'espace. Il n'en est pas plus performant pour autant : aucun modèle linéaire appliqué sur nos données originales ne parvient à apprendre les données. Les données minoritaires étant trop peu nombreuses, la maximisation de la vraisemblance conduit à prédire une surface plane, de probabilité prédite égale à celle globalement observée sur nos données (soit $> 0.03\%$). Le suréchantillonnage synthétique, en densifiant la région des observations minoritaires, permet d'aboutir à la situation présentée figure 5.6. Il en va de même du suréchantillonnage par réplication : le choix du mode de suréchantillonnage est a priori peu important pour les modèles linéaires. Par souci d'harmonie et de comparabilité des modèles, nous avons retenu le suréchantillonnage synthétique pour l'ensemble de notre approche. Le suréchantillonnage par réplication peut d'ailleurs être envisagé comme un suréchantillonnage synthétique "à 0 plus proches voisins". Plus l'on augmente le k du k-NN, plus l'on s'éloigne du cas réplicatif. Nous verrons que le SMOTE présente un effet potentiel indésirable sur le GLM lorsque k est élevé, mais d'impact très limité en pratique sur la performance.

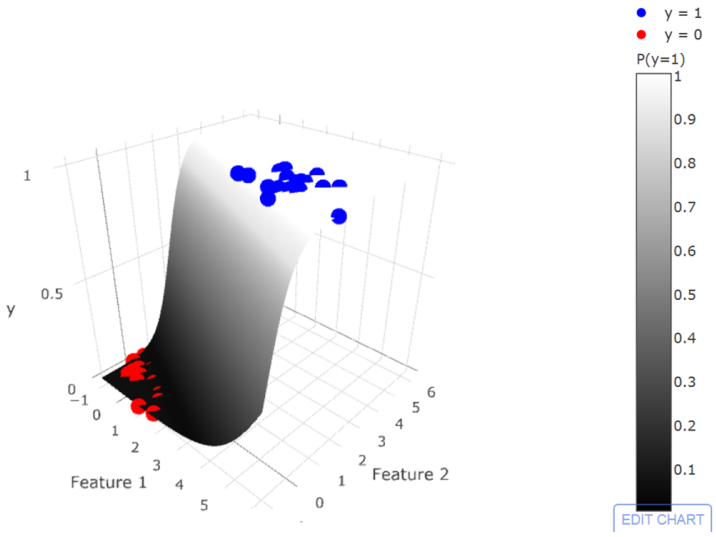


FIGURE 5.6 – Surface des prédictions d'une régression logistique appliquée à deux variables explicatives.
Source : F. Hartl, https ://florianhartl.com/logistic-regression-geometric-intuition.html

5.4 Optimisation du ré-échantillonnage synthétique

Le suréchantillonnage synthétique, par son mode de génération linéaire - un individu est créé sur le segment défini par l'un de ses k plus proches voisins et par lui-même - peut perturber des structures de données non-linéaires. L'effet de ces perturbations est fort sur les CART, qui sont sensibles aux structures non-linéaires, et faible sur les modèles linéaires. Cet effet naît du mode de sélection des k plus proches voisins, et dépend du paramètre k ainsi que de l'espace des variables explicatives choisi.

5.4.1 Effet du nombre de plus proches voisins

La littérature indique qu'il que le nombre k de plus proches voisins est un paramètre susceptible d'influencer la performance, mais il n'y a à notre connaissance que peu de précisions sur l'effet de ce paramètre, effet d'ailleurs fortement dépendant de la structure des données.

Illustrons les effets du k sur les prédictions des CART et des modèles linéaires par un exemple. Considérons l'effet du k sur la position des individus générés, dans un cas à deux variables explicatives présentant une interaction (Figure 5.7). Du fait de l'interaction négative des deux variables - elles ont isolément un effet positif sur la fréquence de grave, mais conjointement un effet faible ou nul - nous observons deux populations d'individus minoritaires, que nous notons A et B.

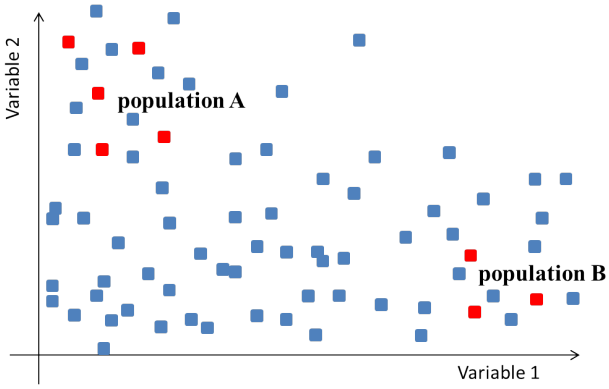


FIGURE 5.7 – Exemple de données sur deux variables explicatives présentant une interaction.

En prenant $k = 2$ (Figure 5.8), chaque individu trouve ses deux plus proches voisins dans sa population, et nous densifions chacun des deux groupes isolément. Notons que nous générons ici deux individus synthétiques par individu original. Avec $k = 3$ (Figure 5.9), les individus de la population B trouvent leur troisième plus proche voisin dans la population A, ce qui génère des graves dans l'espace intermédiaire et réduit la densité de la population B.

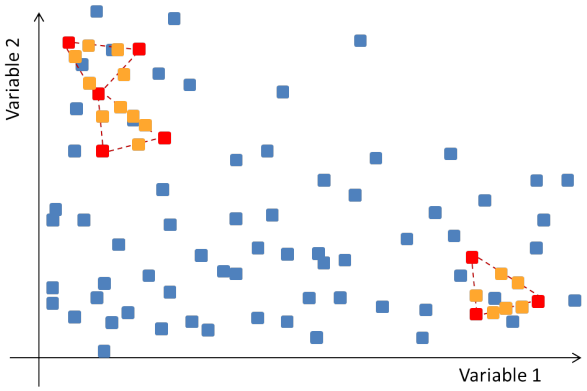


FIGURE 5.8 – Suréchantillonnage synthétique avec les deux plus proches voisins.

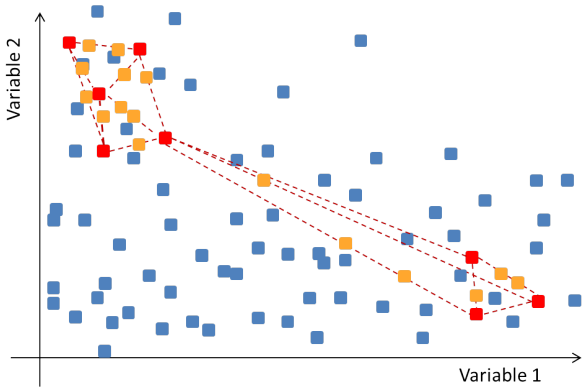


FIGURE 5.9 – Suréchantillonnage synthétique avec les trois plus proches voisins.

Le choix du k conduit ici à deux structures différentes des données d'apprentissage. Quel est son effet en termes de prédiction des modèles ?

Le choix de $k = 3$ conduit le CART à identifier une zone de fréquence moyenne artificielle (Figure 5.11), en distordant une partie de la densité de la population B. Les prédictions du CART sont donc sensiblement affectées : les individus négatifs situés dans la zone de prédiction à 70% se voient

attribuer une probabilité prédite supérieure à celle des individus positifs de la zone de prédiction à 60%, dégradant la qualité de la prédiction par rapport à $k = 2$, qui les prédit au même niveau de risque (Figure 5.10). La présence d'une région de probabilité prédite de 20% peut, elle aussi, être délétère : toute nouvelle observation située en frontière des zones de prédiction des populations A et B se voit prédire un risque moindre que celui de la région intermédiaire, alors qu'elle est selon toute vraisemblance plus risquée. Ainsi, sous réserve que la population de test échantillonnée soit bien représentative de la structure en deux populations des données d'apprentissage, la performance du modèle est plus faible avec $k = 3$.

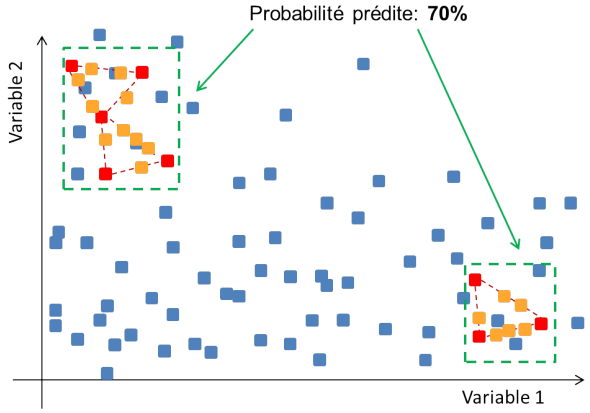


FIGURE 5.10 – Effet du suréchantillonnage (avec k=2) sur les prédictions du CART.

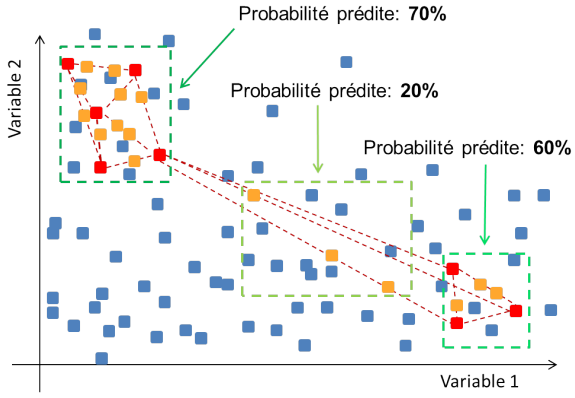


FIGURE 5.11 – Effet du suréchantillonnage (avec k=3) sur les prédictions du CART.

Les figures 5.12 et 5.13 sont analogues de la figure 5.6 et représentent les probabilités prédites d'un modèle linéaire (logit, ou Poisson normalisé). La clarté de la nappe bleue croît avec le niveau de probabilité prédit. Dans le cas linéaire, la moindre densité de la population B issue de $k = 3$ (figure 5.13) conduit à un plus faible coefficient associé à la variable 1 au sein du prédicteur linéaire, et par conséquent à un "aplatissement" des droites d'iso-prédiction. Ceci a pour effet de placer quelques observations de la population B en marge du plateau de haute probabilité - ce qui abaisse leur probabilité prédite - et de considérer qu'une part de la population B est de moindre risque que celle de la région artificielle. Il peut en résulter une réduction de la performance du modèle, et un cas pour lequel le suréchantillonnage réplicatif est plus souhaitable. L'effet du k demeure cependant moins fort que celui observé sur les CART, du fait de la moindre sensibilité des modèles linéaires à la structure du nuage de points, et nous n'observerons en pratique aucun effet sensible.

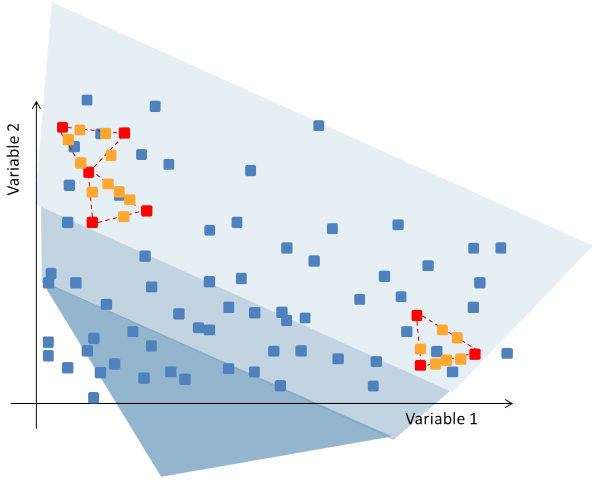


FIGURE 5.12 – Effet du suréchantillonnage (avec k=2) sur les prédictions du GLM.

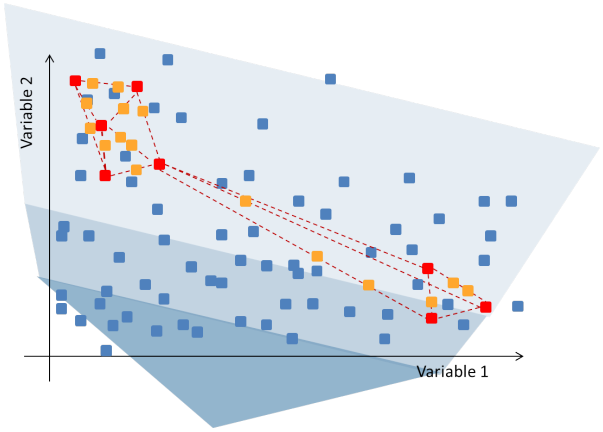


FIGURE 5.13 – Effet du suréchantillonnage (avec k=3) sur les prédictions du GLM

5.4.2 Effet de l'espace sur le suréchantillonnage synthétique

L'espace choisi pour le suréchantillonnage synthétique n'est pas anodin. Avec 150 variables explicatives, nos données sont frappées du "fléau de la dimension" (Bellman, 1961) : lorsque la dimension augmente, le volume de l'espace croît rapidement et les données deviennent isolées. Il n'est pas difficile d'imaginer la solitude de nos 151 incendies graves dans un espace à 150 dimensions, et réduire le nombre des variables en excluant celles qui ne sont pas explicatives apparaît être une piste intéressante. Illustrons l'effet des variables non-explicatives à l'aide de la figure 5.14. La variable 2 n'a pas d'effet sur le risque de grave : la proportion de graves observée est sensiblement la même sur l'ensemble de ses valeurs. La variable 1 a en revanche un effet significatif sur le risque de graves, et permet d'isoler une population à haut risque ainsi qu'une population à risque moyen. Considérons un suréchantillonnage synthétique à deux plus proches voisins et avec un taux de 100% (un nouvel individu par individu original). Le k-NN (figure 5.15) conduit à générer des individus synthétiques entre les deux populations, ce qui pose deux problèmes. D'une part, nous générons des graves dans une région n'en contenant pas, ce qui réduit la performance (nous l'avons vu dans le paragraphe précédent). D'autre part, nous modifions la distribution des graves selon la variable 2 - puisqu'il est généré davantage d'individus dans la zone haute du graphe - ce qui la rend artificiellement explicative. Supprimer la variable 2 avant de procéder au suréchantillonnage permettrait d'éviter ces problèmes, ce que montre la figure 5.16.

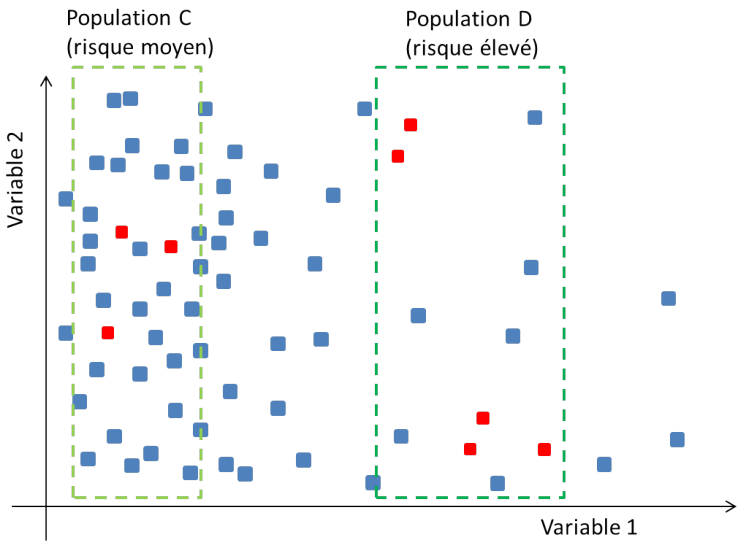


FIGURE 5.14 – Exemple de données sur deux variables, l'une explicative des graves (variable 1), l'autre non.

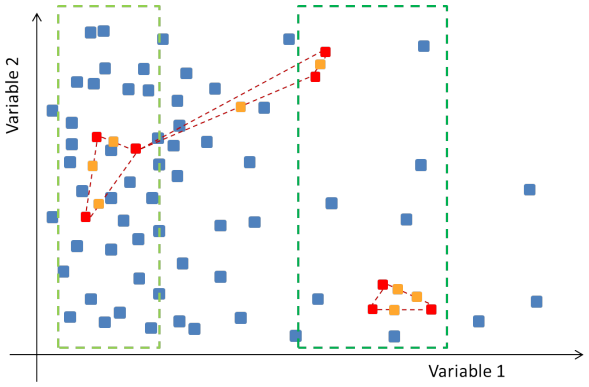


FIGURE 5.15 – Suréchantillonnage sur l'espace des deux variables.

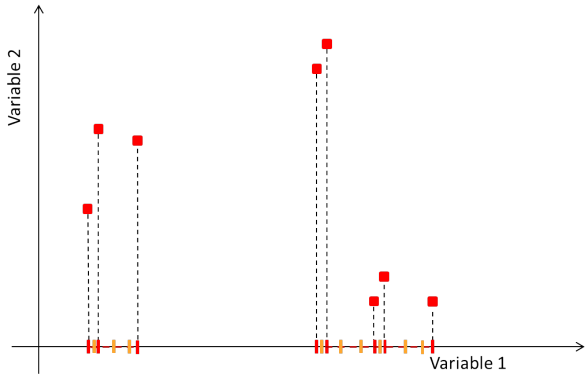


FIGURE 5.16 – Suréchantillonnage sur la variable 1 uniquement.

Il peut donc être opportun de procéder en deux temps lors des modélisations : après un premier suréchantillonnage sur l'ensemble des variables et un premier ajustement du modèle, ne conserver que les variables retenues dans le modèle et procéder à un nouveau suréchantillonnage suivi d'un nouvel ajustement du modèle. Il demeure certes possible qu'une variable peu importante soit rendue plus explicative par le SMOTE et soit donc conservée dans le second SMOTE. Ce phénomène est difficile à mesurer et à contrôler, mais nous pouvons raisonnablement penser qu'il n'a qu'un effet marginal.

5.5 Prise en compte des enjeux de prime commerciale : cost-sensitive learning

L'un des trois objectifs de notre modélisation est d'aboutir à une prime pure pour sinistres graves, en associant un coût moyen des graves au modèle de fréquence. Cette prime pure s'intègre au dispositif existant :

— Un modèle de prime pure attritionnelle, additionnée d'un montant peu discriminant - pour couvrir la charge des sinistres graves sur le portefeuille - que la prime pure grave vient remplacer.
— un ensemble d'éléments économiques (coûts de gestion unitaire), commerciaux (développement du portefeuille, gain de parts de marché, politique de multi-équipement avec des produits d'entrée) et stratégiques (cohérence avec la politique bancaire du groupe Crédit Agricole) conduisant à la prime commerciale.

Il peut être utile d'anticiper les enjeux de la prime commerciale dans le modèle de prime pure, en intégrant des éléments micro-économiques au sein même du processus d'apprentissage. Ce type d'approche est connue sous le terme général de *cost sensitive learning*.

5.5.1 Performance micro-économique d'un modèle de prime pure

Considérons un modèle de prime pure dérivé de notre modèle de fréquence, et constitué de N groupes tarifaires, de niveau de risque croissant. Une première possibilité est de tarifer chaque groupe à sa prime pure observée (fréquence de graves sur le groupe x coût moyen d'un grave). Pour autant, il est possible que certains groupes génèrent une marge - hors sinistralité grave - plus forte que d'autres, du fait d'un ratio sinistres/cotisations attritionnels plus faible. Entre deux groupes de sinistralité grave proche, il peut être préférable d'augmenter la cotisation du groupe le moins profitable davantage que celle du groupe le plus profitable, que l'on souhaite conserver en portefeuille, sous réserve de la prise en compte d'autres éléments - le produit d'assurance multi-risque peut constituer un produit d'entrée pour le groupe peu profitable, qui s'avère profitable sur les produits auxquels il souscrit par la suite.

L'aspect dynamique d'un portefeuille d'assurance rend l'évaluation de la performance difficile : s'il est aisé d'estimer la marge unitaire au sein d'un groupe, il l'est moins de mesurer l'effet d'une variation de cotisation sur le volume du portefeuille, sur son érosion ou son développement du fait de la modification des taux de souscription et/ou de résiliation. Une approche pertinente nécessite de connaître la fonction de demande de chaque groupe tarifaire : un groupe présentant une élasticité-prix élevée (en valeur absolue) réagit fortement aux variations de cotisation.

L'évaluation de la performance économique est plus aisée dans le cas d'une stratégie d'exclusion des hauts risques du portefeuille, qui consiste à filtrer la fraction des assurés de plus haut risque, et constitue l'un des objectifs de notre travail. Cette approche s'affranchit de variations indéterminées de prix et de volumes, et l'on peut calculer simplement la marge espérée sur le portefeuille. Considérons un seuil d'exclusion s : tous les assurés de probabilité de survenance supérieure à ce seuil

sont éliminés du portefeuille. On calcule alors la marge historique réalisée sur le portefeuille des assurés conservés, comme la marge - hors sinistres graves - réalisée sur ces assurés moins la charge (Nombre de graves × Coût moyen) des graves.

La figure 5.17 montre la possibilité d'optimiser la marge globale du portefeuille si le modèle est suffisamment performant. Les hauts risques sont concentrés dans la région 1, et la charge des graves y est supérieure à la marge réalisée : on augmente la marge globale en élargissant la zone d'exclusion. En région 2 la charge des graves est moins importante que les marges réalisées, et l'érosion du portefeuille n'est plus souhaitable. Il existe donc un seuil d'exclusion optimal, représenté par le trait vertical orange.

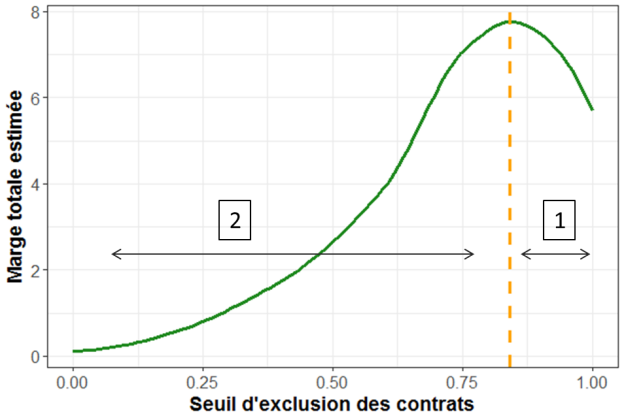


FIGURE 5.17 – Optimisation de la marge du portefeuille par le seuil d'exclusion des contrats

5.5.2 Pondération des graves pour l'amélioration de la performance économique

Pouvons-nous influencer le processus d'apprentissage des graves de façon à moins pénaliser les contrats les plus profitables, pour les conserver et augmenter la marge du portefeuille ? Nous envisageons ici une méthode de pondération différentielle en fonction de la profitabilité des contrats. Nous pouvons en première approche considérer que la profitabilité d'un contrat augmente avec le montant de sa cotisation, en supposant le ratio Sinistres/Cotisations indépendant de la cotisation et la marge brute réalisée proportionnelle à la cotisation.

Le modèle de tarification actuel ne comporte pas de réelle prime pure pour les graves, et son montant dépend principalement de la prime pure attritionnelle. Les variables qui reflètent la taille de l'entreprise (surface, valeur de contenu, capitaux garantis) ont un effet positif sur la prime pure attritionnelle, et par conséquent sur la cotisation. Les contrats à faibles cotisation et à forte cotisation tendent donc à former des régions distinctes dans l'espace de ces variables explicatives (ou à minima un gradient de cotisation). Ainsi, en modifiant la pondération des individus sinistrés, il est possible de moins pénaliser les contrats à forte cotisation.

Examinons en quoi cette méthode peut être utile dans une stratégie d'exclusion des hauts risques. Considérons un cas simple, en supposant qu'il existe deux principaux profils de risque : certains assurés de grande taille, présentant un risque grave plus élevé du fait de leur forte activité, et certains professionnels de petite taille présentant un risque grave du fait d'un manque de contrôle des installations. Nous représentons ces profils de risque dans le plan défini par les variables Surface et Capital garanti, figure 5.18.

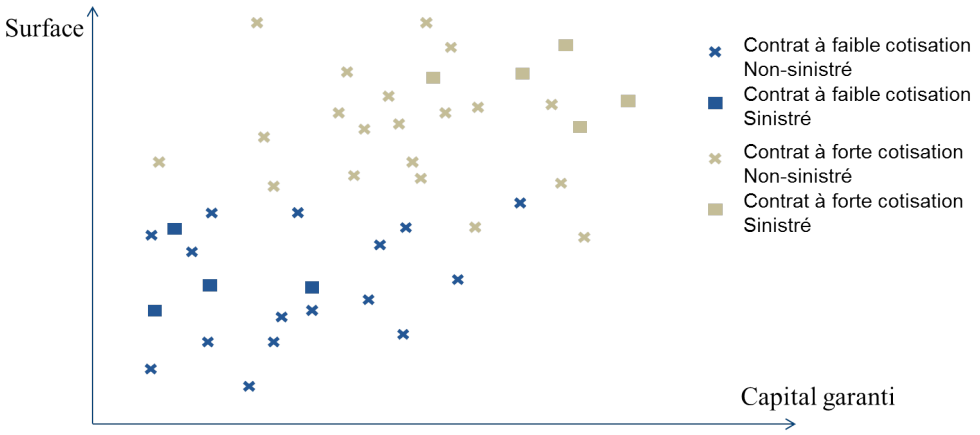


FIGURE 5.18 – Illustration : données avec deux profils de risque, correspondant à deux niveaux de cotisation différents

L'ajustement d'un modèle sur ces données conduit naturellement à deux régions de risque élevé, que nous représentons dans le plan Surface x Capital garanti (figure 5.19). Nous y repérons deux assurés (A et B) de même prime pure grave estimée. Ces deux assurés sont situés au même niveau dans le vecteur ordonné de prédiction des risques, et seront exclus pour un même seuil.

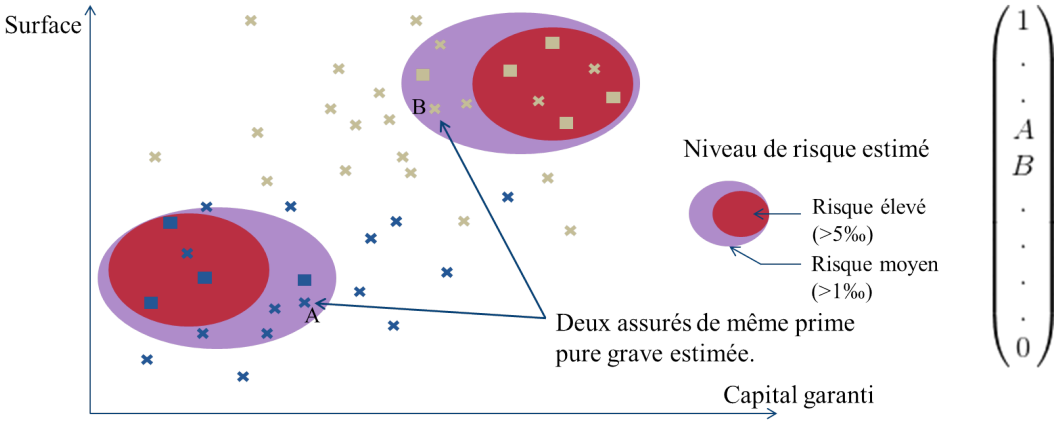


FIGURE 5.19 – Estimation de la prime pure grave sur données équipondérées.

Nous proposons à présent de modifier la pondération des graves : en conservant le même poids cumulé des graves, nous multiplions le poids des assurés sinistrés par un facteur inversement proportionnel à leur cotisation. Nous augmentons ainsi le poids des graves de faible cotisation, et réduisons celui des graves de forte cotisation. Ceci a pour effet de modifier l'apprentissage des graves, ce qui s'observe figure 5.20 par l'accroissement de la zone de risque au sein des assurés de faible cotisation et par sa réduction pour les assurés de forte cotisation. Le contrat A, de faible cotisation, est désormais placé à un niveau supérieur à B dans le vecteur ordonné des probabilités prédites, et sera donc exclu en premier.

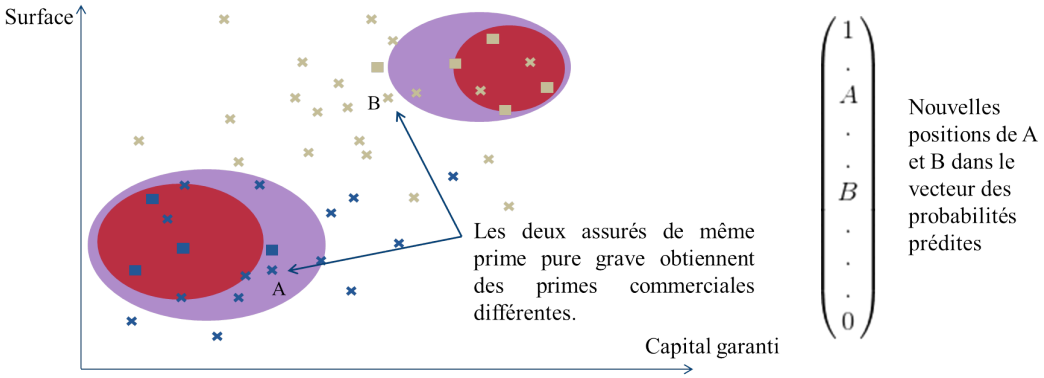


FIGURE 5.20 – Effet de la re-pondération des graves sur le risque estimé - Obtention d'une prime commerciale

Ce nouveau vecteur de probabilités prédites permet une optimisation plus avancée de la marge globale du portefeuille par l'exclusion des risques. On obtient (figure 5.21, courbe jaune) une marge plus élevée à l'optimum, du fait que l'on exclut en région 1 des contrats à plus faible cotisation moyenne qu'avec le modèle équipondéré (courbe verte).

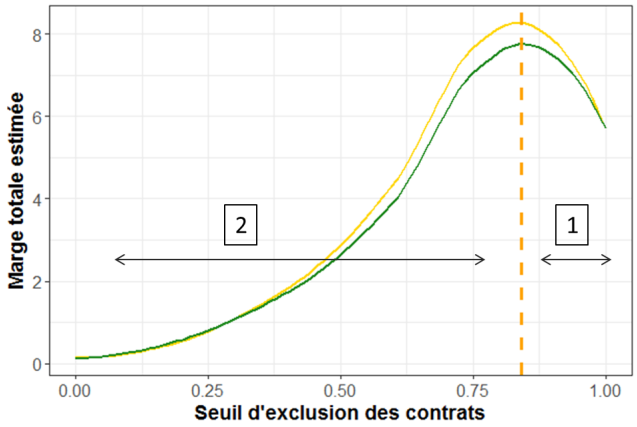


FIGURE 5.21 – Optimisation de la marge du portefeuille par le seuil d'exclusion des contrats

Nous n'avons pas expérimenté cette méthode du fait d'un manque d'éléments immédiatement disponibles sur la profitabilité des contrats, l'hypothèse d'une profitabilité proportionnelle à la cotisation étant trop simpliste sachant la diversité du portefeuille. La mise en oeuvre et l'opportunité de cette pratique demeurent donc à vérifier, et sont fortement dépendantes de la structure des profils de risque.

Notons enfin que cette approche semble moins pertinente dans le cadre d'un modèle général de tarification. Les contrats à faible cotisation sont plus sensibles à l'augmentation de leur cotisation : l'augmentation de la prime pure d'un montant absolu x conduit à une plus grande augmentation relative du prix pour un contrat à faible cotisation que pour un contrat à forte cotisation. À élasticité-prix donnée, la demande relative des contrats à faible cotisation est davantage impactée, et une exclusion sélective s'opère avant même que la prime pure ne soit orientée vers une prime commerciale.

Chapitre 6

Sélection et optimisation des modèles

6.1 Sélection des modèles

Pour l'ensemble des modèles mis en oeuvre, les variables quantitatives ont été normalisées (i.e. rapportées entre 0 et 1) : ceci est essentiel pour l'algorithme des plus proches voisins du SMOTE et pour une pénalisation LASSO équilibrée.

Comparons les performances des modèles envisagés, et leur sensibilité au ré-échantillonnage synthétique - taux de graves et nombre d'observations dans la base d'apprentissage. Nous n'étudions en définitive que cinq des six modèles présentés : le modèle de régression logistique corrigé de l'exposition n'a convergé sur aucune des bases d'apprentissage évaluées, et sera donc d'emblée exclu. La figure 6.1 résume cette première étude.

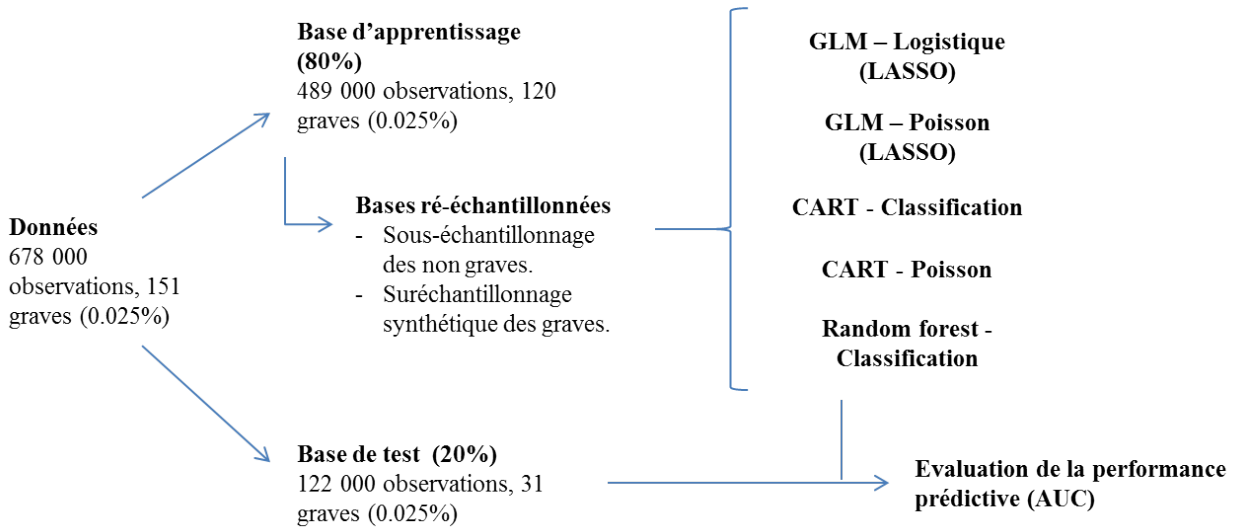


FIGURE 6.1 – Calcul des performances de chaque modèle sur différents rééchantillonnages de la base d'apprentissage.

Nous considérons pour chaque modèle 14 rééchantillonnages différents, avec des taux de graves de 0.1 à 50% et un nombre d'observations dans la base d'apprentissage compris entre 50 000 et 500

000. Nous ne générons pas de base d'effectif supérieur à 500 000 pour des raisons de temps de calcul et de mémoire disponible. Les AUCs obtenus sont présentés figure 6.2. Les AUCs moyens par taux ou par effectif sont calculés en marge, et la meilleure performance moyenne indiquée en orange. Le premier rééchantillonnage (0.1% de graves et 50 000 observations) n'est pas calculé car il nécessite de retirer des graves de la base d'apprentissage, ce qui est à l'opposé de nos objectifs. Les variables catégorielles à grand nombre de modalités sont regroupées par arbre pour cette première étude. Nous avons vu que cette méthode n'était pas la plus souhaitable, et procéderons à des regroupements par modèle linéaire par la suite.

		Taux de graves					
GLM - Logistique		**0,1%**	**1,0%**	**10,0%**	**30,0%**	**50,0%**	**AUC moyen**
Effectif	50 000		47,1%	55,0%	62,5%	54,1%	54,7%
	150 000	50,2%	62,9%	50,6%	60,0%	63,9%	57,5%
	500 000	56,1%	53,3%	59,1%	53,0%	Problème de convergence	-
AUC moyen		53,1%	54,5%	54,9%	58,5%	-	**56,5%**

		Taux de graves					
GLM - Poisson		**0,1%**	**1,0%**	**10,0%**	**30,0%**	**50,0%**	**AUC moyen**
Effectif	50 000		47,9%	60,1%	53,0%	58,4%	54,8%
	150 000	51,2%	64,5%	46,1%	59,4%	66,0%	57,4%
	500 000	53,1%	49,1%	56,5%	50,9%	71,8%	56,3%
AUC moyen		52,1%	53,8%	54,2%	54,4%	65,4%	**56,2%**

		Taux de graves					
CART - Classification		**0,1%**	**1,0%**	**10,0%**	**30,0%**	**50,0%**	**AUC moyen**
Effectif	50 000		65,8%	78,6%	75,6%	71,7%	72,9%
	150 000	50,0%	72,8%	76,2%	63,9%	69,1%	66,4%
	500 000	71,2%	70,4%	63,5%	69,5%	62,4%	67,4%
AUC moyen		60,6%	69,7%	72,7%	69,6%	67,8%	**68,5%**

		Taux de graves					
CART - Poisson		**0,1%**	**1,0%**	**10,0%**	**30,0%**	**50,0%**	**AUC moyen**
Effectif	50 000		67,6%	66,1%	68,2%	64,8%	66,7%
	150 000	57,3%	70,8%	67,9%	70,0%	66,1%	66,4%
	500 000	61,0%	72,4%	69,3%	66,7%	67,2%	67,3%
AUC moyen		59,2%	70,3%	67,8%	68,3%	66,0%	**66,7%**

		Taux de graves					
Random Forest - Classification		**0,1%**	**1,0%**	**10,0%**	**30,0%**	**50,0%**	**AUC moyen**
Effectif	50 000		83,3%	81,6%	79,1%	79,1%	80,7%
	150 000	80,8%	80,1%	81,7%	79,9%	76,6%	79,8%
	500 000	83,2%	78,4%	80,2%	79,1%	76,7%	79,5%
AUC moyen		82,0%	80,6%	81,2%	79,3%	77,5%	**80,0%**

FIGURE 6.2 – Performances mesurées sur la base de test - par modèle et par rééchantillonnage.

Les GLMs sont globalement peu performants, avec des AUCs moyens inférieurs à 0.57. Certains ajustements mésapprennent même les données : avec un AUC inférieur à 0.5, ils sont moins performants qu'un classifieur aléatoire. La régression logistique présente des temps de calcul importants, du fait de difficultés de convergence, et le dernier ajustement n'est pas réalisable du fait du très grand nombre d'itérations nécessaires pour obtenir la convergence (sous réserve qu'elle existe). Nous privilégierons par conséquent la régression de Poisson.

La performance du modèle de Poisson n'évolue pas de façon monotone avec le taux ou l'effectif du rééchantillonnage. On peut pour autant identifier des tendances. Le taux de graves le plus élevé (50%) génère un AUC moyen nettement supérieur aux autres. La performance est moins sensible à l'effectif : le meilleur AUC moyen est obtenu pour 100 000 observations, mais avec de faibles écarts. Le meilleur modèle (71.8%), obtenu avec 500 000 observations, démontre la puissance du rééchantillonnage pour améliorer la capacité prédictive d'un modèle linéaire.

Les CART performent mieux que les modèles linéaires, avec des AUCs moyens proches de %70, Le CART de classification est plus performant que la CART de régression de Poisson, tant pour l'AUC moyen que pour le meilleur ajustement obtenu. Les deux modèles sont moins sensibles au taux de graves que les modèles linéaires, mais présentent comme eux des variations de performance non-monotones. Un taux de 10% et un effectif de 50 000 sont les meilleurs paramètres pour le CART de classification.

La forêt aléatoire surpasse ses concurrents : elle offre un AUC moyen de 80%, avec un maximum à 83%. Ses performances sont maximales pour un taux de 0.1 ou 1%, et un effectif de 50 000.

6.2 Stabilité des performances mesurées sur la base de test

Pour autant, quelle confiance peut-on accorder à la performance mesurée sur une base de test avec seulement 31 sinistres graves ? Nous retrouvons ici le problème de la classe minoritaire, sous une autre forme. Etant donné le faible nombre d'observations positives, la répartition aléatoire des 151 graves dans les bases d'apprentissage (120) et de test (31) peut influencer la performance mesurée. Il semble par conséquent hasardeux de tirer des conclusions à partir d'un seul échantillonnage apprentissage/test.

La loi des grands nombres ne s'appliquant pas sur notre échantillon, nous ne pouvons exclure l'occurrence de tels phénomènes : seuls plusieurs échantillonnages aléatoires pour générer des lots de bases - d'apprentissage et de test - différents permettent d'obtenir une mesure fiable du résultat.

Illustrons ce propos par un cas dans lequel il existerait deux profils de risque. Les sinistres graves sont alors situés dans deux zones distinctes de l'espace des variables explicatives. Il est possible que l'échantillonnage distribue majoritairement les sinistres d'une zone vers la base d'apprentissage et ceux de l'autre zone vers la base de test. La performance du modèle est affaiblie par le moindre apprentissage d'une des deux zones, et par la génération d'individus entre voisins éloignés (cf Partie 5.4.1), et la performance mesurée est encore affaiblie du fait que les observations de test sont majoritairement situées dans la zone peu apprise. La faible performance obtenue n'a pour autant rien d'artificiel : il est possible que si nous n'avions connu que 120 sinistres à ce jour, ils soient ceux de la base d'apprentissage, et que les 31 à venir soient ceux de la base de test. Cette sous-performance témoigne du fait que, n'ayant que peu de données, il est possible que nous n'ayons encore jamais rencontré certains types de graves, qui surviendront avec le temps et l'accroissement

du portefeuille. Les sinistres connus à ce jour peuvent être majoritairement issus de défaillances électriques, tandis que ceux à venir peuvent être causés par des feux de poubelle, correspondant à un profil d'assuré différent de ceux identifiés. Nous pouvons de même observer des surperformances lorsque les profils de risque rare sont distribués dans la base d'apprentissage et que la base de test ne comporte ainsi que des graves faciles à prédire.

Un échantillonnage peut être identifié et reproduit à partir de la graine du processus aléatoire qui lui a donné naissance : une graine = un échantillonnage.

6.3 Sensibilité des modèles au suréchantillonnage synthétique

Nous avons vu que le nombre de plus proches voisins utilisé par l'algorithme de suréchantillonnage synthétique était susceptible d'impacter la performance du modèle. Afin de mesurer la sensibilité des modèles linéaires et des arbres, nous considérons 3 valeurs du paramètre k : 2, 10 et 20. Le taux de graves est fixé à 50% et l'effectif à 100 000, paramètres ayant conduit aux meilleures performances moyennes dans le modèle précédent. Nous ajustons sur chaque base rééchantillonnée une régression de Poisson et une random forest : la random forest étant plus stable que les CART, elle constitue un bon moyen de tester la sensibilité des arbres au paramètre k puisque ses performances sont moins touchées par le bruit.

Les variables à nombreuses modalités sont ici transformées en indicatrices, aussi cette étude est-elle l'occasion de mesurer l'impact délétère de cette méthode sur les performances de la random forest. Dans la mesure où nous ne chercons pas ici à optimiser la random forest mais à connaître sa sensibilité, et où nous approfondirons l'effet du k par la suite, les faibles performances de la random forest ne sont pas un problème.

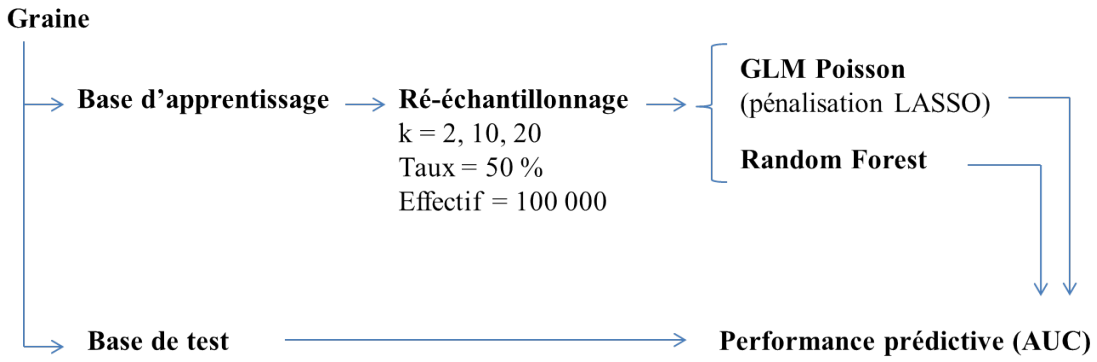


FIGURE 6.3 – Calcul des performances de chaque modèle sur différents rééchantillonnages de la base d'apprentissage.

Les performances du GLM apparaissent peu sensibles au paramètrage du k-NN (figure 6.4), mais varient fortement avec la graine : on observe jusqu'à près de 20 points d'écart sur l'AUC moyen entre deux graines. La random forest est, elle, sensible à la graine et au k. Les valeurs moyennes suggèrent que les valeurs élevées de k (10, 20) conduisent à un meilleur apprentissage.

Le résultat utile de cette analyse est qu'il n'est pas nécessaire d'optimiser le paramètre du k-NN pour la régression de Poisson : on pourra se contenter de prendre une valeur fixe, prise entre 2 et 10. La conclusion diffère pour les CART : puisqu'ils sont plus instables que la random forest et que cette dernière apparaît sensible au k-NN et à la graine, on ne peut écarter l'optimisation du k ni se fier au résultat d'un seul échantillonnage.

graine	k	AUC GLM Poisson	AUC Random Forest
98	2	56%	63%
	10	55%	65%
	20	54%	71%
250	2	69%	61%
	10	68%	64%
	20	67%	60%
484	2	73%	59%
	10	74%	66%
	20	73%	69%
544	2	61%	68%
	10	60%	62%
	20	61%	70%
178	2	69%	62%
	10	66%	71%
	20	68%	66%
Moyenne	2	66%	63%
	10	65%	66%
	20	65%	67%

FIGURE 6.4 – Sensibilité du GLM Poisson et de la random forest à la graine d'échantillonnage et au paramètrage du k-NN

6.4 Optimisation des modèles - Démarche générale

Forts des éléments précédents, nous retenons l'approche décrite figure 6.5 : nous commencons par chercher le ré-échantillonnage optimal (taux, effectif) pour la régression de Poisson (figure 6.6). Une fois l'optimum identifié, nous utilisons les résultats du modèle afin de procéder au regroupement des modalités de mêmes effets purs. Les variables à nombreuses modalités ainsi corrigées, nous procédons à l'optimisation du ré-échantillonnage (taux, effectif, k-NN) pour le CART (figure 6.9). Nous utilisons enfin le résultat des CART (et de random forests) pour identifier les variables d'effet non-linéaire et les interactions de variables. Ces éléments contribueront alors à enrichir la spécification du prédicteur linéaire au sein du GLM.

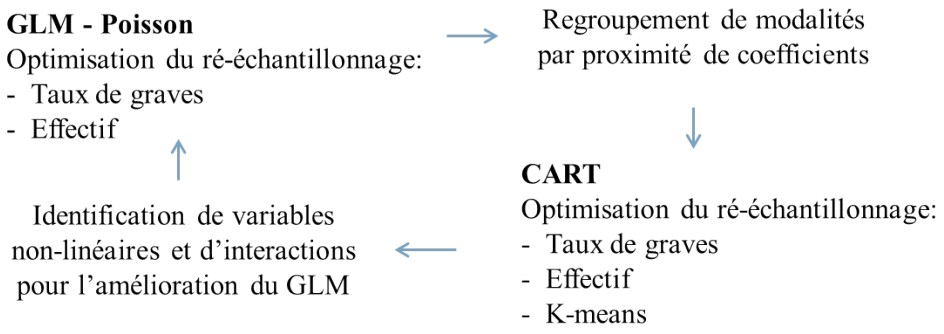


FIGURE 6.5 – Schéma général d'optimisation des modèles linéaire et CART

6.5 Optimisation du GLM Poisson

Pour l'optimisation du GLM, nous reprenons l'étude des paramètres de taux et d'effectif, en nous limitant aux taux 10, 30 et 50 %, ainsi qu'aux effectifs 50 000, 100 000 et 250 000 (nous nous limitons à ce plafond pour des raisons de temps de calcul). Nous explorons en outre le potentiel d'amélioration de la performance par l'élimination des variables non-explicatives "parasites" préalablement au ré-échantillonnage (cf partie 5.4.2) : à l'issue d'un premier ajustement du GLM, nous retirons de la base d'apprentissage les variables qui n'ont pas été sélectionnées par le LASSO (i.e. de coefficient nul). Nous ré-échantillonnons cette base simplifiée puis réajustons le modèle. La figure 6.6 résume la méthode employée. Cette méthode est utilisée pour 10 échantillonnages différents (i.e. 10 graines différentes). Ce sont donc 10 échantillonnages x 3 Effectifs x 3 Taux = 90 modèles qui sont ajustés, ainsi que 90 modèles sur les bases réduites associées.

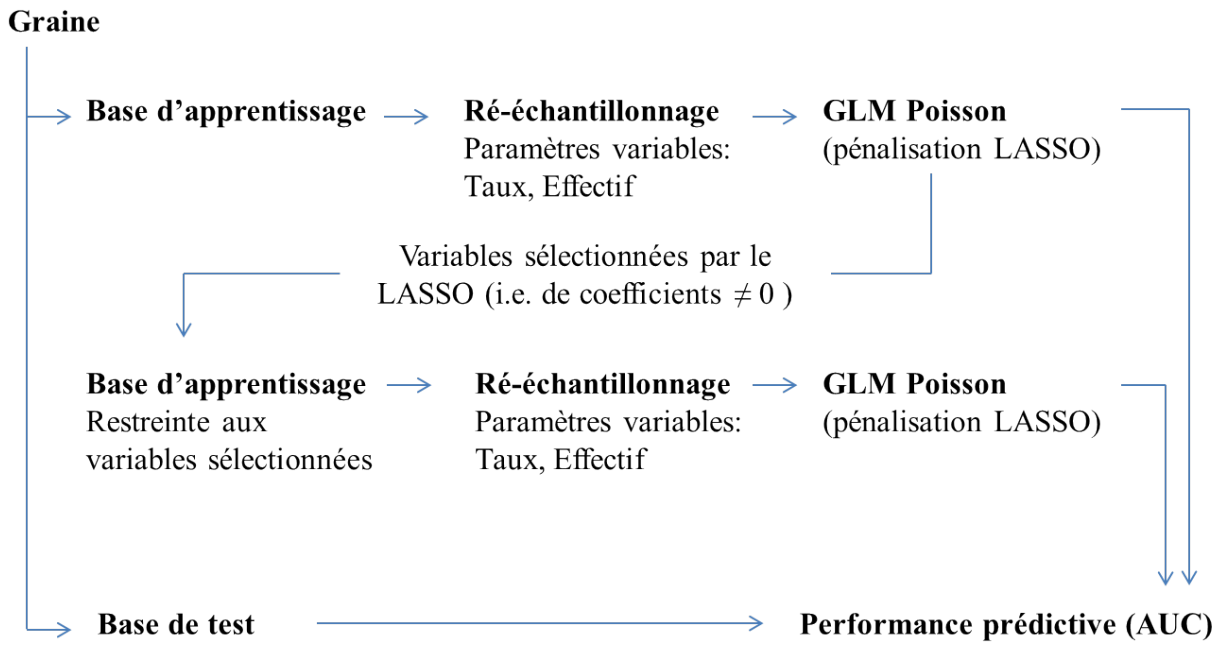


FIGURE 6.6 – Optimisation du GLM Poisson par le rééchantillonnage : taux, effectif et réduction de la dimension

Les résultats de cette analyse indiquent que les possibilités d'optimisation du GLM par le réglage fin du ré-échantillonnage sont faibles :

— La réduction de dimension de la base d'apprentissage n'améliore pas la performance des modèles, voire la dégrade légèrement : sur les 90 ajustements réalisés à partir de bases d'apprentissage non-réduites, on obtient un AUC moyen de 66.50 %, contre 66.30% à partir des 90 bases réduites. Aussi avons-nous laissé les résultats des bases réduites de côté pour nous concentrer sur ceux des bases non-réduites (figures 6.7 et 6.8). Notons qu'une étude du SMOTE (Blagus & Lusa [2]) sur des données de grande dimension (jusqu'à un millier de variables) contraste avec ce résultat, concluant à la nécessité d'une sélection des variables préalablement au SMOTE.
— La variation du taux de graves de 10 à 50 % n'a pas d'effet sensible sur la performance (figure 6.7).
— L'accroissement de l'effectif de la base ré-échantillonnée permet une légère amélioration de la performance : en multipliant l'effectif par 5 (de 50 à 250 000), l'AUC moyen augmente de 1.1 point (figure 6.7).

		Taux de graves			
		10%	**30%**	**50%**	**AUC moyen**
Effectif	50000	66,7%	65,7%	65,3%	65,9%
	150000	66,4%	65,6%	67,7%	66,6%
	250000	66,7%	67,7%	66,6%	67,0%
AUC moyen		66,6%	66,3%	66,6%	66,5%

FIGURE 6.7 – Performances des ajustements sur bases d'apprentissage ré-échantillonnées (non-réduites)
Chaque valeur d'AUC correspond à la moyenne de dix AUCs mesurés sur des bases d'apprentissage ré-échantillonnées à partir de différents échantillonnages apprentissage/test (10 graines différentes).

La figure 6.8 indique, pour chacune des graines utilisées dans l'échantillonnage train/test, la performance moyenne obtenue par les 9 ajustements du GLM sur les bases ré-échantillonnées (3 Taux x 3 Effectifs). La performance est très variable, avec dix points d'AUC entre la plus faible et la plus haute mesure de performance (63% vs 73%). Ces résultats confirment qu'il est essentiel de mesurer la performance à partir de différents échantillonnages pour consolider les résultats d'optimisation du ré-échantillonnage.

Graine	AUC moyen
16	71%
223	65%
287	70%
399	64%
490	63%
575	65%
584	66%
771	65%
944	73%
978	63%

FIGURE 6.8 – Sensibilité de la performance mesurée des GLM à l'échantillonnage de la base d'apprentissage (bases non-réduites).

6.5.1 Stabilité de la sélection de variables et stabilité des coefficients

Parallèlement à la stabilité des performances, nous souhaitons mesurer la stabilité de la sélection des variables principales par le modèle. Pour ce faire, nous ajustons un GLM sur des bases ré-échantillonnées (taux de 50% et effectif de 50 000 pour limiter le temps de calcul) à partir de dix bases d'apprentissage issus d'échantillonnages apprentissage/test différents. Nous examinons ensuite les 20 variables les plus importantes du modèle, i.e. de coefficients les plus élevés en valeur absolue, et calculons la fréquence avec laquelle chaque variable est sélectionnée parmi les 20 plus importantes. La sélection apparaît assez peu stable : 7 variables sont sélectionnées parmi les 20 plus importantes dans 80% des modèles, et 7 autres dans 70% des modèles.

Dans un deuxième temps, en imposant au modèle une pré-sélection des variables les plus importantes (nous ne fournissons au modèle que 5, 10 ou 20 variables), nous examinons la stabilité des coefficients. Un *trade-off* entre nombre de variables et stabilité des coefficients apparaît alors : plus le nombre de variables employé est grand, moins leurs coefficients sont stables.

6.5.2 Stabilité des regroupements de modalités

Nous avons vu que la méthode la plus fiable pour regrouper les modalités d'une variable catégorielle était le GLM. Quelle est la stabilité des regroupements obtenus avec cette méthode ? Pour l'évaluer, nous ajustons un GLM sur 7 bases d'apprentissage différentes (ré-échantillonnées à un taux de 50% et un effectif de 50 000). Nous formons à partir du résultat de chaque ajustement - c'est-à-dire des valeurs des coefficients estimés pour les fonctions indicatrices des modalités - cinq groupes de modalités au sein de la variable considérée. Dans le cas de la variable Activité (231 modalités), 24% des modalités appartiennent systématiquement au même groupe, et 23% appartiennent à deux groupes différents (mais voisins) selon l'ajustement considéré. Il s'agit de modalités d'effet significatif, qui appartiennent de façon stable aux groupes extrêmes, de coefficients fortement positifs ou négatifs. Les 53% restants des modalités se comportent de façon moins stable ; on les trouve dans 3 ou 4 groupes différents, ce qui s'explique par leur moindre significativité. Nous pourrions donc réduire le nombre de groupes, en ne conservant que trois groupes, deux extrêmes et un intermédiaire contenant les modalités instables. Les résultats obtenus sur la variable Département sont plus stables : 96% des observations n'appartiennent qu'à un ou deux groupes (voisins).

6.6 Optimisation des CART

Ayant appris à regrouper les variables Activité et Département, nous étudions (figure 6.9) la sensibilité des CART (Classification et Régression de Poisson) aux paramètres du ré-échantillonnage : le taux, l'effectif et le nombre k de plus proches voisins. Pour chaque échantillonnage apprentissage/test, nous ajustons un nouveau GLM et mettons à jour les groupes de modalités. Fixer les groupes conduirait à un surapprentissage : les groupes ayant été formés par un GLM ajusté sur une base d'apprentissage, un nouvel échantillonnage aléatoire génère une base de test dont quatre observations sur cinq (en espérance) sont issues de la base d'apprentissage. Ainsi les variables catégorielles regroupées sur lesquelles s'appuient les CART contiennent de l'information sur la base de test, et nous retrouvons le problème de surestimation de la performance.

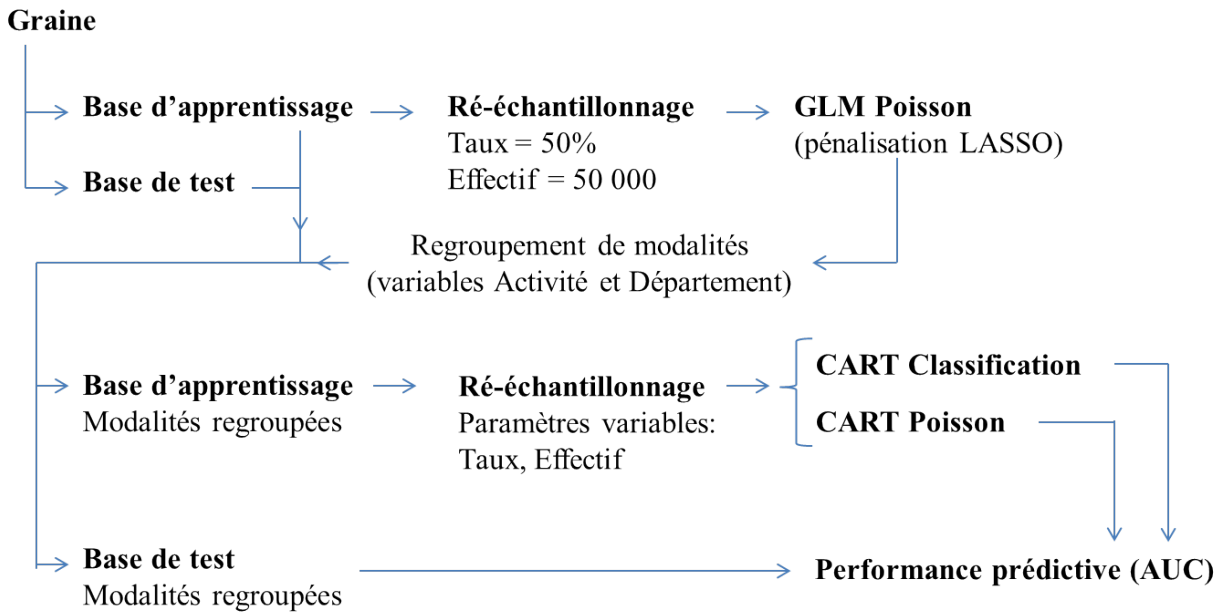


FIGURE 6.9 – Optimisation des CART par le ré-échantillonnage : taux, effectif et k-NN

Les arbres de classification présentent un AUC moyen systématiquement inférieur aux arbres de régression de Poisson : nous les excluons de la suite de l'étude, et ne détaillons que les résultats obtenus sur les CART de régression de Poisson.

Les trois meilleurs modèles sont obtenus par arbres de régression de Poisson (Figure 8.1 en annexe) pour les combinaisons de paramètres suivantes :

— k = 10, taux = 10% et Effectif = 250 000 : AUC moyen de 74.6%.
— k = 10, taux = 10% et Effectif = 150 000 : AUC moyen de 74.4%.
— k = 30, taux = 1% et Effectif = 150 000 : AUC moyen de 74.1%.

Au sein des arbres de régression de Poisson, il apparaît que nous pouvons d'emblée exclure le paramètre k = 5, qui conduit à un AUC moyen sensiblement inférieur aux autres valeurs de k, et dont aucun des AUCs moyens par couple Taux-Effectif ne figure parmi les trois plus élevés. Nous excluons les taux 30 et 50% par le même raisonnement.

Nous recalculons les marges sur ces paramètres restreints (Figure 8.2 en Annexe). Le jeu de paramètres k=10, taux = 10% et Effectif = 150 000 semble alors un bon choix : il correspond au deuxième meilleur modèle (proche du premier en termes de performance et de paramètres), et le taux de 10% est celui qui fournit le meilleur AUC moyen, tout comme l'effectif 150 000.

La figure 6.10 indique, pour chacune des 6 graines utilisées dans l'échantillonnage apprentissage/test, la performance moyenne obtenue par les 12 ajustements du CART Poisson sur les bases ré-échantillonnées (4 Taux x 3 Effectifs). La performance est encore davantage variable que pour le GLM, avec plus de 13 points d'AUC entre la plus faible et la plus haute performance mesurée (65.5% vs 79.0%). Il est donc tout aussi important pour les CART de mesurer la performance à partir de différents échantillonnages - afin de consolider les résultats d'optimisation du ré-échantillonnage - que pour les GLMs. Notons que cette forte sensibilité à la graine contraste avec les résultats observés sur les random forests, et rappelle la faible stabilité des CARTs.

Graine	AUC moyen
16	70,2%
223	66,4%
399	70,0%
490	65,5%
584	68,4%
944	79,0%

FIGURE 6.10 – Sensibilité de la performance mesurée des CARTs à l'échantillonnage de la base d'apprentissage

Ayant réuni suffisamment d'information sur les variables explicatives (à travers les CART et les GLMs), nous mesurons enfin l'effet de la réduction de la dimension. Pour ce faire nous ne retenons que les 70 variables les plus importantes : nous divisons ainsi par deux le nombre de variables utilisées dans le k-NN. Procéder de la même façon que lors de l'optimisation des modèles linéaires - en utilisant la sélection de variables d'un CART pour la réduction de dimension - n'aurait pas été judicieux étant donnée la faible stabilité du CART. Nous reprenons les 6 bases d'apprentissage utilisées pour l'optimisationet les ré-échantillonnons avec les paramètres optimaux (k=10, taux = 10%, Effectif = 150 000). Les performances des 6 CART ajustés sont fournies en annexe (figure 8.3). Nous obtenons un AUC moyen de 73.9%, inférieur à celui obtenu sur les bases non-réduites (74.4%) : la réduction n'a pas d'effet sensible sur la performance de nos données.

6.6.1 Stabilité des sélections de variables

De même que dans le cas du GLM, nous souhaitons examiner la stabilité de la sélection de variables opérée par le CART. Pour ce faire, nous reprenons l'ajustement du CART sur les 6 bases d'apprentissage ré-échantillonnées de façon optimale (k=10, taux = 10%, Effectif = 150 000). Bien qu'ils soient élagués après validation croisée et ainsi protégés du surapprentissage, les arbres obtenus demeurent profonds (fréquemment plus d'une dizaine d'étages). Nous les élaguons de façon plus poussée afin d'identifier les 15 premières variables utilisées dans chaque arbre. La stabilité est moindre que dans le modèle linéaire : seules 6 variables sont présentes dans plus de la moitié des modèles. Ceci nous rappelle que les CART sont peu stables dans leur sélection de variables.

6.6.2 Benchmark des random forests

Si les random forests ne peuvent répondre directement à nos objectifs, elles demeurent une référence intéressante en termes de sélection de variables par leur stabilité. Nous appliquons donc aux random forests la même approche qu'aux CARTs, à l'optimisation près : pour des raisons de

temps, nous n'optimisons pas le ré-échantillonnage pour les random forests et nous utilisons le jeu de paramètres optimal retenu pour les CARTs (k=10, taux = 10%, Effectif = 150 000). Nous obtenons un AUC moyen de 71% pour la random forest ajustée sur bases d'apprentissage ré-échantillonnées, soit une performance inférieure à celle du CART à l'optimum. Cette faible performance de la random forest est imputable à son absence de prise en compte de l'exposition, à son faible nombre d'arbres (200, pour des raisons de capacité mémoire) et à ce qu'elle n'a pas fait l'objet d'une optimisation de la méthode de ré-échantillonnage.

Malgré cette performance limitée, nous atteignons l'objectif d'une sélection de variables plus stable. Nous considérons pour chaque ajustement la sélection des 15 variables plus importantes (en termes de décroissance moyenne de l'indice de Gini). La sélection est nettement plus stable que celle du CART : 10 variables sont sélectionnées par plus de la moitié des modèles, et 7 par la moitié d'entre eux.

6.7 Amélioration du modèle linéaire

Le modèle linéaire est sensiblement moins performant que le CART : son AUC moyen pour le meilleur ré-échantillonnage ne dépasse pas 67.7%, contre 74.4% pour le ré-échantillonnage optimal du CART. Cette différence indique que le prédicteur linéaire ne permet pas de modéliser pleinement l'effet des variables : les variables présentent des effets non-linéaires et/ou des effets d'interaction.

Les non-linéarités sont couramment intégrées aux modèles linéaires à l'aide de splines, i.e. de fonctions polynomiales par morceaux. Ces splines ajoutent des termes au prédicteur linéaire : un spline de fonction carrée en deux morceaux comporte ainsi 5 termes, auxquels sont associés des coefficients qui seront estimés dans l'algorithme de maximisation de vraisemblance. Les interactions sont, elles, prises en compte à l'aide de termes multiplicatifs : un premier moyen de mesurer l'interaction entre deux variables est d'ajouter au prédicteur linéaire un troisième terme correspondant au produit de ces deux variables. Intégrer directement au prédicteur linéaire l'ensemble des non-linéarités et des interactions possibles conduit à une trop forte inflation du nombre de paramètres : prendre en compte l'ensemble des termes d'interactions de 150 variables conduit par exemple à $150^2 = 22500$ coefficients à estimer, ce qui pose fréquemment des problèmes de mémoire et de temps de calcul, y compris pour le LASSO (notons qu'il demeure possible de réduire l'intervalle des valeurs de λ utilisées pour le calibrage du LASSO en ne conservant que des valeurs élévées, permettant une convergence rapide).

Nous nous appuyons sur la sélection des variables réalisée par les random forests pour retenir les meilleurs candidats à une modélisation avancée : ces variables sont les plus à même d'être fortement explicatives tout en ayant un effet non-linéaire et des interactions avec d'autres variables. L'hypothèse de non-linéarité est encore plus forte si une variable n'est pas conservée dans le modèle linéaire mais est retenue par les arbres.

Nous utilisons les 17 variables les plus importantes des random forests (cf ci-dessus) pour améliorer le modèle : un terme au carré pour chaque variable et un terme d'interaction pour chaque couple de variables sont ajoutés au prédicteur linéaire. Une variable catégorielle à m modalités ayant $m-1$ indicatrices, chaque interaction génère $m-1$ nouveaux termes au prédicteur. Aussi nous regroupons au préalable les modalités des variables Activité et Département. Nous comparons ainsi les performances de la régression de Poisson améliorée, sur 5 bases d'apprentissage (5 graines), selon le schéma décrit figure 6.11.

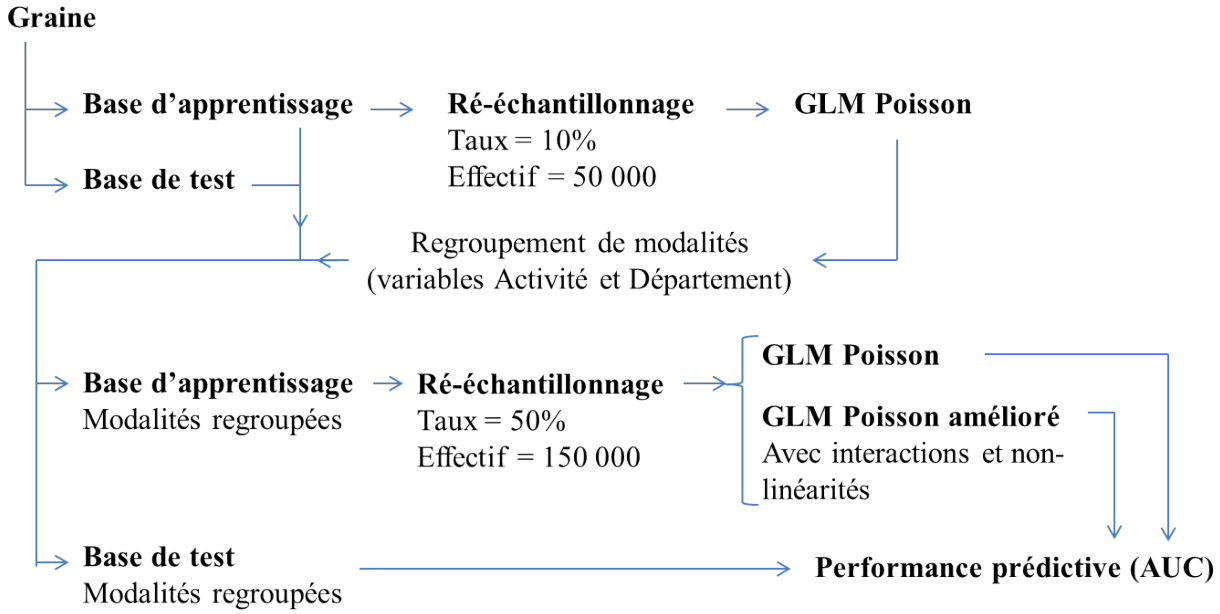


FIGURE 6.11 – Intégration d'interactions et de non-linéarités dans le GLM Poisson

Nous obtenons un AUC moyen de 70,0% du modèle amélioré contre 67.7% pour le modèle sans interaction ni non-linéarité : le potentiel d'amélioration du modèle linéaire demeure limité, et ne rejoint pas les performances du CART.

Analyse des effets univariés et multi-variés par les random forests

La random forest peut apporter bien davantage qu'une sélection de variables. Antoine Guillot illustre dans son mémoire [9] sur l'apprentissage statistique en tarification non-vie l'utilisation qui peut en être faite pour observer des effets univariés (détection des non-linéarités) ou multivariés (nature de l'interaction entre deux variables) afin de faire un ajout ciblé et intelligent des termes de non-linéarités et d'interactions dans le modèle linéaire.

Cette méthode repose sur la définition d'un profil-type (profil médian ou moyen). Ainsi pour étudier l'interaction de deux variables, on ajuste la random foret sur la population d'apprentissage puis l'on génère des individus couvrant l'ensemble des valeurs possibles pour le couple des deux variables étudiées, mais de valeurs constantes pour les autres variables (égales à leur médiane ou à leur moyenne selon le profil choisi). On prédit alors la fréquence de ces deux individus, et l'on obtient une surface de prédiction (figure 6.12) décrivant la nature de l'interaction pour le profil-type.

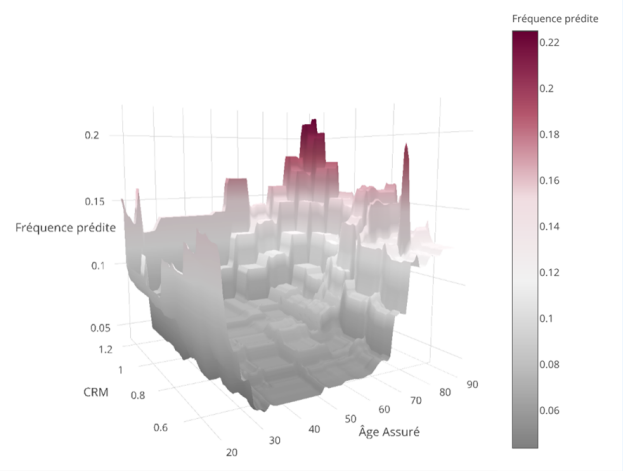


FIGURE 6.12 – Surface de fréquence prédite montrant l'interaction de deux variables pour un profil médian.

Source : Antoine Guillot [9]

Chapitre 7

Résultats des modèles

Pour l'ajustement final des modèles, nous cessons d'isoler une base de test et utilisons toutes les observations. Nous exploitons ainsi toute l'information disponible à ce jour. En ajoutant 20% au volume de la base d'apprentissage, nous améliorons l'espérance de performance du modèle sur de futures données. Nous ne pouvons pour autant affirmer quelle sera la performance du modèle au vu de la variabilité des profils de risque (dont témoignent les variations observées selon l'échantillonnage apprentissage/test). Il est possible que les sinistres dont nous disposons soient bien représentatifs de la distribution des profils de risque, auquel cas notre modèle présentera une bonne performance - à long terme, pour que la loi des grands nombres s'applique aux nouveaux sinistres. Il est également possible que l'échantillon de 151 sinistres dont nous disposons soit peu représentatif de la distribution des risques, auquel cas nos modèles seront en dessous de la performance espérée.

Si nous nous affranchissons ici de la base de test, nous continuons de nous prémunir contre le surapprentissage par validation croisée pour le calibrage des paramètres du LASSO et du CART.

La réalisation des trois modèles finaux est présentée figure 7.1. Les modèles GLM sont des régressions de Poisson pénalisées par LASSO, et le modèle CART est celui fondé sur la régression de Poisson. Le ré-échantillonnage préalable aux GLM est paramétré avec le couple (Taux = 50%, Effectif = 150 000), l'un des deux couples apportant la meilleure performance moyenne mesurée (AUC de 67.7%), et le nombre de plus proches voisins est maintenu à sa valeur par défaut, k=10. Le ré-échantillonnage préalable au CART est lui paramétré avec le triplet optimal (k=10, Taux = 10%, Effectif = 150 000).

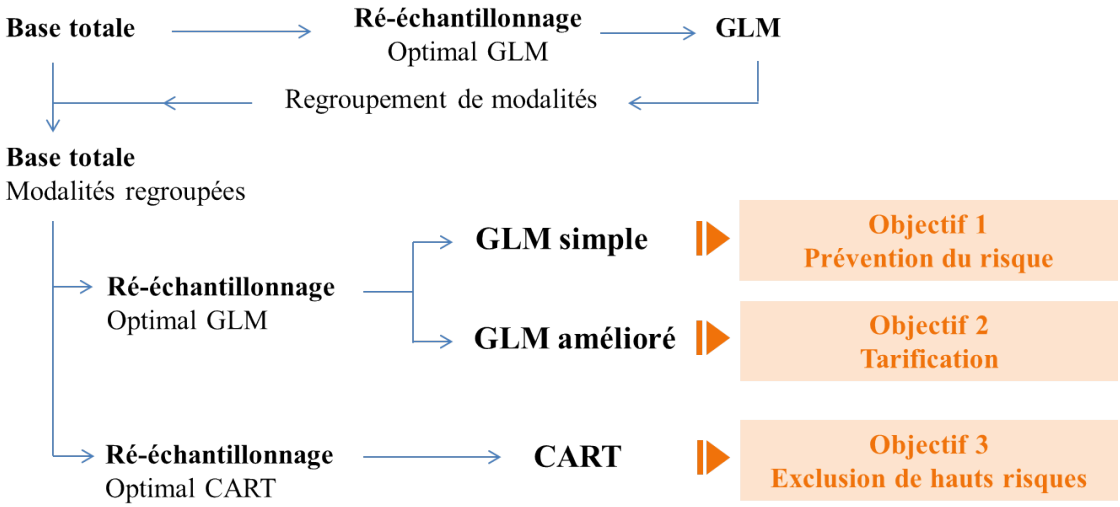


FIGURE 7.1 – Ajustement final des modèles GLM et CART

7.1 Régression de Poisson

7.1.1 Analyse des corrélations pour l'interprétation

La présence de variables fortement corrélées n'est pas souhaitable pour créer un outil tarifaire à partir d'un modèle linéaire généralisé, même lorsque la sélection de variables est assurée par le LASSO. Ces variables réduisent l'interprétabilité du modèle et sa stabilité. Entre deux variables fortement corrélées, nous choisirons la plus interprétable, ou la plus simple à obtenir.

L'analyse fait apparaître deux zones principales de fortes corrélations :

— Au sein des variables externes, les variables indiquant les taux de délits par département sont fortement corrélées, ce que montre la thermocarte figure 7.2. Nous choisissons de ne conserver que la variable de taux d'incendies volontaires de biens privés.

— Les variables d'antécédents (Figure 8.4 en Annexe) de long terme et court terme présentent de fortes corrélations. Ceci naît du fait que les antécédents de court terme sont inclus dans les antécédents de long terme. Nous choisissons de ne conserver que les antécédents de long terme. La variable de fréquence agrégée étant construite à partir des autres variables d'antécédents, nous la retirons également.

Nous n'observons pas de corrélations problématiques au sein des variables contractuelles (Figure 8.5 en Annexe) hormis pour la variable de "Sinistre Raisonnablement Escompté", construite à partir d'autres variables, et que nous retirons.

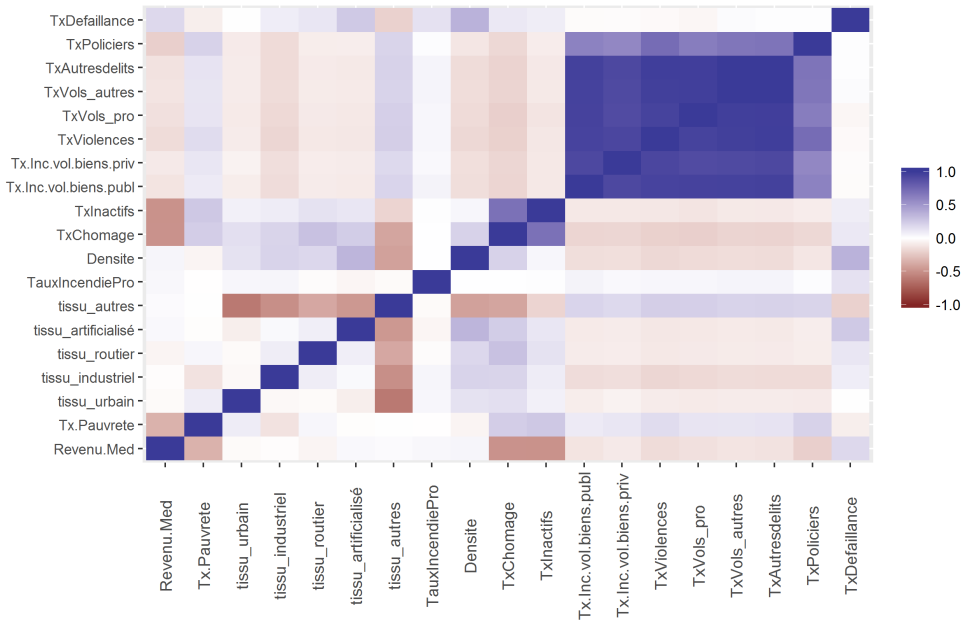


FIGURE 7.2 – Coefficient de corrélation - Variables externes

7.1.2 Guide d'interprétation d'une régression de Poisson pénalisée par LASSO

Notre régression de Poisson pénalisée par LASSO est ajustée sur des variables normalisées (pour les besoins du LASSO et du suréchantillonnage synthétique). Il est nécessaire de conserver ces éléments à l'esprit lors de l'interprétation des coefficients estimés par la régression.

La figure 7.3 détaille la méthode de calcul de la fréquence prédite par la régression de Poisson sur de nouvelles données. Les variables sont normalisées, puisque les coefficients sont estimés sur des variables normalisées. De plus les fréquences prédites doivent être corrigées d'un ratio : la régression a été ajustée sur une base suréchantillonnée, qui présente une fréquence bien supérieure à la fréquence réelle. Nous corrigeons donc la fréquence prédite pour chaque observation par le ratio

$$\frac{\text{Fréquence moyenne réelle}}{\text{Fréquence moyenne prédite}}$$

.

Données à prédire

Antécédents d'incendie (fréquence annuelle)	Surface (m2)
3	1000
0	800
1	100

Données à prédire, variables normalisées

Normalisation: $\dfrac{\boldsymbol{x - min(var)}}{\boldsymbol{max(var) - min(var)}}$

Antécédents d'incendie (fréquence annuelle)	Surface (m2)
1	1
0	0.8
0.33	0.1

Fréquences prédites sur les données

log(Y) = β0 + x1*β1 + x2*β2

Antécédents d'incendie (fréquence annuelle)	Surface (m2)	Fréquence prédite
1	1	= exp(β0 + 1*β1 + 1*β2) * Ratio
0	0.8	= exp(β0 +0*β1 + 0.8*β2) * Ratio
0.33	0.1	= exp(β0 +0.33*β1 + 0.1*β2) * Ratio

Correction de la sur-fréquence liée au ré-échantillonnage:

$$\text{Ratio} = \frac{Fréquence\ moyenne\ réelle}{Fréquence\ moyenne\ prédite}$$

FIGURE 7.3 – Coefficient de corrélation - Variables externes

Nous pouvons comparer les coefficients de la façon suivante : à position égale de l'observation entre le min et le max (i.e. à même valeur normalisée), une variable de coefficient $\beta_1 = k \cdot \beta_2$ a un effet k fois plus important sur le logarithme de la fréquence. Nous montrons ci-dessous une autre méthode, plus explicite.

7.1.3 Résultats et interprétation du modèle linéaire simple

Le nombre de variables sélectionnées dans le modèle par le LASSO (par validation croisée, avec la *one standard error rule*) conduit à conserver 54 variables numériques et fonctions indicatrices. Nous savons pour autant que seules les variables les plus importantes de la régression sont stables : nous l'avons observé partie 6.5.1 avec la variabilité des sélections du LASSO et le *trade-off* entre nombre de variables et stabilité des coefficients. Nous décidons par conséquent de ne retenir que les 12 premières variables du modèle, soit les 12 variables de coefficients les plus élevés en valeur absolue, qui représentent plus de 75% du pouvoir explicatif de la régression (ratio des sommes de coefficients en valeur absolue). Nous présentons le résultat obtenu figure 7.4. Les coefficients sont tous positifs pour ces variables, mais certaines des variables non-retenues présentent des coefficients négatifs.

Variable	Coefficient
Fréquence d'antécédents d'incendie	2,48
Fréquence d'antécédents de sinistres divers (bris de machine, dommages divers, frais fixes, pertes de marchandises périssables...)	2,23
SURFACE	1,80
Valeur de marge brute	1,71
Valeur de contenu	1,34
Valeur vénale du fond	1,31
Capital de marchandises transportées	1,06
Taux de chômage	1,06
Groupe d'activités 5	0,93
Groupe de départements 5	0,72
Groupe d'activités 4	0,70
Capital de marchandises périssables	0,57

FIGURE 7.4 – Coefficients de régression pénalisée

Notre régression de Poisson livre les informations suivantes :

- La variable la plus explicative de la sinistralité est la variable d'antécédents d'incendies, de coefficient égal à +2.48. Ce résultat semble pleinement cohérent, et incite à une prévention ciblée des profils ayant déjà connu un incendie.
  Mesurons l'effet de cette variable sur la fréquence prédite. Considérons que la fréquence annuelle d'incendie maximale observée sur le portefeuille est d'un incendie tous les deux ans, soit 0.5 (ce chiffre n'est pas exact, pour des raisons de confidentialité). La fréquence minimale observée est naturellement de 0. Mesurons l'augmentation de la fréquence prédite à un assuré de profil "intermédiaire" (i.e. dont les valeurs sont situées à mi-chemin du minimum et du maximum de chaque variable) ayant connu un incendie au cours des trois dernières années. Sa fréquence est égale à 0.33, ce qui donne après normalisation $(0.33 - 0)/(0.5 - 0) = 0.67$. Son profil étant "intermédiaire", les valeurs de ses variables normalisées sont égales à 0.5 . La fréquence prédite de cet individu s'il n'avait pas eu d'incendie est donc égale à 0.165 .

$$exp(\beta_0 + 0 \cdot \beta_{\text{Antécédents d'incendie}} + 0.5 \cdot \beta_{\text{Antécédents divers}} + 0.5 \cdot \beta_{\text{Surface}} + ...)$$

$$= exp(-8.5 + 0 \cdot 2.48 + 0.5 \cdot 2.23 + 0.5 \cdot 1.80 + ...) = exp(-1.8) = 0.165$$

  Sa fréquence prédite sachant qu'il a eu un incendie est égale à 0.903.

$$exp(\beta_0 + 0.67 \cdot \beta_{\text{Antécédents d'incendie}} + 0.5 \cdot \beta_{\text{Antécédents divers}} + 0.5 \cdot \beta_{\text{Surface}} + ...)$$

$$= exp(-8.5 + 0.67 \cdot 2.48 + 0.5 \cdot 2.23 + 0.5 \cdot 1.80 + ...) = exp(-0.1) = 0.903$$

  Ainsi, la survenance d'un sinistre incendie pour un profil intermédiaire ayant souscrit il y a 3 ans multiplie le risque d'incendie grave par 5.5 (0.903/0.165).
  Nous observons des fréquences prédites élevées : ceci tient au ré-échantillonnage de notre base d'apprentissage, qui a un taux de graves de 50%. Les fréquences réelles sont obtenues

par l'application du ratio vu figure 7.3. Ce ratio ne modifie pas le facteur de 5.5 obtenu ; l'interprétation faite sur la base ré-échantillonnée reste vraie sur la base originale.

- La deuxième variable la plus explicative est la variable agrégée d'antécédents de sinistres sur un ensemble de garanties (+2.23). Cette dernière variable agrège les antécédents sur des garanties autres que les vols, les responsabilités civiles, les dommages électriques et les dégâts des eaux. Elle comporte par exemple les garanties de bris de machine, dommages divers, frais fixes et pertes de marchandises périssables. Ce résultat suggère que le niveau de sinistralité "de fond" d'un assuré est un bon indicateur de son risque de grave, ou que certaines de ces garanties sont explicatives du risque grave. Il conviendra d'examiner l'impact individuel de ces garanties, que nous avions regroupées en considérant qu'elles auraient peu d'effet, hypothèse qui s'avère fausse.

- La valeur de marge brute (coefficient égal à +1.71), la valeur vénale du fonds (+1.31) et la surface (+1.80) sont très explicatives. Les capitaux garantis sont également explicatifs : la valeur de contenu (+1.34), le capital de marchandises transportées (+1.06) et le capital de marchandises périssables (+0.57).
  L'interprétation des variables qui déterminent le coût du sinistre (capitaux garantis, valeur vénale et marge brute) peut être de deux natures. La première concerne leur effet sur la survenance de l'incendie, la seconde concerne l'effet des variables sur le coût du sinistre. La première interprétation envisage l'effet positif de ces variables comme un effet taille, hypothèse confortée par l'effet également fort et positif de la surface : plus le lieu d'une activité est grand et développé, plus le risque de survenance est fort. La seconde tient à l'augmentation mécanique du coût d'un sinistre avec ces variables : nous avons vu que la garantie de perte d'exploitation génère fréquemment une augmentation de la charge, de 5 à 200%, et que la garantie de perte de valeur vénale du fonds occasionne aussi une augmentation de la charge pouvant atteindre 200%. L'effet des variables peut donc transiter par une action à la marge sur les sinistres de montant inférieur au seuil de 70 k€ : à cause et circonstances égales, elles peuvent faire la différence entre un sinistre à 50 k€ et un sinistre à 90 k€, entre un non-grave et un grave.
  D'autre part, la présence de marchandises périssables nécessite souvent des installations frigorifiques, source possible de départ de feux. Le fort effet de la marge brute et de la valeur vénale du fonds peut également transiter par un aléa moral : ces variables déterminent le montant de l'indemnité au titre des garanties de perte d'exploitation et de perte de valeur vénale du fonds.

- L'activité intervient comme une variable assez explicative : les deux groupes d'activités de plus haut risque sont sélectionnés. Le groupe d'activités 4 présente un coefficient inférieur au groupe 5, ce qui est cohérent avec la façon dont ils ont été construits. Le groupe des départements les plus risqués est également sélectionné. Une grille des activités et un zonier des départements sont donc utiles dans le modèle de prime pure grave.

7.1.4 Résultats et interprétation du modèle linéaire amélioré

Le modèle linéaire amélioré conserve le carré de la valeur de contenu (-9.54) et de la surface (-5.45) : l'effet de ces variables est donc concave. Ces résultats peuvent indiquer que si le risque augmente avec la taille d'une activité (ce que nous avons vu dans le modèle linéaire simple), le niveau de prévention augmente également, réduisant l'effet marginal de la taille. Le modèle conserve deux interactions. La valeur de marge brute et le capital dommages électriques présentent une interaction négative (-3.63). Ceci peut à nouveau indiquer un lien entre taille et actions préventives : lorsqu'une activité présente à la fois une forte rentabilité et d'importantes installations électriques, la prévention du risque y est plus développée. La valeur de marge brute et l'effectif total présentent aussi une interaction négative, de plus faible valeur (-0.86) : les entreprises dont la marge brute repose avant tout sur un capital humain sont moins risquées.

7.1.5 Création de groupes tarifaires

Quel est le degré de discrimination tarifaire fourni par notre modèle ? Combien de groupes pouvons-nous générer, et avec quel écart de tarif ? La figure 7.5 fournit les premiers éléments de réponse à cette question. Nous découpons les données originales en dix groupes, définis par leurs fréquences prédites. Les bornes qui délimitent ces groupes correspondent aux déciles (arrondis à l'entier le plus proche). Nous observons en rouge la fréquence moyenne prédite au sein du groupe (corrigée du ratio Fréquence moyenne réelle / Fréquence moyenne prédite), et en bleu la fréquence moyenne réellement observée au sein de ce groupe. Fréquences prédites et observées sont ensuite multipliées dans ce graphique par un facteur arbitraire, pour des raisons de confidentialité.

Les 5 premiers groupes tarifaires figure 7.5 ne présentent aucun sinistre grave. Notre modèle discrimine donc efficacement une moitié de la population très peu risquée. La discrimination tarifaire au sein de cette première grande population semble cependant peu utile. Les trois groupes suivants (de 35-50 à 60-75) présentent une évolution non-monotone de la fréquence de graves observée, qui tient au faible nombre de graves qu'ils contiennent (moins de 15 graves par groupe). La discrimination entre ces groupes semble inopportune du fait de l'incohérence entre les évolutions des fréquences prédites et observées, et il semble plus judicieux d'en faire un deuxième grand groupe. Enfin, les deux derniers groupes se démarquent nettement par une fréquence élevée et croissante de graves, et peuvent constituer chacun un groupe tarifaire, voire être affinés.

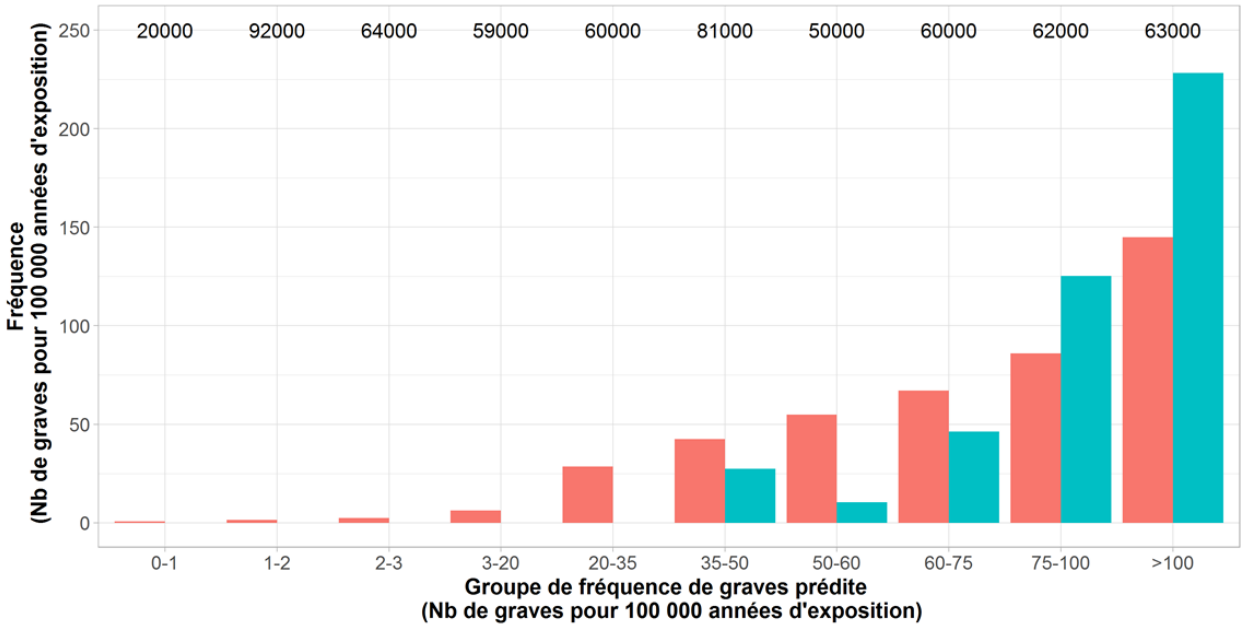


FIGURE 7.5 – Exemple de groupes tarifaires générés à partir du modèle linéaire sur les données originales
En rouge : fréquence moyenne prédite au sein du groupe. En bleu : fréquence moyenne observée

7.2 CART - Résultats et interprétation

Le CART ajusté sur les données ré-échantillonnées est élagué avec la valeur du paramètre de complexité correspondant à la *one standard error rule*. La lecture des arbres issus de bases ré-échantillonnées n'est pas immédiate. Nous générons un grand nombre d'individus fictifs, et pour tirer des conclusions de l'arbre, il est nécessaire de le reconstruire sur la base de données réelles afin de connaître le véritable nombre d'observations et la véritable fréquence de graves au sein de chaque feuille.

Ainsi une première lecture de l'arbre - avant reconstruction sur la base originale - conduit à identifier rapidement un noeud (à gauche de l'arbre) contenant 12 000 observations avec une fréquence de 99%. La reconstruction de ce noeud sur la base originale fournit une feuille avec quelques opbservations seulement, et aucun sinistre grave. Ceci vient du problème de la réduction de variables vu précédemment, qui touche fortement les variables discrètes. Nos variables d'antécédents de sinistres sont discrètes car elles sont le quotient de variables de comptage prenant des valeurs entières et peu dispersées (couramment entre 1 et 5) et d'une variable d'exposition prenant elle même des valeurs entières peu dispersées. Ces variables explicatives comportent donc des intervalles vides ne contenant presque aucune observation, mais dans lesquels seront générés des individus synthétiques. L'arbre identifie alors rapidement ces zones "pures". Dans la mesure où les observations futures ne seront jamais rattachées à ces zones, ces zones ne sont pas en elle-mêmes des sources

d'erreur prédictive. Elles correspondent à une masse d'individus synthétiques "produits pour rien". Notre suréchantillonnage synthétique a ici conduit à 12 000 individus sans intérêts à cause de la souplesse prédictive du CART, mais les 3000 restants seront suffisants pour extraire l'information disponible des données, grâce à cette même souplesse.

Nous représentons un extrait de l'arbre reconstruit sur les données réelles figure 7.6. La première variable utilisée par l'arbre est la fréquence d'incendies depuis l'entrée en portefeuille : elle isole dans les données originales un groupe de plus de 400 000 observations (dont fréquence d'incendies est inférieure au seuil). La variable Surface divise alors ce groupe en un groupe (> 200 000 observations) de petites surfaces, de fréquence de graves divisée par 4.5 par rapport à la fréquence globale de la population, et un autre groupe (>200 000 observations) de grandes surfaces, de fréquence multipliée par 1.7 (par rapport à la fréquence globale de la population). Ainsi, grâce à deux critères seulement, le CART génère deux grands groupes tarifaires, le premier 7 fois moins risqué que le deuxième.

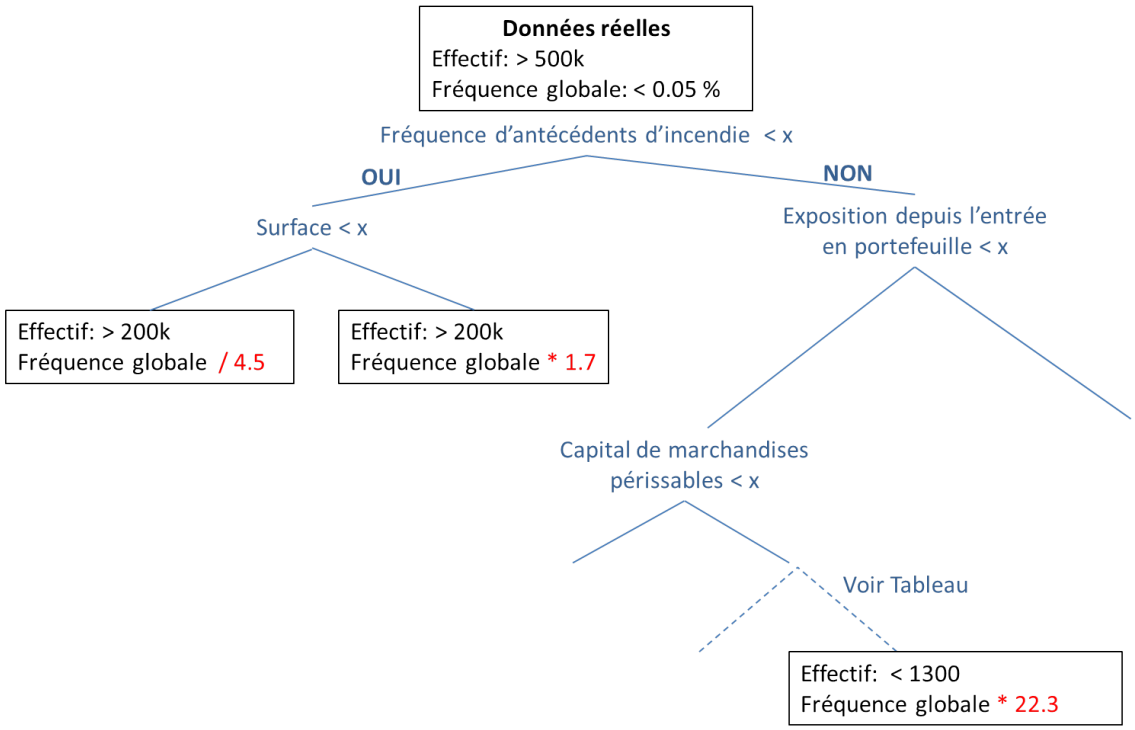


FIGURE 7.6 – Extrait du CART reconstruit sur les données réelles

Notre recherche d'une sous-population d'individus particulièrement risqués nous conduit à une population d'observations ayant connu des sinistres incendies, de faible ancienneté et dont le capital de marchandises périssables dépasse un certain seuil. Il s'agit d'une population de 16 900 observations, avec une fréquence multipliée par 2.9 (par rapport à la fréquence globale). L'arbre permet d'affiner encore ce groupe, ce que nous résumons figure 7.7.

Critères successifs	Nombre d'observations de la feuille	Fréquence observée dans la feuille / Fréquence globale
Fréquence d'antécédents d'incendie > x	< 150 000	0.97
Exposition depuis l'entrée en portefeuille < x	<100 000	0.86
Capital de marchandises périssables > x	< 20 000	2.9
Appartenance aux groupes de départements les plus risqués	< 5000	7
Taux de policiers par habitant < x	< 4000	9.3
Surface >= x	< 3000	13.4
Fréquence d'antécédents de vol depuis l'entrée en portefeuille > x	< 2000	17.2
Appartenance aux groupes d'activités les plus risqués	< 1500	22.3

FIGURE 7.7 – Groupe de hauts risques identifié par le CART

Nous obtenons à partir de 8 critères de décision un groupe de moins de 1500 observations, d'une fréquence plus de 22 fois supérieure à la fréquence globale. Ce groupe est d'effectif suffisamment faible pour faire l'objet d'un traitement particulier par les Caisses Régionales. Les critères contractuels utilisés sont intuitifs, et par conséquents simples à argumenter. Nous ciblons des contrats récents présentant déjà des antécédents d'incendie et de vol (ou un risque élevé si l'observation correspond à une entrée en portefeuille, auquel cas sa valeur est issue de l'imputation par arbre de régression décrite partie 1.7.3), avec une surface supérieure à la moyenne et un capital de marchandises périssables important. Notons que notre zonier de département et nos groupes d'activités entrent en jeu dans la détection de ce profil risqué, preuve de l'intérêt d'un traitement adapté de ces variables. Les variables externes nous permettent également d'affiner le profil, avec le taux de policiers.

Il est possible d'affiner davantage le groupe, à partir de critères tels que le taux de pauvreté et le nombre d'apprentis : avec ces deux critères nous obtenons un groupe d'un millier d'observations, d'une fréquence multipliée par 26.6. La possibilité de sur-apprendre par un trop grand nombre de critères, et la plus faible interprétabilité de ces critères rend toutefois hasardeuse la recherche d'un profil encore plus risqué.

Conclusion

L'apport des méthodes d'apprentissage statistique à la modélisation des sinistres graves apparaît majeur dans notre travail. Confrontés à des données de faible volume, très hétérogènes et extrêmement déséquilibrées, les modèles linéaires sont inopérants. Nous parvenons, grâce au suréchantillonnage synthétique, à développer leur capacité d'apprentissage jusqu'à un niveau de performance satisfaisant (jusqu'à 68% d'AUC). Nous atteignons ainsi l'objectif d'un modèle de prime pure pour les incendies graves.

Les arbres de régression apportent, par la souplesse de leurs capacités d'apprentissage, des informations complémentaires à celles du modèle linéaire. Ils bénéficient également du rééquilibrage des données par le suréchantillonnage synthétique : celui-ci accroît leur performance au-delà de celle des modèles linéaires (jusqu'à plus de 74 %). Ils identifient alors des groupes de haut risque où la fréquence de graves est plus de 20 fois supérieure, et ce en moins d'une dizaine de règles de décision. Les arbres sont également susceptibles de dégager des critères simples et discriminants sur de grands volumes : en seulement deux critères, le CART nous permet d'isoler deux grands groupes tarifaires, l'un 7 fois plus risqué que l'autre.

Arbres et modèles linéaires n'apparaissent pas seulement complémentaires, mais en interaction. Le modèle linéaire est la meilleure méthode pour retraiter intelligemment les variables catégorielles très hétérogènes, qui ne peuvent être utilisées telles quelles par les arbres. En retour, les arbres offrent une perspective non-linéaire sur les données, qui constitue un bon outil pour améliorer la spécification du modèle.

Ces méthodes présentent un coût en termes de mise en oeuvre : elles ne sont implémentées que sur des logiciels de data science (R, Python) et leur transfert aux outils de tarification actuarielle n'est pas immédiat. En outre, les résultats des modèles linéaires ne sont plus directement interprétables, du fait de la normalisation des variables et de la présence d'individus fictifs dans les données d'apprentissage.

La régression de Poisson apparaît in fine peu sensible aux paramètres du suréchantillonnage synthétique. Nous pourrions par conséquent envisager d'ignorer l'étape d'optimisation, et de limiter ainsi le recours à l'apprentissage statistique, en substituant au LASSO une sélection stepwise (plus coûteuse en temps de calcul). L'utilisation des CART nécessite, elle, l'optimisation des paramètres, et doit être menée dans son intégralité sur les logiciels de data science avant que les règles de décision utiles ne puissent en être extraites.

Nos résultats montrent l'intérêt d'une révision bayésienne du risque des assurés année après année : les antécédents d'incendie et d'autres sinistres sont les deux variables les plus explicatives du modèle linéaire. Un dispositif de surveillance et de ré-évaluation du risque à partir de la sinistralité attritionnelle réduirait par conséquent la fréquence des graves. Ils montrent aussi l'intérêt des variables externes, avec un effet du taux de chômage. Les variables externes et d'antécédents sont d'autant plus utiles qu'elles sont obtenues en interne et ne nécessitent pas d'être intégrées au dispositif commercial ni aux questionnaires.

Bibliographie

[1] Younes Bensalah et al. *Steps in applying extreme value theory to finance : A review.* Bank of Canada, 2000.

[2] Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14, 2013.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[4] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.

[5] Arthur Charpentier. Exposure with binomial responses, url : https ://freakonometrics.hypotheses.org/3318. 2013.

[6] Arthur Charpentier. Ce que la courbe roc ne raconte pas, url : https ://freakonometrics.hypotheses.org. 2016.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote : synthetic minority over–sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[8] Paul Embrechts, Claudia Kluppelberg, and Thomas Mikosch. Modelling extremal events. *British Actuarial Journal*, 5(2) :465–465, 1999.

[9] Antoine Guillot. Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? *Mémoire pour l'admission à l'Institut des Actuaires*, 2015.

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

[11] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity : the lasso and generalizations.* CRC Press, 2015.

[12] Nathalie Japkowicz. The class imbalance problem : Significance and strategies. In *Proc. of the Int Conf. on Artificial Intelligence.* Citeseer, 2000.

[13] Christian Y. Robert. Cours de théorie des valeurs extrêmes, ensae m2 actuariat. 2016.

[14] Richard L Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, pages 67–90, 1985.

[15] Terry M Therneau and Elizabeth J Atkinson. An introduction to recursive partitioning using the rpart routines. 2015.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[17] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov) :2541–2563, 2006.

Table des figures

Chapitre 8

Annexes

8.1 Algorithme SMOTE

Algorithm 1: Algorithme SMOTE

Input: Nombre d'individus de la classe minoritaire T, taux de SMOTE N (<1 ou entier non–nul), Nombre de plus proches voisins k.
Output: N·T individus synthétiques de la classe minoritaire.
Si N < 100%, on choisit un échantillon aléatoire de N·T individus, dont chacun servira à générer un individu synthétique ;
if $N < 1$ **then**
    Randomiser le vecteur des T individus;
    T = N.T;
    N = 1 ;
end
$Nvariables$ = Nombre de variables;
$ORIGINAUX$ $(T \times Nvariables)$: matrice des individus originaux de la classe minoritaire;
$SYNTHETIQUES$ $(N \cdot T \times Nvariables)$: matrice des individus synthétiques générés ;
$NouvIndex \leftarrow 1$: variable de comptage des individus synthétiques générés;
for $i \leftarrow 1$ **to** T **do**
    Calculer les k plus proches voisins de l'individu i, dont les index correspondants (dans la matrice $ORIGINAUX$) sont stockés dans un vecteur $Voisins(k \times 1)$;
    for $j \leftarrow 1$ **to** N **do**
        Choisir aléatoirement un nombre nn entre 1 et k, qui indexe l'un des k voisins dans le vecteur $Voisins$;
        for $var \leftarrow 1$ **to** $Nvariables$ **do**
            $Distance = ORIGINAUX[Voisins[j], var] - ORIGINAUX[i, var]$;
            $Ecart = Nombre\ aléatoire\ entre\ 0\ et\ 1$;
            $SYNTHETIQUES[NouvIndex, var] = ORIGINAUX[i, var] + Ecart \cdot Distance$;
        end
        $NouvIndex = NouvIndex + 1$;
    end
end

8.2 Optimisation du CART

K= 5		Taux de graves 1%	10%	30%	50%	AUC moyen
Effectif	50 000	69,5%	68,6%	66,9%	65,6%	67,6%
	150 000	71,3%	67,9%	66,1%	65,6%	67,7%
	250 000	65,1%	67,8%	66,3%	65,7%	66,2%
AUC moyen		68,6%	68,1%	66,4%	65,6%	**67,2%**

K=10		Taux de graves 1%	10%	30%	50%	AUC moyen
Effectif	50 000	69,4%	71,4%	70,8%	65,7%	69,3%
	150 000	72,6%	74,4%	70,1%	66,8%	71,0%
	250 000	72,5%	74,6%	70,1%	64,9%	70,5%
AUC moyen		71,5%	73,5%	70,3%	65,8%	**70,3%**

K=20		Taux de graves 1%	10%	30%	50%	AUC moyen
Effectif	50 000	72,5%	73,7%	69,8%	66,8%	70,7%
	150 000	73,7%	71,6%	71,8%	69,1%	71,5%
	250 000	73,3%	71,4%	71,1%	69,3%	71,3%
AUC moyen		73,2%	72,2%	70,9%	68,4%	**71,2%**

K=30		Taux de graves 1%	10%	30%	50%	AUC moyen
Effectif	50 000	71,6%	73,9%	71,3%	69,2%	71,5%
	150 000	74,1%	73,3%	68,8%	67,8%	71,0%
	250 000	71,7%	72,6%	69,4%	67,7%	70,4%
AUC moyen		72,5%	73,3%	69,8%	68,2%	**71,0%**

FIGURE 8.1 – Sensibilité des CART au ré-échantillonnage : taux, effectif et k-NN
Chaque valeur d'AUC correspond à la moyenne de six AUCs mesurés sur des bases d'apprentissage ré-échantillonnées à partir de différents échantillonnages train/test (6 graines différentes).

K = 10		Taux de graves 1%	10%	AUC moyen
Effectif	50 000	69,4%	71,4%	70,4%
	150 000	72,6%	74,4%	73,5%
	250 000	72,5%	74,6%	73,5%
AUC moyen		71,5%	73,5%	**72,5%**

K = 20		Taux de graves 1%	10%	AUC moyen
Effectif	50 000	72,5%	73,7%	73,1%
	150 000	73,7%	71,6%	72,6%
	250 000	73,3%	71,4%	72,3%
AUC moyen		73,2%	72,2%	**72,7%**

K = 30		Taux de graves 1%	10%	AUC moyen
Effectif	50 000	71,6%	73,9%	72,8%
	150 000	74,1%	73,3%	73,7%
	250 000	71,7%	72,6%	72,2%
AUC moyen		72,5%	73,3%	**72,9%**

FIGURE 8.2 – Sensibilité des CART au ré-échantillonnage : taux, effectif et k-NN- Après exclusion des paramètres peu performants

Chaque valeur d'AUC correspond à la moyenne de six AUCs mesurés sur des bases d'apprentissage ré-échantillonnées à partir de différents échantillonnages train/test (6 graines différentes).

Graine	AUC
490	66,8%
16	70,8%
944	84,9%
584	73,1%
399	78,0%
223	70,0%
AUC moyen	**73,9%**

FIGURE 8.3 – Effet de la réduction de la dimension du suréchantillonnage synthétique sur la performance du CART

8.3 Corrélation des variables

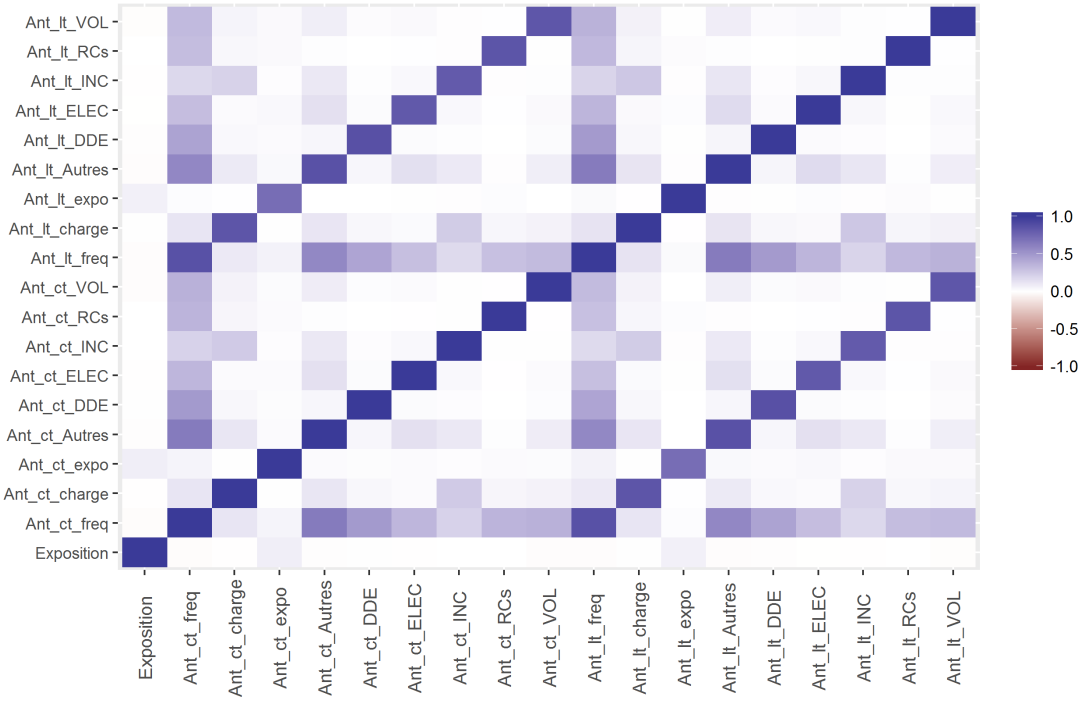


FIGURE 8.4 – Coefficient de corrélation - Variables d'antécédents de sinistres

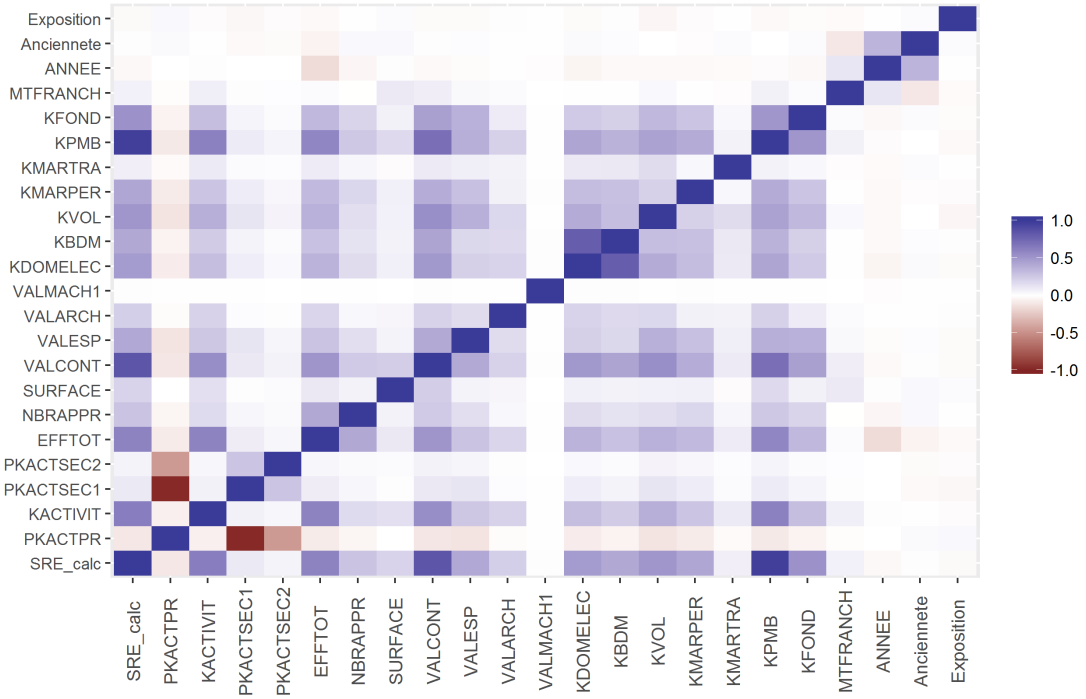


FIGURE 8.5 – Coefficient de corrélation - Variables de souscription