Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame / Monsieur DIALLO Abdourahmane

Titre du mémoire : Mise en relief d'une nouvelle méthode de tarification adaptée aux risques volatiles pour la garantie incendie en risque industriel

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 : Signature :	Entreprise : Nom : GENERALI Signature :
	Directeur de mémoire en entreprise Nom : DEBOUDT Hélène Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

Mémoire d'actuariat

Mise en relief d'une nouvelle méthode de tarification adaptée aux risques volatiles pour la garantie incendie en risque industriel

Réalisé par : Abdourahmane Diallo
Chez Generali

Encadrement
Tuteur académique : Maud Thomas
Tuteur en entreprise : Helene Deboudt

ISUP
03 Novembre 2022

Résumé

Le marché de l'assurance des risques industriels, caractérisé par une forte volatilité induit par les sinistres graves, connaît ces dernières années une dégradation des résultats techniques causée par l'accroissement des sinistres extrêmes. Chez Generali France comme sur le marché, cette situation devient préoccupante et nécessite d'être étudiée afin de comprendre les causes de ces types d'incidents et de proposer une nouvelle méthode de tarification prenant en compte les sinistres graves.

L'incendie constitue la garantie principale de la branche et représente plus de la moitié de la charge globale des sinistres. Dans ce mémoire, nous avons développé une approche permettant de mettre en place un tarif technique de la garantie incendie du risque industriel intégrant la modélisation des sinistres graves. La modélisation des sinistres graves s'est faite à travers la propension de grave, i.e. la probabilité qu'un sinistre survenu conditionnellement aux risques associés soit grave, et la sévérité des sinistres graves conditionnellement aux risques associés. L'avantage de s'intéresser à la queue de distribution des événements extrêmes est la suivante : au lieu de mettre tous les assurés dans un même groupe, ils seront classifiés en groupe homogène en fonction de la lourdeur de la queue de distribution, ce qui permettra de ne pas sur-évaluer la gravité des sinistres sur certains types d'assurés.

La première étape de cette étude consiste à la création de la base de données sur une profondeur d'historique de neuf ans (2013-2021) et aux retraitements des données. Ensuite, un seuil de grave a été déterminé à l'aide de la théorie des valeurs extrêmes. Et enfin, nous avons procédé à la modélisation. Dans cette partie, un modèle de propension, un modèle de fréquence, un modèle de sévérité attritionnelle et un modèle de sévérité grave ont été élaborés. Sur les trois premiers un glm puis un modèle CART et ses extensions (le Random forest et le Gradient boosting) sont ajustés à chaque fois. Ainsi, au regard des indicateurs de performance, la régression linéaire généralisée a été retenue pour la propension et la sévérité attritionnelle et pour la fréquence, la méthode xgboost a été retenue. Pour la modélisation de la sévérité grave, un arbre de régression de Pareto généralisé a été mis en place afin d'estimer dans chaque feuille de l'arbre le paramètre ξ indiquant la lourdeur de la queue et le paramètre d'échelle σ. Ces paramètres aideront à calculer le coût moyen des sinistres pour chaque classe de risque.

Mots-clés : Risques industriels, dommages aux biens, tarif technique, incendie, GLM, CART, Random forest, Gradient Boosting, fréquence, sévérité, propension, théorie des valeurs extrêmes, arbre de régression de Pareto généralisé.

Abstract

The industrial risk insurance market, characterised by a high volatility induced by serious claims, has been experiencing a deterioration in technical results in recent years due to the increase in extreme claims. At Generali France, as in the market, this situation is becoming worrying and needs to be studied in order to understand the causes of these types of incidents and to propose a new pricing method that takes serious claims into account.

Fire is the main coverage in the industry and accounts for over half of the overall claims burden. In this thesis, we have developed an approach to develop a technical tariff for fire cover for industrial risks that incorporates severe loss modelling. The modelling of severe claims was done through the propensity of severe claims, i.e. the probability of a claim occurring conditional on the associated risks being severe, and the severity of severe claims conditional on the associated risks. The advantage of looking at the tail of the distribution of extreme events is that, instead of putting all policyholders in the same group, they will be classified into a homogeneous group according to the severity of the tail of the distribution, which will make it possible not to overestimate the severity of claims on certain types of policyholders.

The first stage of this study consists of creating the database over a nine-year historical depth (2013-2021) and making adjustments to the data. Then, a severe threshold was determined using the extreme value theory. Finally, we proceeded with the modelling. In this part, a propensity model, a frequency model, an attritional severity model and a severe severity model were developed. For the first three, a glm and then a CART model and its extensions (Random forest and Gradient boosting) are fitted each time. Thus, with regard to the performance indicators, generalized linear regression was used for propensity and attritional severity, and for frequency, the xgboost method was used. For the modelling of severe severity, a generalized Pareto regression tree was set up in order to estimate in each leaf of the tree the parameter ξ indicating the heaviness of the tail and the scale parameter σ. These parameters will help to calculate the average cost of claims for each risk class.

Keywords : Industrial risks, property damage, technical tariff, fire, GLM, CART, Random forest, Gradient Boosting, frequency, severity, propensity, extreme value theory, generalized Pareto regression tree.

Note de synthèse

Contexte et objectif

L'assurance des risques industriels est une branche au sein de l'assurance dommages aux biens et responsabilité des professionnels. Elle est caractérisée par une forte volatilité induit par les événements extrêmes, c'est-à-dire les sinistres graves. La rentabilité de cette branche est fortement dépendante des sinistres extrêmes. De ce fait, la gestion et le pilotage du risque doivent d'avantage s'orienter sur la maîtrise et le calibrage de ces types d'incidents. De nos jours, les compagnies d'assurance sont confrontées à une concurrence très forte. Pour cette raison, la tarification et l'attractivité des "bons" clients représentent un défi majeur pour l'assureur. La dégradation des résultats techniques du portefeuille de Generali entraîné par l'augmentation des sinistres graves notamment les sinistres millionnaires (sinistres supérieurs à 1M€) est devenue préoccupante au sein de la compagnie. Ce phénomène de grave est également constaté sur le marché et nécessite une étude pointue permettant de comprendre les facteurs impactant leur réalisation.
L'incendie, qui constitue la garantie principale, représente plus de 55% de la charge globale du portefeuille. Pour cette raison, la calibration de la sinistralité de cette garantie permettra d'améliorer les résultats techniques de la branche. Ainsi, l'objectif de ce mémoire est de mettre en relief une méthode adaptée aux risques volatiles permettant la mise en place d'un tarif technique pour la garantie incendie.

base d'étude

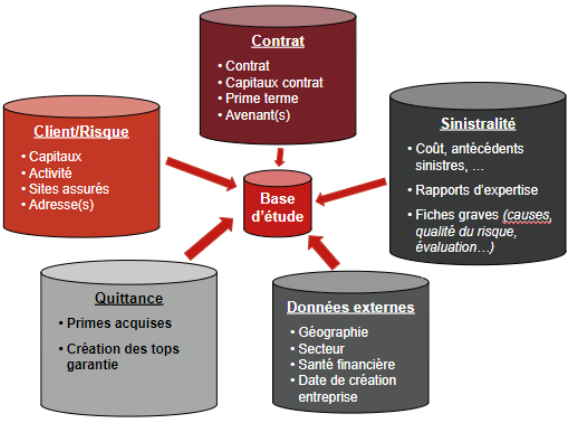

FIGURE 1 – bases de données exploitées

La figure ci-dessus schématise les différentes tables servant à constituer la base d'étude. Cinq bases principales sur une profondeur d'historique de neuf ans (2013-2021) ont été utilisées. Ceci a permis d'avoir un jeu de données de 66 081 lignes avec une ligne par contrat et par année, comportant 9 340 sinistres. Comme les sinistres sont survenus à des années différentes, l'indice RI (Risque Industriel) a été utilisé pour les revaloriser afin de prendre en compte l'inflation. Pour les contrats multi-sites, le sinistre est attribué au site qui a l'engagement maximal car les systèmes d'information de Generali ne permettent pas d'identifier le site sinistré. La qualité des données joue un rôle fondamental sur la fiabilité des résultats obtenus. De ce fait, des méthodes permettant d'imputer les valeurs manquantes sur les variables liées à la taille du risque (engagement, capitaux, surface, etc.) et celles liées aux caractéristiques du risque (Qualité, Zone géographique, etc.) sont appliquées afin de limiter les données manquantes.

Approche de modélisation

La nouvelle méthode permettant de mettre en place un tarif technique adapté aux risques volatiles est la suivante :

$$\pi = E[N|X = x] * [E[Y|Y \leq u; X = x] * (1 - P_G) + E[Y|Y > u; X = x] * P_G]$$

où

- $P_G = P(Y > u|X = x)$ est la probabilité d'avoir un sinistre grave qui sera modélisée par une regression logistique et des méthodes de machine learning
- $E[N|X = x]$ est la fréquence moyenne de sinistre qui sera modélisée par la méthode de glm et des méthodes de machine learning.
- $E[Y|Y \leq u; X = x]$ est le coût moyen des sinistres attritionnels qui sera modélisé par la méthode de glm et des méthodes de machine learning.
- $E[Y|Y > u; X = x]$ est le coût moyen des sinistres graves qui sera modélisé par la méthode d'arbre de régression de Pareto, élaborée par Olivier Lopez, Maud Thomas et Sébastien Farkas.

Pour ce faire, un seuil d'écrêtement des sinistres a été déterminé en premier lieu. Ensuite, pour chaque modèle, une pré-sélection des variables explicatives est faite en effectuant un Random forest afin de garder les vingt variables les plus importantes. Des retraitements sont appliqués sur les variables explicatives retenues des modèles (discrétisation des variables quantitatives, regroupement de modalité des variables qualitatives si nécessaire). Une étude de corrélation a été effectuée sur ces dernières afin de voir s'il y a des dépendances entre elles. Lorsque deux variables sont corrélées, nous sélectionnons la plus pertinente qui permet d'obtenir le modèle le plus performant tout en veillant au sens opérationnel.

Etude du seuil de grave

Quatre méthodes de la théorie des valeurs extrêmes sont utilisées pour le choix du seuil d'écrêtement des sinistres. Sur l'ensemble de ces méthodes les seuils candidats obtenus sont représentés dans le tableau ci-dessous :

Méthodes	seuils
Mean excess plot	150000
Hill plot	[100000,155000]
Pickand's plot	[130000,180000]
Gerstengarbe plot	150000

Étant donné que le seuil u doit être assez grand pour garantir la convergence de la queue de distribution vers une GPD. Et il ne doit pas être trop élevé afin de garder un nombre d'éléments suffisants permettant de réaliser les analyses. Ainsi, la valeur de **u=150 000** a été choisie comme seuil de grave pour la suite de l'étude. Cette valeur est candidate sur l'ensemble des quatre méthodes établies et l'adéquation des données au-delà du seuil à une GPD semble être vérifiée.

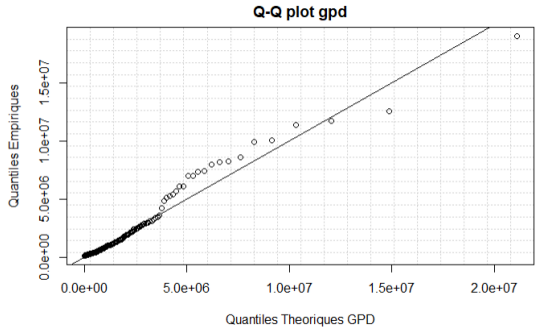

FIGURE 2 – QQ-plot au delà du seuil

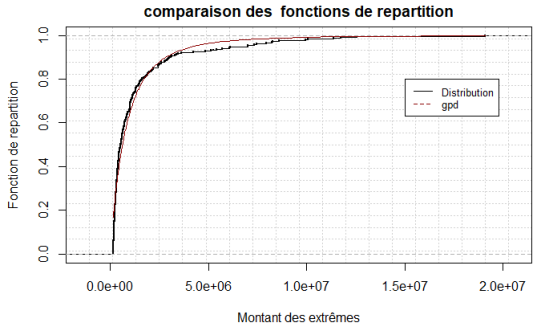

FIGURE 3 – des Fonctions de répartition au delà du seuil

Résultats modélisation

1.Modèle de propension

Afin de modéliser la survenance des sinistres graves, plusieurs méthodes ont été testées dans le but de choisir la plus performante. Aux vues des résultats sur la base test de l'ensemble des modèles établis, la régression logistique est le modèle le plus performant en termes de qualité de prédiction avec un AUC de 76%

Méthodes	AUC	Gini
Régression logistique	76%	52%
CART	62%	25%
Random forest	72%	44%
xgboost	74,5%	49%

2.Modèle de Fréquence

En ajustant un modèle linéaire généralisé avec une loi de poisson et quelques méthodes de machine learning, les résultats ci-dessous sont obtenus sur la base test :

Méthodes	Erreur globale	Gini
glm poisson	-3%	61,4%
CART	3,9%	57,6%
Random forest	1,25%	63,28%
xgboost	-1%	65%

Avec un Gini de 65% et une erreur globale du modèle de -1%, il apparaît clairement que le xgboost est plus performant que les autres méthodes au regard de ces indicateurs. Pour cette raison, il est choisi pour la modélisation de la fréquence du portefeuille.

3.Modèle de Sévérité attritionnelle

Une étude sur le choix de la loi de modélisation entre log-normal et gamma montre que le glm gamma est plus adapté pour la modélisation de la sévérité attritionnelle. Suite à cette étude, nous avons ajusté un glm gamma, un modèle CART et ses extensions. Ainsi la performance de ces méthodes sur la base test est donnée dans le tableau ci-dessous :

Méthodes	Erreur globale	Gini
glm gamma	8%	34%
CART	8,9%	27%
Random forest	7,9%	29%
xgboost	8,5%	30%

La régression gamma est le modèle le plus discriminant avec un Gini de 34%, ce qui signifie qu'il segmente mieux les risques. Par conséquent, cette méthode a été choisie pour la modélisation de la sévérité attritionnelle.

4.Modèle de Sévérité grave

La sévérité des sinistres en risque industriel est très hétérogène, encore plus au niveau de la queue de distribution, cela peut amener à des pertes très différentes. Par conséquent, il est nécessaire de séparer les situations dans le but de ne pas mettre tous les cas de figure dans une même classe. Ceci permettra de voir s'il y a des risques non "assurables" (qui n'admettent pas d'espérance, i.e. $\xi > 1$). Cette approche expose également les assurés susceptibles d'avoir des sinistres très graves. Ainsi, les arbres de régression de Pareto généralisés permettront de déterminer des typologies de classe dont la sévérité des sinistres est homogène. Cette méthode utilise les arbres de régression CART en modifiant le critère permettant de définir les divisions optimales de l'arbre par la perte quadratique de la log-vraissemblance du Pareto généralisé. Dans chacune des feuilles de l'arbre, le paramètre de forme ξ et le paramètre d'échelle σ sont estimés en fonction des facteurs de risque (variables explicatives). Les paramètres σ et ξ estimés par la régression de Pareto généralisée au-delà du seuil u et les intervalles de confiance à 95% sont donnés dans le tableau ci-dessous :

feuilles	ξ	$\sigma 10^{-}2$
feuille1	0.41 [0.01 ;0.8]	6578,74
feuille2	0.06 [0.002 ;0.32]	9942,6
feuille3	0.02 [0.001 ;0.3]	4622,04
feuille4	0.13 [0.02 ;0.3]	6905,09
feuille5	0.17 [0.01 ;0.52]	23978,33
feuille6	0.83 [0.4 ;1.3]	7113,03

Six groupes d'assurés de comportements différents ont été obtenus au niveau de la queue de distribution. Le groupe ayant la queue la plus lourde ($\xi = 0.83$) représente 14% de nos événements extrêmes. Comme $\xi = 0.83 > 0.5$, la variance de ce groupe d'assuré est infinie. De ce fait, il est important de mettre plus de restriction à ces groupes d'assurés afin de calibrer les pertes financières suite à un incident.

En plus de proposer une méthode de tarification, ce mémoire pourrait servir également comme outil de pilotage. En effet, l'arbre de régression de Pareto généralisé a construit des classes d'assurés en fonction de la sévérité des sinistres extrêmes du portefeuille. De plus, la modélisation de la propension de graves estime une probabilité d'avoir un sinistre grave pour chaque assuré. En combinant ces deux études, des classes de risque peuvent être créées afin de définir les risques cibles et les risques à éviter.

Executive summary

Context and objective

Industrial risk insurance is a branch within property and liability insurance for professionals. It is characterised by a high volatility induced by extreme events, i.e. serious claims. The profitability of this branch is highly dependent on extreme claims. As a result, risk management and monitoring must focus more on the control and calibration of these types of incidents. Nowadays, insurance companies are faced with very strong competition. For this reason, pricing and attracting the "right" customers is a major challenge for the insurer. The deterioration of the technical results of Generali's portfolio due to the increase in serious claims, especially millionaire claims (claims above €1M), has become a concern for the company. This phenomenon of serious claims is even observed in the market and requires a detailed study to understand the factors impacting their occurrence.
Fire, which is the main cover, represents more than 55% of the overall portfolio load. For this reason, the calibration of the loss experience of this cover will enable us to improve the technical results of the branch. Thus, the objective of this thesis is to highlight a method adapted to volatile risks allowing the setting up of a technical tariff for the fire cover.

study basis

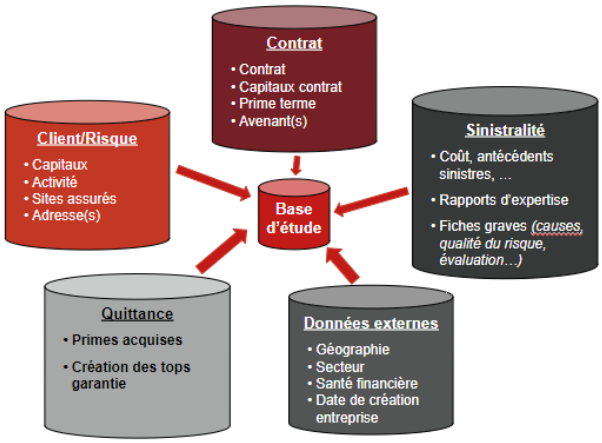

FIGURE 4 – databases used

The figure above shows the different tables used to form the study base. We used five main databases with a history depth of nine years (2013-2021). This gave us a dataset of 66081 rows with one row per contract per year with 9340 claims. As the claims occurred in different years, we used the RI index to inflate them. In the case of multi-site contracts, the claim is allocated to the site with the highest liability because the information systems of Generali do not allow us to identify the location of the claim. The quality of the data plays a fundamental role in the further work, and consequently in the reliability of the results obtained. Therefore, we have implemented methods to impute missing values on variables related to the size of the risk (commitment, capital, surface, etc.) and those related to the characteristics of the risk (Quality, Geographical area, etc.) in order to limit missing data.

Modelling approach

The new method for setting up a technical tariff adapted to volatile risks is the following :

$$\pi = E[N|X = x] * [E[Y|Y \leq u; X = x] * (1 - P_G) + E[Y|Y > u; X = x] * P_G]$$

where

- $P_G = P(Y > u|X = x)$ is the probability of having a serious claim which will be modelled by logistic regression and machine learning methods $E[N|X = x]$ is the average frequency of claims which will be modelled by GLM and machine learning methods.
- $E[Y|Y \leq u; X = x]$ is the average cost of attritional claims which will be modelled by the GLM method and machine learning methods.
- $E[Y|Y > u; X = x]$ is the average cost of severe claims which will be modelled by the Pareto regression tree method developed by Olivier Lopez,Maud Thomas and Sébastien.

To do this, we started by determining the threshold of severity. Then, for each model, we first proceeded to a pre-selection of the explanatory variables by carrying out a random forest and keeping the twenty most important variables, then we reprocessed the explanatory variables (discretization of the quantitative variables, grouping of modality of the qualitative variables if necessary) of the models. A correlation study was carried out on these variables to see if there was any dependence between them. When two variables are correlated, we select the most relevant one which allows us to obtain the best performing model while taking care of the operational meaning.

Severe threshold study

Four methods of extreme value theory are used for the selection of the capping threshold. Of these methods, the candidate thresholds obtained are shown in the table below :

Méthods	threshold
Mean excess plot	150000
Hill plot	[100000,155000]
Pickand's plot	[130000,180000]
Gerstengarbe plot	150000

Given that the threshold u must be large enough to guarantee the convergence of the tail distribution to a GPD but also not too high in order to keep a sufficient number of elements to carry out a study. Thus, we have chosen **u=150000** as a severe threshold for the rest of our work. Indeed, this value is a candidate for all four established methods and the adequacy of the data beyond the threshold to a GPD seems to be verified.

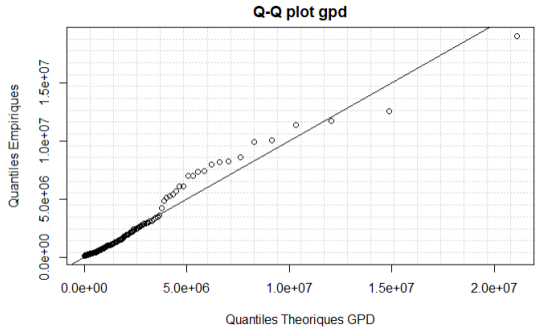

FIGURE 5 – QQ-plot above the threshold

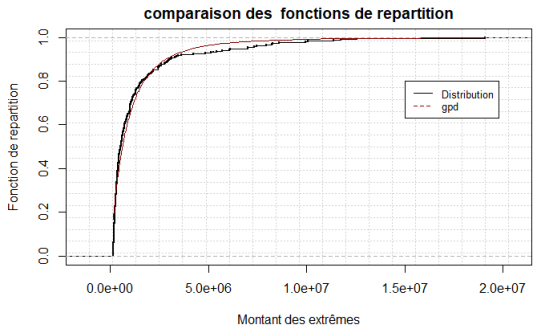

FIGURE 6 – Distribution function above the threshold

Modelling results

1.Propensity model

In order to model the occurrence of serious claims, several methods were tested in order to choose the most efficient one. Based on the results of the test of all the models, logistic regression is the best performing model in terms of prediction quality with an AUC of 76%.

Methods	AUC	Gini
Logistic regression	76%	52%
CART	62%	25%
Random forest	72%	44%
xgboost	74,5%	49%

2.Frequency Model

By fitting a generalized linear model with a Poisson distribution and some machine learning methods, the following results are obtained on the test basis :

Methods	Global Error	Gini
glm	-3%	61,4%
ìline CART	3,9%	57,6%
Random forest	1,25%	63,28%
xgboost	-1%	65%

With a Gini of 65% and a global error of the model of -1%, we can clearly see that the xgboost is more efficient than the other methods regarding these indicators. For this reason, it will be chosen for the modelling of the portfolio frequency.

3. Attritional Severity Model

A study on the choice of the modelling law between log-normal and gamma shows that the gamma glm is better adapted for the modelling of attritional severity. This being done, we have fitted a gamma glm and some machine learning models. Thus the performance of these methods on the test base is given in the table below :

Methods	Global Error	Gini
glm gamma	8%	34%
CART	8,9%	27%
Random forest	7,9%	29%%
xgboost	8,5%	30%

Gamma regression is the most discriminating model with a Gini of 34%, which means that it better segments the risks. Therefore, this method was chosen for modelling attritional severity.

4.Severity Model

The severity of industrial risk claims is very heterogeneous, even more so at the distribution tail, this can lead to very different losses. Therefore, it is necessary to separate the situations in order not to put all cases in the same package. This will allow us to see if there are any "uninsurables" risks (that do not admit of expectation, i.e. $\xi > 1$) in our portfolio but also among the insurable ones which are likely to have very serious losses. Thus, generalized Pareto regression trees will allow us to determine class typologies with homogeneous claim severity. This method uses regression trees by modifying the criterion for defining the optimal tree splits by the quadratic loss of the generalized Pareto log-likelihood. In each leaf of the tree, the shape parameter ξ and the scale parameter σ are estimated as a function of the risk factors (explanatory variables). The parameters σ and ξ estimated by the generalized Pareto regression over the threshold u and the 95% confidence intervals are given in the table below

leaves	ξ	$\sigma 10^{-}2$
leaf1	0.41 [0.01 ;0.8]	6578,74
leaf2	0.06 [0.002 ;0.32]	9942,6
leaf3	0.02 [0.001 ;0.3]	4622,04
leaf4	0.13 [0.02 ;0.3]	6905,09
leaf5	0.17 [0.01 ;0.52]	23978,33
leaf6	0.83 [0.4 ;1.3]	7113,03

We obtain 6 groups of policyholders with different behaviours at the level of the distribution queue. The group with the heaviest tail ($\xi = 0.83$) represents 14% of our extreme events. As $\xi = 0.83 > 0.5$, the variance of this group of insured is infinite. As a result, we will need to put more restrictions on these groups of insureds in order to calibrate the financial losses following an incident.

In addition to proposing a pricing method, this thesis could also serve as a steering tool. Indeed, the generalized Pareto regression tree has constructed classes of insureds according to the severity of extreme claims in the portfolio, and the severe propensity modelling estimates a probability of having a severe claim for each insured. By combining these two studies, we can create risk classes and determine target risks and risks to avoid.

Remerciement

Je tiens tout d'abord à remercier toutes les équipes de la Direction Technique Non Vie de Generali France de m'avoir accueilli plus particulièrement toute l'équipe Dommage aux Biens dont je fais partie.

Je souhaiterais exprimer ma reconnaissance à Jean Sébastien Vieu de m'avoir fait confiance pour me donner l'occasion d'intégrer son équipe, mais aussi sur la qualité de nos échanges et ses conseils au moment de la calibration de mon sujet de mémmoire.

J'adresse également mes remerciements à Hassan Sedraoui et Helene Deboudt de leurs conseils et leurs accompagnements pour la réussite de ce mémoire.
Je remercie aussi à Nicolas Martineau, Océane Lin et Massile Mourah pour leurs disponibilités et leurs soutiens durant cette avanture.

J'aimerais aussi remercier à Maud Thomas pour ses conseils et ses remarques.

En fin, je tiens à remercier ma famille d'être, malgrès la distance, toujours à mes côtés pour me soutenir et m'encourager. votre amour m'a permis de tenir bon tout au long de ces années de dur labeur.

Table des matières

1 Introduction

En assurance, l'assureur fixe le prix de vente de sa prestation, c'est-à-dire la prime, sans pour autant connaître le prix de revient (le montant des sinistres) : c'est le phénomène de **l'inversion du cycle de production**. Par conséquent, l'assureur est exposé aux risques de sous-tarification (ou sur-tarification) ou de dérive de la sinistralité. Sur ce, l'assureur est, en quelques sorte, obligé de prévoir la sinistralité de son portefeuille en utilisant des méthodes statistiques et économétriques afin de garantir la rentabilité de son portefeuille.

Dans un contexte concurrentiel comme celui de l'assurance des risques industriels caractérisée par une forte volatilité due aux sinistres extrêmes (peu en nombre mais qui pèsent fortement en charge), la tarification représente un défi de taille pour l'assureur. Afin d'éviter le risque d'anti-sélection et de faire payer, pour chaque assuré, une prime correspondant à son niveau de risque ; l'assureur doit segmenter aussi finement que possible son portefeuille, dans le but d'établir des classes de risque homogènes. Cependant, une segmentation très fine pourrait contraindre au principe de base de l'assurance : **la mutualisation**. De ce fait, l'assureur doit trouver le juste équilibre pour respecter à la fois le principe de la mutualisation et la segmentation de son portefeuille en groupes homogènes. Ceci permettra d'élaborer des modèles tarifaires qui refléteront au mieux le risque auquel il sera exposé.

L'assurance des risques industriels est l'une des branches de l'assurance Dommage qui nécessite les techniques les plus pointues pour calibrer la sévérité des coûts de sinistres à cause de l'hétérogénéité de la branche. La rentabilité du produit est fortement liée à la sinistralité dite grave. Par conséquent, l'assureur des risques industriels doit prévoir l'occurrence de ces événements extrêmes et d'estimer le coût moyen de ces dernièrs afin de s'assurer des bons résultats de la branche. Toutefois, notons que pour ce type de produit, la quantification des risques est fondamentale mais pas suffisante ; nous avons tout l'aspect de la prévention qui est très important et qui est un facteur non négligeable pour la rentabilité de la branche.

Dans ce mémoire, nous allons mettre en relief une nouvelle méthode de tarification de la garantie incendie en risque industriel qui prendra en compte les sinistres extrêmes et qui permettra d'évaluer « **l'assurabilité** » d'un client pour lui attribuer une prime correspondant à son niveau de risque. Ainsi, ce travail se déroulera en six grandes parties :

1. La construction de la base d'étude et le retraitement des données,

2. L'étude du seuil de grave, fixé actuellement à 150 000 par Generali, par la méthode de la théorie des valeurs extrêmes afin de vérifier la validité de celui utilisé actuellement,
3. La modélisation de la propension de sinistres graves i.e la probabilité qu'un sinistre survenu conditionnelement aux risques associés soit grave,
4. La modélisation de la fréquence de sinistre,
5. La modélisation des coûts attritionnels,
6. La modélisation des coûts graves

Dans chaque étape, nous élaborerons plusieurs méthodes de modélisation et nous les comparerons afin de choisir le meilleur modèle en termes de performance et de précision.

Utilisant jusqu'à présent une méthode de tarification basée sur l'expertise et l'analyse du souscripteur mais aussi sur les taux de prime marché fourni par la FFA (Fédération Française de l'assurance), ce mémoire va permettre à **Generali** à la fois d'avoir un tarif technique plus adapté à son portefeuille mais aussi d'améliorer leur politique de souscription.

2 L'assurance des Risques industriels

2.1 Description de la branche

L'assurance des Risques industriels est une branche au sein de l'assurance dommages aux biens et responsabilité civile qui permet aux entreprises de se couvrir contre les dommages susceptibles d'atteindre tous leurs biens mobiliers et immobiliers et les conséquences financières suite à une interruption d'activité ou de la responsabilité causé à des tiers qu'ils s'agissent des dégâts matériels, corporels ou immatériels. Ce produit d'assurance a été créé vers la fin du dix septième siècle suite aux violents incendies qui avaient ravagé le centre de la ville de Londres en 1666, et couvrait principalement les industries. Aujourd'hui, l'assurance des risques industriels est devenue, en d'autres termes, l'assurance des risques d'entreprise. Ces dernières font face à de nombreux risques qui peuvent endommager leurs biens et mettre en péril leurs activités de façon permanente. De ce fait, l'assurance des risques industriels est importante pour les entreprises afin qu'elles mènent sereinement leurs activités. Elle propose aux entreprises plusieurs garanties :

- Incendie ou explosion, qui couvre les dommages directement causés aux bâtiments, aux équipements (matériels, installation,etc.) et aux marchandises suite à la survenance de ces événements.
- Perte d'exploitation, qui couvre les pertes financières dues à l'interruption de l'activité de l'entreprise après un sinistre.
- Responsabilité civile, qui couvre les dommages matériels, immatériels, corporels causés par l'entreprise à une tierce personne
- Vol et vandalisme, qui couvre tout dommage ou perte causé par ces événements.
- Dommage électrique, qui couvre les dommages matériels suite à une surtension électrique
- Bris de machine, dégât des eaux, événements climatiques, catastrophes naturelles couvrant les dégâts entraînés par ces événements.

Les risques industriels (RI) communément appelés **Risques relevant du Traité des Risques d'Entreprise (TRE)** ne considèrent que les entreprises ayant un capital supérieur à 152 fois l'indice RI et dont l'activité est présente dans le TRE. Dans ce traité, chaque entreprise est attribuée à un code TRE composé de trois chiffres et chaque code TRE appartient à une famille appelée fascicule correspondant au premier chiffre du code TRE. Les fascicules sont au nombre de 10 allant de 0 à 9 correspondant aux activités suivantes :

- Fascicule 0 : Extraction et préparation de minerais et minéraux divers, de combustibles minéraux solides, Métallurgie,

- Fascicule 1 : Production de matériaux de construction, Industrie des céramiques, Industries du verre,
- Fascicule 2 : Travail des métaux, Industries électriques et électroniques, Construction automobile, aéronautique et navale, Carrosserie et réparation de véhicules, Garages et stations-service
- Fascicule 3 : Industries chimiques et parachimiques, Transformation de matières plastiques et de caoutchouc,
- Fascicule 4 : Industries textiles, Bonneterie, Confection de vêtements et autres articles textiles,
- Fascicule 5 : Industries du papier et du carton , Imprimeries , Industries du cuir et du délainage,
- Fascicule 6 : Industries du bois,
- Fascicule 7 : Industries agro-alimentaires,
- Fascicule 8 : Traitement des déchets urbains et industriels , Production et distribution d'énergie,
- Fascicule 9 : Autres risques d'entreprises.

2.2 Le Marché des risques industriels

Le marché de l'assurance des risques industriels a connu ces dernières années une dégradation de son ratio sinistres à primes. Ce phénomène est dû à l'accroissement de la fréquence des sinistres extrêmes dans la branche. Cependant, une forte augmentation des cotisations (7% injecté entre 2015 et 2021 et 1,8% entre 2020 et 2021) et la stabilisation du nombre de sinistres supérieurs à 10M€ ont contribués à l'amélioration des résultats de l'année 2021.

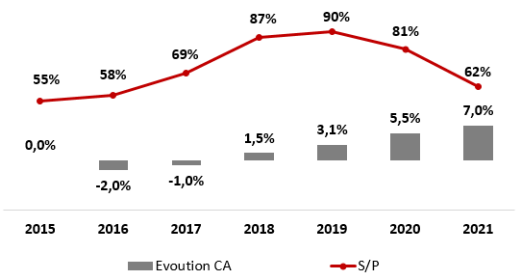

FIGURE 7 – Evolution des cotisations et ratio sinistres à primes entre 2015 et 2021, données FFA

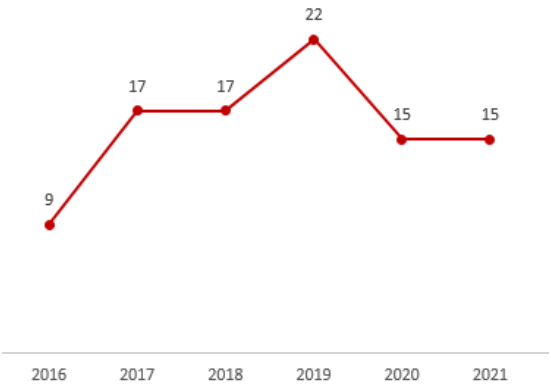

FIGURE 8 – Evolution du nombre de sinistre supérieur à 10M€ entre 2016 et 2021,données FFA

Dans l'assurance de biens des professionnels et agricoles, l'assurance des risques industriels représente près de 27% des cotisations en 2021. Ainsi, il apparaît clairement que la branche apporte des primes importantes aux sociétés d'assurance permettant d'augmenter rapidement leurs flux futurs. De ce fait, la branche joue un rôle très important dans l'activité d'assurance. Une maîtrise de la volatilité importante de la branche engendrée par les sinistres d'intensité pourrait faire de l'assurance des risques industriels la branche la plus attirante aux sociétés d'assurance à cause des primes importantes qu'elle génère. Cette maîtrise pourrait se faire par l'augmentation de la prévention, mais aussi par une gestion des risques avec des techniques pointues.

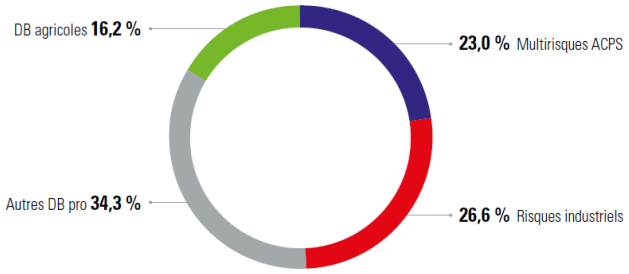

FIGURE 9 – Répartition des cotisations en DAB PRO 2021 et agricole, FFA

2.3 Le groupe Generali et Generali France

Avec la volonté de créer une grande compagnie capable de rivaliser avec les assureurs naissant en Lombardie-Vénétie ou dans le reste de l'Europe, en 1831 en Italie, les «Assicurazioni Generali Austro-Italiche» ont donné naissance aux premières assurances généralistes. C'est la première compagnie d'assurance à avoir proposé à ses clients une couverture multirisque, d'où son nom Generali.

Un an après la création du groupe, la première agence française de Generali voit le jour à Bordeaux. Il s'agit de la plus ancienne implantation étrangère du groupe et marque le début d'une histoire riche de créations, d'évolutions, d'acquisitions et de fusions. Generali est aujourd'hui le troisième assureur européen, dont le premier en vie et l'un des dix premiers groupes d'assurance au monde. Cette présence mondiale s'appuie sur une implantation directe dans plus de soixante pays ainsi que sur de nombreux partenariats qui permettent au Groupe d'opérer

dans soixante-dix pays sur les cinq continents et plus spécifiquement en Europe, en Asie et en Amérique latine. La France occupe la troisième place derrière l'Italie et l'Allemagne dans le marché du groupe. Generali France dispose de 7 000 collaborateurs et de plus de 8 millions de clients. Elle propose des services en assurance Dommages, Épargne , Retraite, Prévoyance et santé etc.

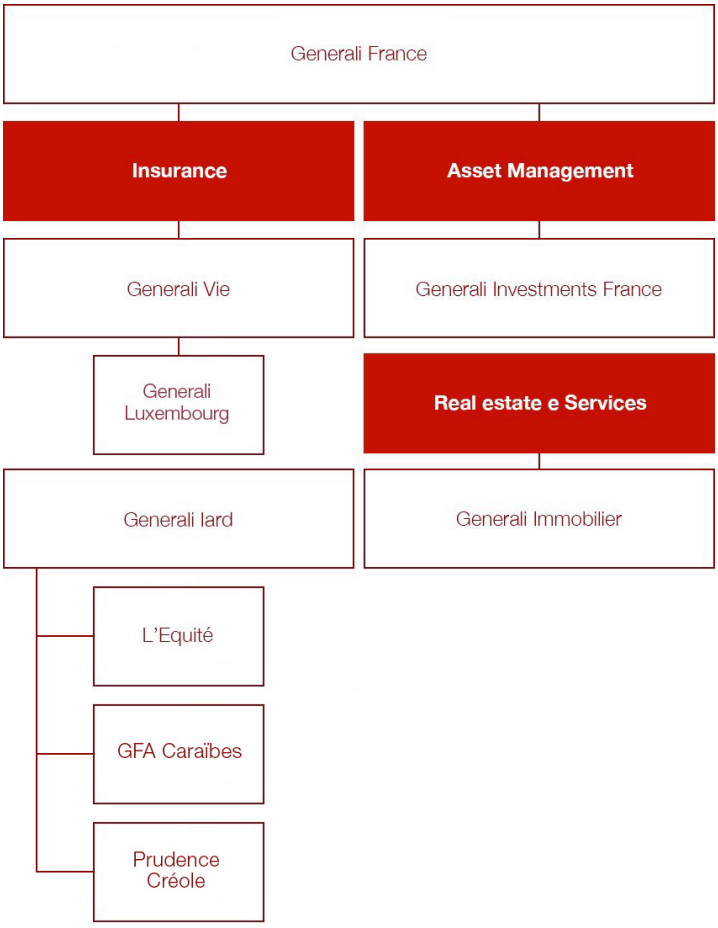

FIGURE 10 – Organigramme Generali France

Le service Dommages et Responsabilités des Professionnels est rattaché à la direction Technique Assurance non-vie. Cette dernière comprend quatre services à savoir : Études indemnisation, Études Produit, Auto & Dommages aux Biens des Particuliers et Dommages & Responsabilités des Professionnels. Cette dernière dispose de huit branches principales : Risque Industriel, Risque Technique, Responsabilité Civile, Construction, Professionnels de l'Auto, Agricole, Transport de marchandise hors plaisance et Cyber risque.

2.4 Les risques industriels chez Generali

L'assurance des risques industriels chez Generali est constituée des Entreprises Industrielles et Commerciales (hors garages et assimilés) dont le chiffre d'affaires est inférieur à trois cents millions d'euros et supérieur à cent soixante fois l'indice RI. Les autres entités dont le chiffre d'affaires est supérieur à trois cents millions d'euros sont gérées par le service "Generali global corporate" et celles ayant un chiffre d'affaires inférieur à cent soixante fois l'indice RI sont gérées par l'équipe multirisque commerce (MRC). Par conséquent, elles sont exclues dans le périmètre de notre étude.

La garantie principale de cette offre d'assurance est l'incendie, c'est-à-dire aucune entreprise ne pourra souscrire un contrat sans la garantie incendie. Les activités sont classées, selon leurs niveaux de risque, en quatre groupes principaux que sont :

- Les cibles, qui constituent les activités les moins risquées sur lesquelles Generali souhaite se développer,
- Les autorisés, constituent les activités ayant un niveau de risque non négligeable et que Generali souhaite se développer,
- Les résérvés, qui constituent les activités dont la souscription nécessite beaucoup d'analyses et de contrôles,
- Les exclus, constituent les activités dont leurs souscriptions sont restreintes.

Chaque contrat est attribué à un code TRE correspondant à son activité, une franchise qui correspond à un seuil de montant en dessous duquel l'assureur n'indemnise pas, une LCI (Limitation Contractuelle Indemnité) c'est-à-dire le montant maximal que l'assureur indemnisera en cas de sinistre et une estimation du Sinistre Maximum Possible (SMP).

Le tarif technique appliqué, jusqu'à présent, repose sur l'analyse des souscripteurs et sur les principes du Traité d'Assurance Incendie Risques d'Entreprise dommages directs mis en place par la FFA. Cette dernière utilise la théorie de la crédibilité hiérarchique de JEWELL à trois niveaux (niveau portefeuille, niveau fascicule et niveau code tre) pour élaborer un taux de prime pure par activité.
Or, le portefeuille risques industriel de Generali a évolué au cours de ces dernières années, de ce fait, la refonte du tarif technique devient primordiale.

3 Problématique et objectif de notre étude

Le portefeuille risque industriel de Generali France a connu ces dernières années une forte dégradation de ses résultats à cause d'une augmentation de la survenance des sinistres graves (ie les sinistres qui dépassent 150 000, seuil fixé par Generali) notamment les sinistres millionnaires. Ce fléau est constaté, même, au niveau marché et nécessite d'être étudié avec des techniques très pointues. Sur les quatre dernières années, la charge des sinistres graves représente 1% en nombre et près de 60% de la charge globale dans notre portefeuille. De ce fait, il apparaît clairement que la rentabilité de la branche est fortement dépendante des sinistres graves et il est primordial de trouver et de calibrer les facteurs de risque impactant leurs survenances afin d'améliorer notre critère tarifaire et de garantir la rentabilité de la branche. Le graphe ci-dessous montre l'évolution des résultats entre 2013 et 2021 sur la partie attritionnelle et grave.

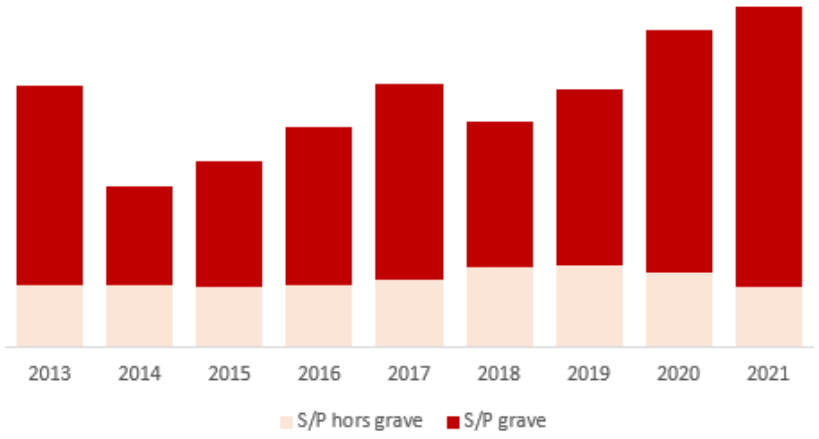

FIGURE 11 – Evolution des résultats entre 2013 et 2021

On voit clairement une nette dégradation du ratio sinistres à primes du portefeuille notamment les graves entre 2018 et 2021 avec une tendance linéaire.
Avec l'évolution des méthodes statistiques notamment les apprentissages machine (« machine learning ») et le nombre de données massives disponibles, il est important d'exploiter ces outils tout en prenant en compte leurs limites et leurs sensibilités afin d'essayer de capter ou de comprendre les facteurs de risque entrainant la survenance des sinistres extrêmes et de mettre en place un tarif technique adapté à ces types d'incidents.

Ainsi dans ce mémoire, nous allons essayer de mettre en place le modèle suivant :

$$\pi = E[N|X = x] * [E[Y|Y \leq u; X = x] * (1 - P_G) + E[Y|Y > u; X = x] * P_G]$$

où

- $P_G = P(Y > u|X = x)$ est la probabilité d'avoir un sinistre grave qui sera modélisée par une regression logistique et des méthodes de machine learning
- $E[N|X = x]$ est la fréquence moyenne de sinistre qui sera modélisée par la méthode de glm et des méthodes de machine learning.
- $E[Y|Y \leq u; X = x]$ est le coût moyen de sinistre attritionnels qui sera modélisé par la méthode de glm et des méthodes de machine learning.
- $E[Y|Y > u; X = x]$ est le coût moyen des sinistres graves qui sera modélisé par la méthode d'arbre de régression de Pareto généralisé élaborée par Olivier Lopez, Maud Thomas et Sébastien Farkas dans leur étude de sévérité du risque cyber. Cette méthode permettra de distinguer ce qui est « assurable » ou non en fonction de la valeur du parmétre de queu de distribution « tail index ».

En effet, la prime pure en assurance est, par définition, donnée par la formule suivante :

$$\begin{aligned}
\pi &= E[N|X = x] * E[Y|X = x] \textit{ avec } N \textit{ independant de } Y \\
&= E[N|X = x] * E[Y_{attri} + Y_{grave}|; X = x] \\
&= E[N|X = x] * [E[Y|Y \leq u; X = x] * (1 - P_G) + E[Y|Y > u; X = x] + *P_G]
\end{aligned}$$

Cette méthode aidera à instaurer un tarif technique du risque industriel pour la garantie incendie qui, pour la première fois, tiendra compte à la fois la probabilité de survenance des sinistres graves, mais aussi des facteurs de risque indiquant « l'assurabilité » d'un client et le coût moyen de la charge attendu en cas de sinistre grave.

L'objectif est de mettre en exergue une nouvelle méthode de tarification plus adéquate aux risques volatiles.

4 Présentation et analyse des données

Notre étude portera sur les contrats qui ont vécu au moins un jour dans le portefeuille d'assurance risque industriel de Generali France entre le 1^{er} janvier 2013 et le 31 décembre 2021. Ainsi, dans cette partie, une présentation des données qui serviront à mener notre étude et la façon dont elles sont collectées seront effectuées en premier lieu. Ensuite, nous intéresserons aux traitements des données manquantes. Et enfin, quelques statistiques descriptives seront faites sur certaines variables.

4.1 Les bases de données

4.1.1 Données Sinistres

La base sinistre est composée de l'ensemble des sinistres survenus entre 2013 et 2021 pour la garantie incendie. Elle est constituée d'une ligne par sinistre et par date de survenance. Un sinistre est principalement décrit par les données suivantes :

- **le numéro de sinistre**
- **la date de survenance du sinistre**.
- **L'état du sinistre** (En cours, Clos, Sans suite, Annulé, Reouvert, etc..)
- **Le réglement**
- **La provision**
- Le **recours**

Ainsi, le montant du sinistre à charge de l'assureur est obtenu par la formule suivante :

$$Coût\ sinistre = Règlement + Provision - Recours$$

Comme les sinistres sont survenus à des années différentes, donc pour prendre en compte l'inflation, les charges des sinistres ont été revalorisées en fonction de leur date de survenance avec une mise en "AS if". On a utilisé pour cela l'indice RI fourni par la FFA. Cet indice permet de prendre en compte l'évolution des prix relatifs aux bâtiments, aux matériels, aux marchandises et à la main d'œuvre.

Les systèmes d'information de Generali ne permettent pas d'identifier le site sinistré. Cependant, étant donné qu'un contrat de risque industriel peut avoir plusieurs sites, nous attribuerons le sinistre au site qui détient l'engagement le plus élevé. En faisant cette hypothèse, notre étude expliquera le comportement global du client et non le site industriel sinistré. La problématique d'identification du site

sinistré nous empêchera d'effectuer une étude plus fine. Néanmoins, en prenant cette hypothèse, nous parviendrons à capter les informations les plus importantes du client. Dans ce mémoire, seul les sinistres clos ont été pris avec l'hypothèse qu'il n'y aura pas de réouverture et donc par ricochet les sinistres n'évolueront pas dans le temps. De ce fait, grâce à cette hypothèse, le vieillissement de sinistre ne sera pas effectué. Ainsi, sur la profondeur d'historique de neuf ans, 9 340 sinistres incendies sont recueillis au total.

4.1.2 Données contrats

La base contrat est composée des différentes caractéristiques ou informations sur la police. Elle est constituée d'une ligne par contrat et par année assuré. Toutes les informations principales des contrats qui ont été assuré au moins un jour dans l'exercice considéré sont récupérées. Par exemple, considérons l'année 2020, tous les contrats dont la date de résiliation est supérieure au 1^{er} janvier 2020 et inférieure au 31 décembre 2020 seront pris. Voici quelques exemples de variable que nous avons conservées :

- **La prime terme**
- **Le capital dommage directe du contrat**
- **L'engagement du contrat**
- **Le nombre de site**
- **le reseau de distribution**

4.1.3 Données risques par site industriel

Cette base permet d'avoir la description plus détaillée du risque par site industriel. Elle est constituée d'une ligne par site. Les données risques sont les plus importantes pour la quantification du risque industriel. Un site est principalement décrit par les variables suivantes :

- **Le code activité**
- **Type de site** (principale/secondaire)
- **Le capital du Contenu**
- **Le capital du Bâtiment**
- **Engagement**
- **La surface du bâtiment et le Prix par mettre carré**
- **La qualité de l'occupant**(Propriété occuppant, Locataires, etc.)
- **L'effectif du site**
- **visite avant souscription**(oui/non)
- **etc..**

4.1.4 Variables créées

Quelques variables ont été créées notamment sur les antécédents de sinistres, mais aussi la présence d'une garantie autre que l'incendie pour le contrat. Cette détection de garantie est faite de la façon suivante : si une prime est acquise dans l'année pour un contrat sur une garantie donnée, nous considérerons que cette dernière est acquise dans l'année. Voici la liste des variables crées :

- l'antécédent sinistre des trois dernières années
- TOP_RC : indiquant si la garantie RC est souscrite (oui/Non)
- TOP_PE : indiquant si la garantie PE est souscrite (oui/Non)
- TOP_BDM : indiquant si la garantie BDM est souscrite (oui/Non)
- TOP_DEL : indiquant si la garantie DEL est souscrite (oui/Non)
- TOP_VOL : indiquant si la garantie VOL est souscrite (oui/Non)

4.1.5 Les données rapports d'expertises et fiches graves

A Generali, les sinistres dont le montant dépasse 300 000€,en risque industriel, sont délégués à une structure spécifique qui établit un bilan détaillé du sinistre sur un rapport appelé **rapport d'expertise/Fiche grave**. Ainsi, étant donné que les informations présentent sur ces rapports ont été rédigées en texte libre, plusieurs techniques de type « text minning[1] »ont été déployées afin d'isoler un certain nombre de mots-clés pour créer des tops suivants :

- **Le stockage**,
- **L'électricité**,
- **Perte contrôle du Process**,
- **La combustibilité**,
- **L'entretien**,
- **L'équipement d'intervention** ,
- **La ventilation**,
- **La propreté**,

Ainsi, pour chaque contrat présent dans le portefeuille, un taux d'apparition des tops est calculé sur l'historique des sinistres survenus les trois années précédents l'année d'observation. Par exemple, pour l'électricité, ce top est déclenché à partir du moment où l'un des mots-clés est apparu : panne, coupure, disjoncteur, batterie, interrupteur, surchauffe, conduction, électrique, multiprise, électricité, court circuit. Ensuite, un taux d'apparition de ce top est calculé à la mannière

1. algorithme de traitement de texte permettant de transformer un document non structuré en données structurées.

d'une fréquence, soit le rapport entre le nombre de sinistres sur les trois ans ayant déclanché le top et l'exposition du contrat sur les trois ans.

4.1.6 Données externes

Quelques informations externes qui pourraient expliquer la sinistralité de notre portefeuille notamment les sinistres extrêmes ont été collectées. Parmi ces informations, nous avons :

- **La distance aux pompiers :** cet indicateur permet de nous fournir la distance entre le site et le campement des pompiers le plus proche. Ça peut-être un facteur important pour évaluer le niveau de risque d'un site pour l'incendie. En effet, si le site est proche d'une station de pompier alors, en cas d'incendie, les pompiers pourront vite intervenir pour limiter les dégâts
- **La santé financière de l'entreprise :** cette information est importante en risque industriel du fait du poids important que pèse la prévention et de la sécurité pour cette branche. En effet, il est tout à fait légitime de croire qu'une entreprise dans des situations financières difficiles aura des difficultés à mettre les moyens nécessaires sur la prévention et la sécurité.
- **La date de création de l'entreprise**
- **La catégorie de l'entreprise (PME[2]/ ETI[3]/GE[4])**
- **La région**

4.1.7 Qualité des données et traitement des valeurs manquantes

La présence de données manquantes dans les bases est un fléau auquel on est souvent confronté notamment en assurance non-vie. De ce fait, il y a plusieurs façons de traiter ces données manquantes :

1. La suppression des lignes contenant des données non renseignées.
2. Remplacer les données manquantes par une valeur qui peut être la **la moyenne, la médiane, le minimum, le maximum ou par zéro**
3. Pour les variables catégorielles, remplacer les valeurs manquantes par la modalité la plus courante
4. Remplacer les données manquantes en utilisant des algorithmes de machine Learning notamment les k-plus proches voisins. l'idée est la suivante : considérons un jeu de donnée contenant n lignes $(X_i)_{1\leq i\leq n}$ et p variables

2. Petites et Moyennes Entreprises.
3. Entreprises de Tailles Intermédiaires.
4. Grandes Entreprises.

$(X_j)_{1 \leq j \leq p}$. Ainsi, pour chaque individu x ayant des valeurs non renseignées, on recherche les k-individus les plus proches en minimisant la quantité $||x - X_i||$ où $||.||$ est une norme définie sur R^p, puis on remplace la donnée manquante par la moyenne de ces k-individus.

Toutefois, notons que toutes ces méthodes présentent des avantages et des inconvénients, car il est pratiquement impossible de retrouver l'information manquante de façon exacte. En effet, prenons l'exemple de la suppression des lignes contenant des valeurs manquantes, l'avantage est qu'on travaillera sur des données sûres et fiables, mais le problème sera une perte d'information qui peut être considérable dans le cas où on n'a pas beaucoup de ligne dans notre jeu de données ; c'est le cas en risque industriel. Donc nous n'utiliserons pas cette technique pour gérer les valeurs manquantes. Tout d'abord, nous allons présenter la qualité de nos données et ensuite, nous expliciterons les méthodes de rehaussement adoptées pour imputer nos valeurs manquantes.

4.1.8 La qualité des données

Le jeu de données constitué contient à la fois des variables qualitatives et quantitatives. Les figures suivantes schématisent la qualité de chaque variable présente dans la base d'étude.

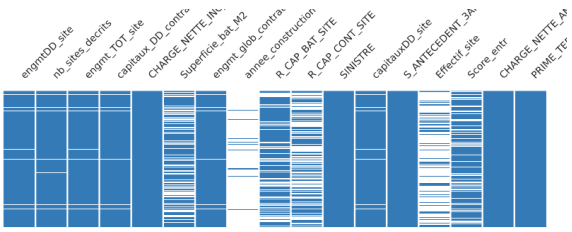

FIGURE 12 – qualité des variables quantitatives

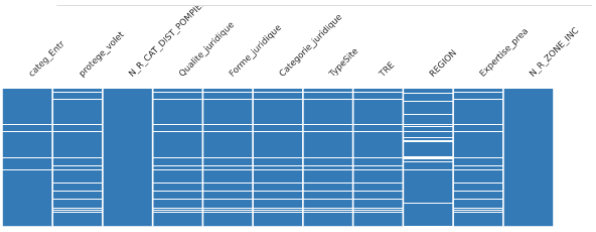

FIGURE 13 – qualité des variables qualitatives

Cette figure s'interprète de la façon suivante : si la barre est totalement remplie, alors la variable ne présente pas de donnée manquante sinon elle contient des données manquantes. De ce fait, on constate qu'il y a des valeurs manquantes aussi bien dans nos variables numériques que dans nos variables catégorielles. La présence de données non renseignées n'est pas identique d'une variable à une autre. Par exemple, on s'aperçoit que la variable **année de construction** est celle qui a un taux de renseignement le plus faible parmi toutes les autres variables. Elle a un pourcentage de valeurs manquantes de 90%. De ce fait, cette variable sera retirée dans la base de modélisation.

Ensuite, le graphique montre que les données permettant d'évaluer la taille du risque notamment **les Capitaux, les engagements, la superficie du bâtiment, etc.** contiennent des valeurs manquantes. Ces dernières constituent des informations importantes pour le bon déroulement de l'étude. De ce fait, nous verrons par la suite la procédure adoptée pour renseigner ces valeurs manquantes.

4.1.9 Traitement des valeurs manquantes

Pour les variables qualitatives, la méthode d'imputation des données maquantes retenues est la suivante :

1. Si l'information est présente à l'année n mais absente à l'année n+k alors la valeur manquante sera remplacée par la modalité renseignée à l'année n.
2. Sinon la donnée manquante sera affectée à la modalité la plus représentative de la variable.

Pour les variables quantitatives notamment celles liées à l'évaluation de la taille du risque, une régression par activité (TRE) en utilisant la prime est appliquée pour renseigner les valeurs manquantes. En effet, la prime est, en général, proportionnelle à la taille du risque et à l'activité en risque industriel. Le schéma ci-dessous l'illustre bien.

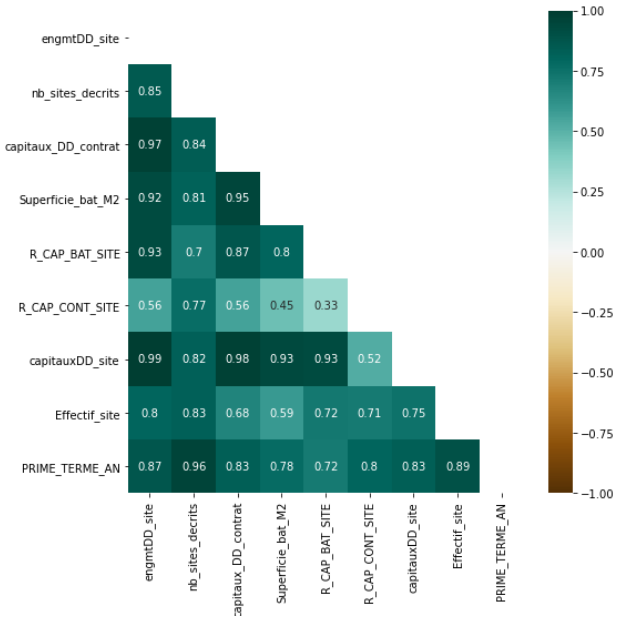

FIGURE 14 – corrélation par activité

De ce fait, nous calculons d'abord un taux moyen par activité par la formule suivante : considérons l'individu x non renseigné de la variable X, P sa prime annuelle et N_k le nombre d'activités k dans notre jeu de données.

$$\tau = \frac{\sum_{i=1}^{N_k} X_i}{\sum_{i=1}^{N_k} P_i}$$

Ensuite la valeur est calculée par la formule suivante : $x = \tau * P$

4.2 Synthèse

Cette partie nous a permis de construire notre jeu de donnée pour la modélisation. Il est constitué des données sinistres, risques, contrats, rapports d'expertise/fiches de grave, et des données externes. La base ainsi construite contient 66 081 lignes avec une ligne par contrat et par année de vision. Pour les contrats multi-sites, le sinistre est attribué au site qui a l'engagement maximal à cause du problème de la non-identification du site sinistré. Par la suite, nous avons vérifié la qualité des données. Ainsi, pour les variables contenant des valeurs manquantes, des méthodes permettant de les traiter sont élaborées.

5 Analyse Descriptive du portefeuille

Dans cette section, quelques analyses exploratoires sont effectuées sur la base de données de modélisation. Cette étape permettra d'avoir, d'une part, une vision d'ensemble de la sinistralité, et d'autre part, l'état de santé du portefeuille sur les 9 ans (2013-2021).

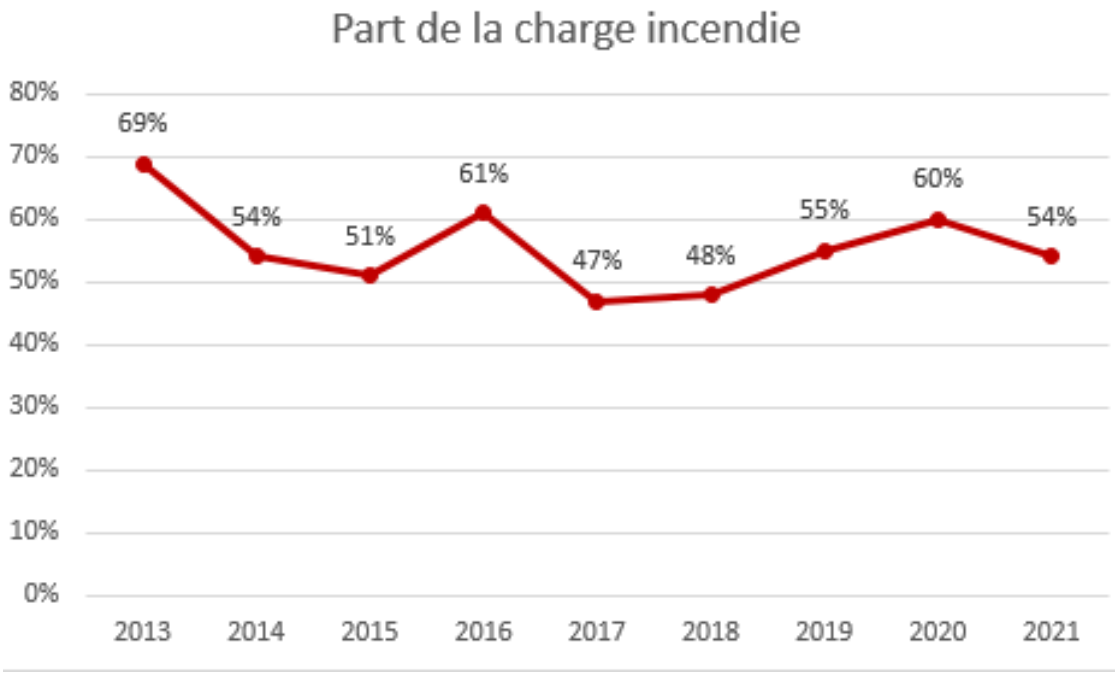

FIGURE 15 – part de la charge incendie

Cette figure montre l'évolution de la part de la charge incendie entre 2013 et 2021, c'est-à-dire le rapport entre la charge totale des sinistres incendie et l'ensemble des sinistres du portefeuille survenus dans l'année. Cet indicateur nous a permis de voir le poids de cette garantie au sein de la branche. De ce fait, on constate qu'en moyenne, sur les 9 ans, la charge des sinistres incendie représente 55,4% de la charge totale. La part maximale est constatée en 2013 avec 69% et la part minimale en 2017 avec 47%. Sur ce, la maîtrise ou le calibrage de la sinistralité incendie permettra d'améliorer les résultats de la branche.

Commençons par regarder l'évolution du Taux de Destruction et du Taux de Prime de l'incendie dans le portefeuille. Ces indicateurs sont définis comme suit :

$$TD = \frac{Charge\ du\ sinistre}{Engagement}$$

$$TP = \frac{Primeacquise}{Engagement}$$

Un Taux de Destruction (TD) égale à 0,01 veut dire qu'on a indemnisé 1 centime pour 1 euro assuré et un Taux de Prime (TP) égale à 0,01 signifie qu'on reçoit 1 centime pour 1 euro assuré. Donc un portefeuille en bonne santé doit avoir un TD inférieur au TP

.

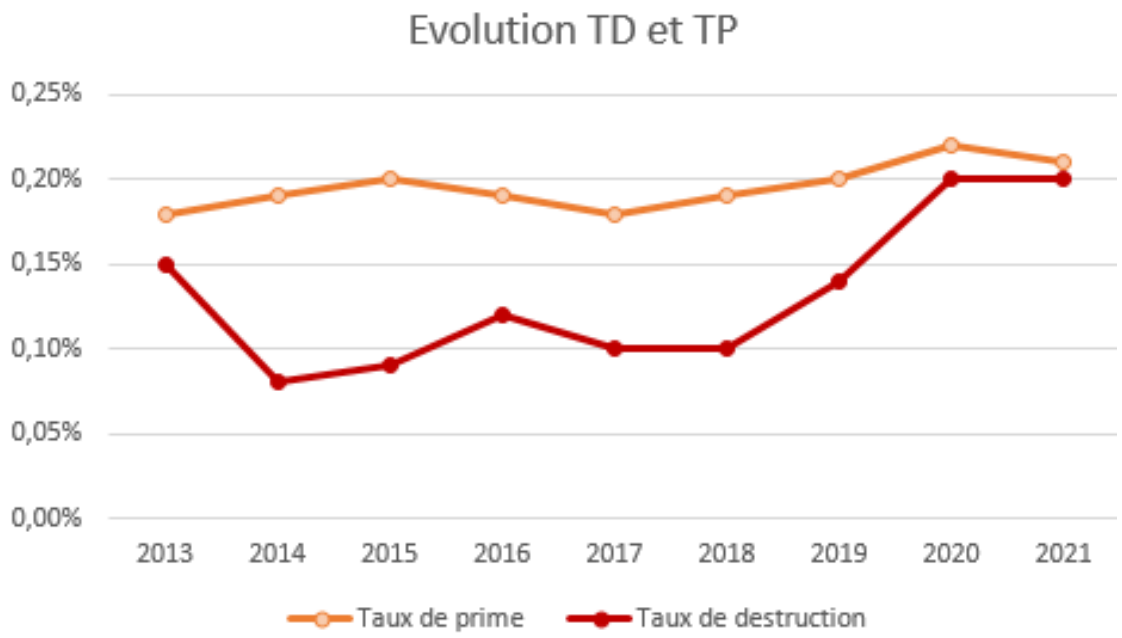

FIGURE 16 – Evolution du taux de prime et du taux de destruction entre 2013 et 2021 du portefeuille

On observe une évolution croissante du Taux de Destruction sur les 5 dernières années. Il converge vers le Taux de Prime, ce qui est un mauvais signe pour la santé de la branche. Cette convergence est due d'une part à la survenance des sinistres graves ces dernières années, mais aussi à une faible augmentation du Taux de Prime comparé au Taux de Destruction.

Par la suite, nous allons faire un zoom sur quelques variables explicatives afin d'avoir quelques idées précises de la dégradation du portefeuille.

Fascicules

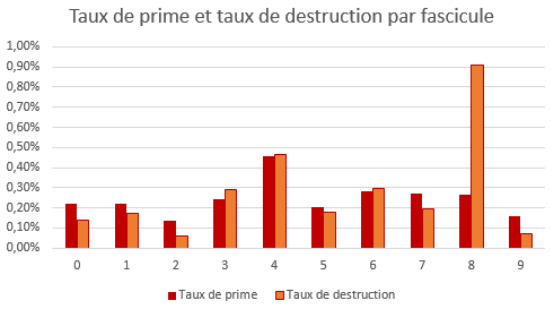

FIGURE 17 – Taux de destruction et Taux de prime par fascicule

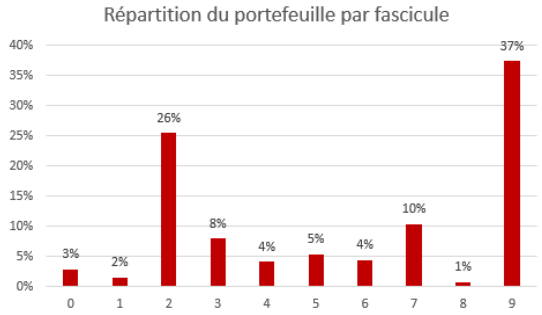

FIGURE 18 – Répartition du portefeuille par fascicule

On constate que le fascicule 8 (traitement des déchets urbains et industriels, production et distribution d'énergie) à un taux de destruction largement supérieur au taux de prime et ne représente que 1% dans le portefeuille. La non-rentabilité de cette famille d'activité peut être expliquée non seulement par le niveau de risque élevé du secteur, mais aussi par un manque d'expérience (historique) profonde. Ce serait intéressant de voir dans notre modélisation quelles sont les activités les plus risquées dans le fascicule 8 et les moins risquées afin de développer ces dernières dans notre portefeuille. Par contre les fascicules 2 (Travail des métaux, Industries électriques et électroniques, Construction automobile, aéronautique et navale, Carrosserie et réparation de véhicules, Garages et stations-services) et 9 (Autres risques d'entreprises) représentent 63% du portefeuille et ont un taux de destruction faible. Ce sont des activités dont Generali a un historique profond . Cette expérience profonde du secteur peut justifier le bon calibrage et la rentabilité du secteur. Le fascicule 0 (Extraction et préparation de minerais et minéraux divers, de combustibles minéraux solides, Métallurgie) qui n'est présent que 3% dans le portefeuille alors qu'il a un bon Taux de destruction et un bon Taux de prime. De ce fait, ce serait important de regarder si nous pouvons développer cette famille d'activité dans le futur.

Qualité juridique

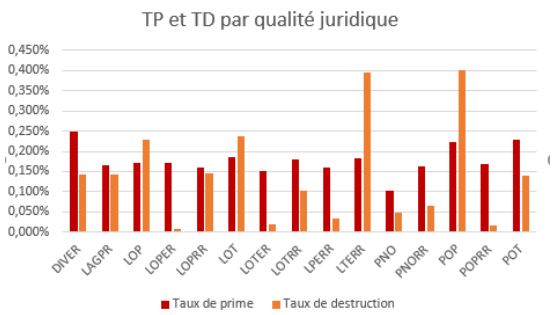

FIGURE 19 – Taux de destruction et Taux de prime par Qualité juridique

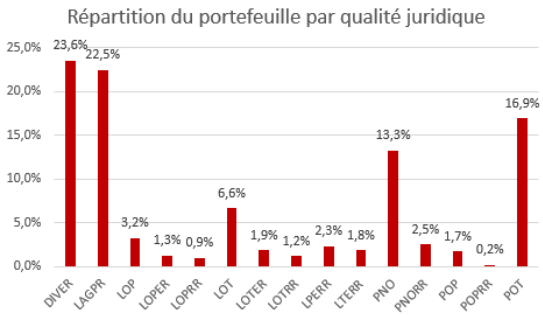

FIGURE 20 – Répartition du portefeuille par Qualité juridique

Les sites qui ont comme qualité POP(Propriétaire occupant partiel) et LTERR (Locataire occupant total, exonéré de ses risques locatifs, avec renonciation à recours) représentent 3,5% du portefeuille et ont un taux de destruction très élevé. Par contre, les POPRR (Propriétaire occupant partiel, avec renonciation à recours contre le locataire), LOTER (Locataire occupant total, exonéré de ses risques locatifs) et LOPER (Locataire occupant partiel, exonéré de ses risques locatifs) ont un taux de destruction très faible et un bon taux de prime et ils représentent 5,3% du portefeuille. Les remarques établies sur les fascicules seront valables aussi sur la qualité juridique et une attention spécifique pourrait être faite sur la modélisation.

Zone géographique

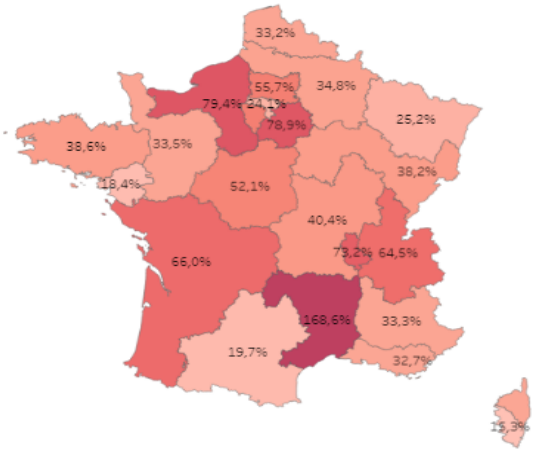

FIGURE 21 – SP par zone geographique

Cette carte schématise le ratio sinistre à prime du portefeuille par zone géographique. Et on constate que les régions d'occitanies et du hauts-de-France ont un ratio dégradé par rapport au reste.

5.1 Synthèse

Cette analyse exploratoire a permis d'avoir une meilleure idée de la composition de notre jeu de données, mais aussi une tendance de la sinistralité du portefeuille. En effet, on a constaté que la charge totale des sinistres incendies représente en moyenne un peu plus de 55% du coût global de notre base de données et que le taux de destruction ne cesse de se dégrader année en année depuis 2017 alors que le taux de prime n'a pas augmenté dans les mêmes proportions dans un contexte de marché fortement concurentiel. Cette situation explique, en partie, le souci de la rentabilité de la branche. Ce constat nous a conduit à faire un zoom sur quelques variables explicatives notamment les fascicules, la qualité juridique et la zone géographique afin de voir les facteurs de risque dégradants de ces dernières.

6 Détermination du seuil de grave

La détermination du seuil de grave consiste à trouver un montant à partir duquel un sinistre est considéré comme étant grave à l'aide des outils mathématiques notamment **la théorie des valeurs extrêmes**. Cette théorie a été développée par Fisher et Tippet (1928) puis par Gnedenko (1943) mais a connue un essor fulgurant à partir des années cinquante. Cette branche s'intéresse à l'étude du comportement asymptotique de la queue de distribution des données et se repose sur des propriétés des statistiques d'ordre et des méthodes d'extrapolation.
cette étape est primordiale pour la suite de notre travail, car elle permettra de déterminer les sinistres attritionnelles et graves afin de les modéliser séparément.
Dans cette partie, nous allons d'abord faire un bref rappel de la théorie mathématique des valeurs extrêmes et ses résultats et ensuite déterminer notre seuil de grave.

6.1 Théorie des valeurs extrêmes

6.1.1 Comportement asymptotique du maximum

Soient $X_1, X_2, ..., X_n$ un échantillon aléatoire indépendant et identiquement distribué (iid) d'une variable aléatoire X de fonction de répartition F. On note le maximum de cette échantillon par M_n définie comme suit :

$$M_n = max\left\{X_1, X_2, ..., X_n\right\}$$

Par indépendance, nous obtenons pour toute réalisation x la fonction de répartition :

$$\begin{aligned} F_n(x) &= P(M_n \leq x) \\ &= P(X_1 \leq x, ...X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) \\ &= \prod_{i=1}^{n} F(x) \\ &= (F(x))^n \end{aligned}$$

Ce résultat montre que si nous connaissons la loi de X alors nous pouvons déduire la loi du maximum. Cependant, on ne connaît pas, en général, la distribution de X. Cette difficulté nous conduit à regarder le comportement asymptotique du maximum afin de chercher la famille de loi vers laquelle M_n converge.

6.1.2 Distribution généralisée des valeurs extrêmes

Définition 1 : Domaine d'attraction
On dit qu'une distribution F appartient au domaine d'attraction de H_ξ et on note $F \in D(H_\xi)$, s'il existe des suites réelles $(a_n) > 0$ et $b_n \in R$ telles que :

$$\lim_{n\to+\infty} F^n(a_n(x) + b_n) = H_\xi(x)$$

Théorème 1 : Fisher-Tippet-Gnedenko
soit $(X_n)_{n>=1}$ une suite de n variables aléatoires *iid* de loi de probabilité F telle que $F(x) = P(X <= x)$ et $M_n = max\{X_1, X_2, ..., X_n\}$. S'il existe une suite $a_n > 0$ et une suite $b_n \in R\ \forall n >= 1$ et une loi non dégénérée G telle que :

$$\begin{aligned} \lim_{n\to+\infty} P(\tfrac{M_n - b_n}{a_n}) &= \lim_{n\to+\infty} F^n(a_n(x) + b_n) \\ &= G(x)\ \forall x \in R \end{aligned}$$

alors G est définie comme suit :

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

où μ est un paramètre de position, ξ le paramètre de forme ou l'indice de queue et σ le paramètre d'échelle. En fonction des valeurs que prennent ξ, G peut appartenir à l'une des trois types de loi suivante :

Loi de Gumbel ($\xi = 0$) :

$$\Lambda_{\mu,\sigma} = \exp(-(\exp(\frac{x-\mu}{\sigma})))$$

Les lois appartenant au domaine d'attraction de Goumbel sont appelées des lois à queue légère, car ils présentent une décroissante exponentielle au niveau de la queue de distribution. Parmi ces lois, nous pouvons citer entre autre la loi **Normal,Gamma,Log-Normale,Exponentielle, etc..**

Loi de Fréchet ($\xi > 0$) :

$$\Phi_{\mu,\sigma,\xi} = \begin{cases} \exp(-[1+\xi(\frac{x-\mu}{\sigma})]^{\frac{-1}{\xi}}) & si\ x\ >\ \mu \\ 0 & sinon \end{cases}$$

Les lois appartenant au domaine d'attraction de Fréchet sont appelées des lois à queue lourde car ils présentent une décroissante lente au niveau de la queue de distribution. Parmi ces lois, nous pouvons citer entre autre la loi **Cauchy, Pareto, Student, Log-Gamma etc..**

Loi de Weibull ($\xi < 0$) :

$$\Phi_{\mu,\sigma,\xi} = \begin{cases} \exp(-[1+\xi(\frac{x-\mu}{\sigma})]^{\frac{1}{\xi}}) & si\ x\ <\ \mu \\ 1 & sinon \end{cases}$$

Les lois appartenant au domaine d'attraction de Weibull sont appelées des lois à queue fine et sont bornées à droite. Parmi ces lois, nous pouvons citer entre autre la loi **Uniforme, Beta, etc..**

Propriété 1 : une distribution généralisée des valeurs extrêmes possède :

1. une esperance finie si et seulement si $\xi\ < 1$
2. une variance finie si et seulement si $\xi\ <\ \frac{1}{2}$
3. un moment d'ordre k si et seulement si $\xi < \frac{1}{k}$

Ce graphe ci-dessous nous montre la structure de la queue des distributions des lois de **Goumbel, Weibull et Fréchet**

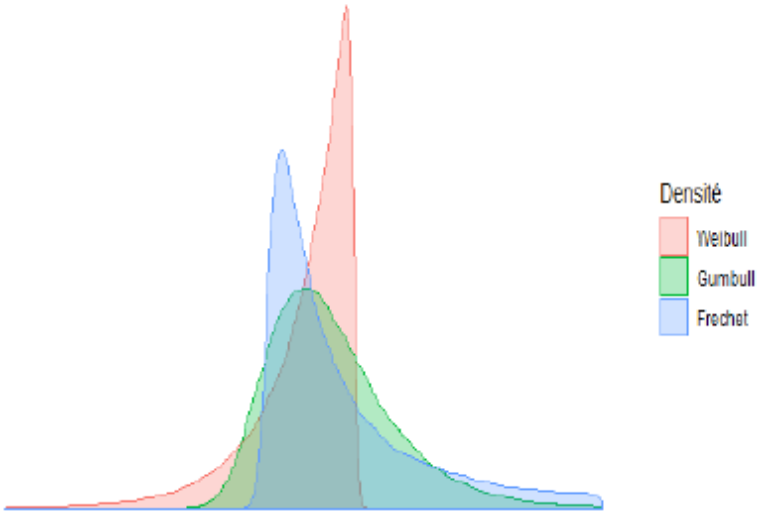

FIGURE 22 – Densité Goumbel, Weibull et Fréchet

6.1.3 Distribution au-delà du seuil

Théorème 2
Soient X une variable aléatoire de fonction de répartition F et u le seuil de grave. Si F appartient au domaine d'attraction d'une distribution généralisé des valeurs extrême (GEV) alors la variable aléatoire $Y = X - u|X > u$ suit une distribution de Paréto généralisé $GPD(\sigma, \xi)$. Par conséquent,

$$P(X-u|X>u) \overrightarrow{u \to x_F} \begin{cases} [1+(\frac{\xi x}{\sigma})]_+^{\frac{-1}{\xi}} & si\ \xi \neq 0 \\ \exp(\frac{-x}{\sigma}) & si\ \xi = 0 \end{cases}$$

où

— $\sigma > 0$ est une paramètre d'échelle.
— $\xi \in R$ est une paramètre d'échelle

de plus, si $0 < \xi < 1$ alors la fonction moyenne des excès est donnée par la formule ci-dessous u

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

Remarque1 : discution sur la fonction moyenne des excès
Par la définition de la la fonction moyenne des excès $e(u)$ on a :

1. si $\xi > 1$ alors $e(u)$ n'est pas définie car son espérence n'est pas finie
2. si $\xi = 1$ alors $e(u)$ est contante et est égale à $\frac{\sigma}{1-\xi}$
3. si $0 < \xi < 1$ alors $e(u)$ est une fonction affine linéaire en u

6.2 Méthodes de détermination du seuil

6.2.1 Quantile-Quantile plot (QQ-plot)

Le graphe des quantiles ou QQ-plot a pour objectif d'approcher et de comparer une distribution théorique à une distribution empirique. Ce graphique est obtenu en traçant le nuage de point :

$$(X_{(i)}, F^{-1}(1 - \frac{i}{n}))_{1 \leq i \leq n}$$

où $X_{(i)}$ est la $i - ème$ statistique d'ordre de l'chantillon aléatoire $(X_1, X_2, ..., X_n)$

Ainsi, si les données sont adéquates à la distribution théorique alors les points seront alignés sur la droite d'équation $y = x$ appelée la première bissectrice.
Cette étape nous permettra d'avoir une idée sur la lourdeur de notre queue de distribution de nos données. En effet, en prenant comme distribution théorique la loi exponentielle de paramètre $\lambda = \frac{1}{E[X]}$ et en traçant le nuage de point suivant :

$$(X_{(i)}, F^{-1}\frac{1}{\lambda} ln(\frac{i}{n}))_{1 \leq i \leq n}$$

on a les propriétés suivantes :

— si le graphique est concave alors nos données ont une distribution à queue lourde : on est dans le domaine d'attraction de **Fréchet** ($\xi > 0$)
— si le graphique est convexe alors nos données ont une distribution à queue légère : on est dans le domaine d'attraction de **Goumbel** ($\xi < 0$)
— si le graphique est aligné à la prémière bissectrice alors nos données sont adéquates à la loi exponentielle ($\xi = 0$)

Aplication
Ci-dessous, nous comparons la distribution théorique exponentielle à la distribution empirique de la charge des sinistres :

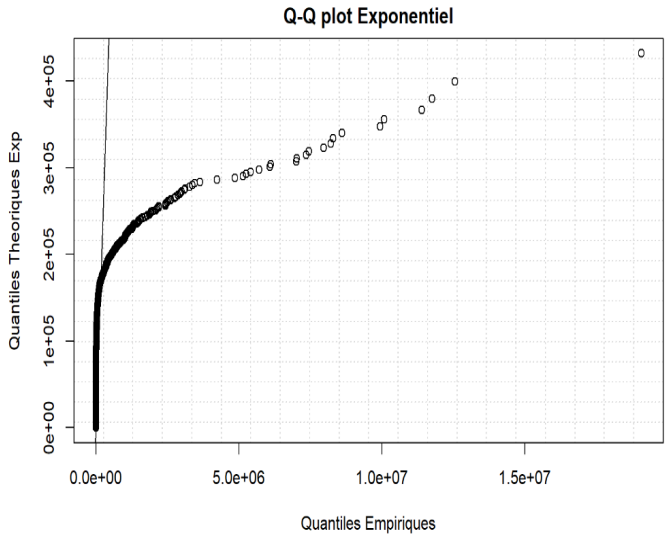

FIGURE 23 – Q-Qplot Charge sinistre

On constate que les points ne sont pas alignés à la droite d'équation $y = x$, donc la distribution de nos charges de sinistre n'est pas adéquate à la loi exponentielle. De plus on voit que le graphe est concave, ce qui signifie que la queue de distribution des charges est plus épaisse que celle de l'exponentielle. Donc on est dans le domaine d'attraction de **Fréchet**.

6.2.2 La Fonction moyenne des excès

La fonction moyenne des excès ou mean excess plot est obtenue en traçant le nuage de point suivant :

$$\{(u, \hat{e_n}(u)),\ x_{(1)} < u < x_{(n)}\}$$

où

$$\hat{e_n}(u) = \frac{\sum_{i=1}^{n}(x_i - u)1_{x_{(i)}>u}(x_{(i)})}{\sum_{i=1}^{n} 1_{x_{(i)}>u}(x_{(i)})},\ l'estimateur\ de\ e(u)$$

et

$x_{(1)}$, $x_{(n)}$ sont respectivement le minimum et le maximum de l'échantillon aléatoire $(X_1, X_2, ..., X_n)$.

Pour déterminer le seuil u, on choisit celui à partir duquel la courbe est approximativement linéaire.

Application

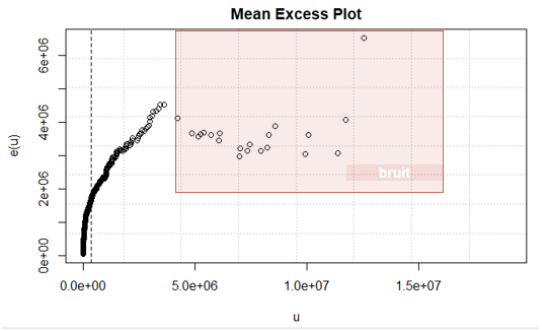

FIGURE 24 – graphe mean exess fonction, zone de recherche seuil

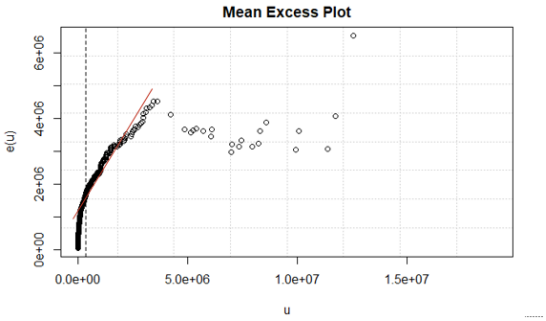

FIGURE 25 – graphe mean exess fonction, seuil retenu

Sur le graphe de la Figure 24, on a défini une zone qui représente du bruit, c'est-à-dire la zone où il n'est pas pertinent de recherché notre seuil de grave car d'une part on n'a pas assez de points dans cette zone et d'autre part nous n'observons pas une tendance linaire des points. De ce fait, nous rechercherons le seuil dans la zone non encadrée. Cela étant dit, on voit sur la Figure 25 une tendance linéaire à partir de **150000**. Donc la valeur $u = 150000$ a été choisie comme notre seuil de grave pour le mean exess plot.

6.2.3 Méthode de Hill

La méthode QQ_plot effectuée au niveau de la section 6.2.1 nous a permis de déterminer l'épaisseur de la queue de distribution de nos charges. À ce stade, nous avions constaté que nous sommes dans le domaine d'attraction de Fréchet($\xi > 0$). Donc, il est légitime de regarder la méthode de Hill.

Théorème 4 : Stabilité des GPD
Si une variable aléatoire $X \sim GPD(\sigma, \xi)$ alors $X - u|X > u \sim GPD(\tilde{\sigma}, \xi)$: les lois GPD sont stables par troncature à gauche.

La méthode de Hill ou Hill-plot est une méthode graphique fondée sur la propriété de stabilité d'une loi GPD permettant d'obtenir un seuil.

Le Hill-plot trace en fonction de u l'estimateur de Hill $\tilde{\xi}$ de ξ défini comme suit :

$$\tilde{\xi_k} = \frac{1}{k}\sum_{i=1}^{n} log(\frac{X_{(i)}}{X_{(k+1)}})$$

Par la suite, pour trouver le seuil u, on recherche la zone pour laquelle la courbe est sur un plateau.

Théorème 5

Pour tout $k, n > 1$ avec $1 \leq k < n$ tel que $k \to \infty$ et $\frac{k}{n} \to 0$ quand n est assez grand $(n \to +\infty)$ alors l'estimateur de Hill vérifie les propriétés suivantes :

— $\tilde{\xi}$ est consistent, *i.e* $\tilde{\xi}$ converge en probabilité vers ξ

— $\tilde{\xi}$ est assymptotiquement normal, *i.e* $\sqrt{k}(\tilde{\xi} - \xi) \sim N(0, \xi^2)$

Application

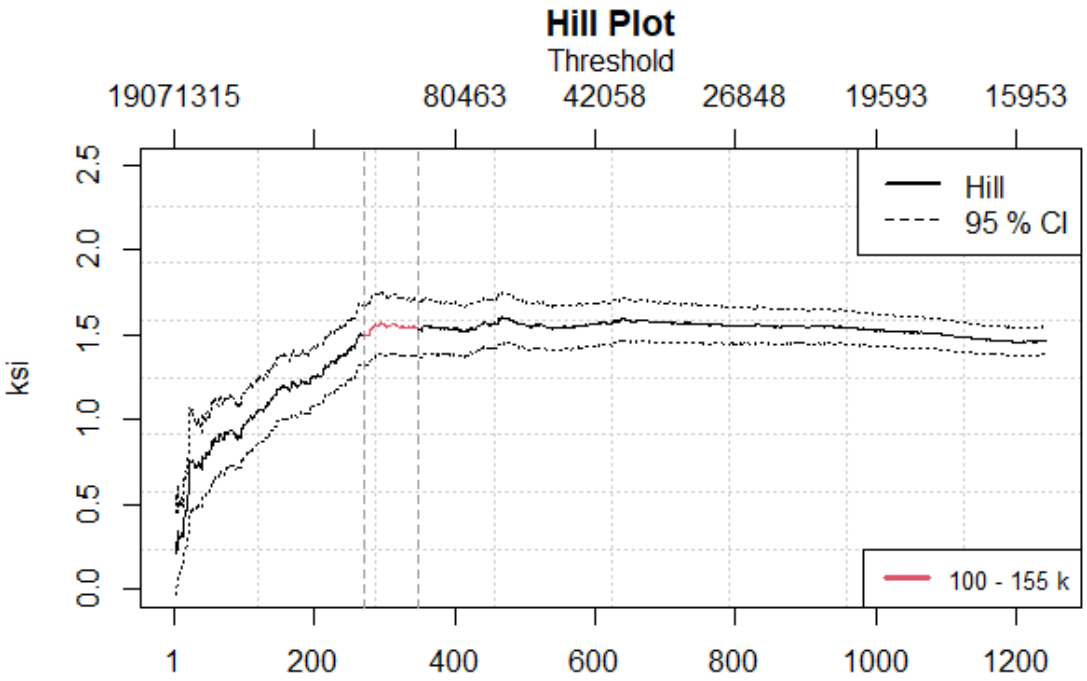

FIGURE 26 – Hill plot

On observe sur ce graphe qu'un seuil candidat est dans l'intervalle [100000, 155000] car il semble y avoir une stabilité dans cette zone.

6.2.4 Méthode de Pickands

Comme pour la méthode de Hill, la méthode de Pickands s'appuie sur la propriété de stabilité de la loi GPD.
L'estimateur de pickands n'est pas très robuste, car il est très sensible à la taille de l'échantillon. En revanche, son avantage est qu'il est valable quelque soit la valeur de ξ et par conséquent, quelque soit le domaine d'attraction de la distribution. Il est basé sur le calcul des quantiles et est défini par la formule suivante :

$$\hat{\xi}^{P}_{k_n} = \frac{1}{log(2)} log(\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}) \ , \ k_n = 1, ..., \lfloor \frac{n}{4} \rfloor une \ suite \ d'entier$$

Théorème 6

Soit k_n une suite d'entiers avec $1 \leq k_n < n$ tel que $k_n \to \infty$ et $\frac{k_n}{n} \to 0$ quand n est assez grand $(n \to +\infty)$ alors l'estimateur de Pickands vérifie les propriétés suivantes :

- $\hat{\xi}^{P}_{k_n}$ est consistent, *i.e* $\hat{\xi}^{P}_{k_n}$ converge en probabilité vers ξ
- sous certains condition sur la suite k_n et sur la fonction de répartition F $\hat{\xi}^{P}_{k_n}$ est assymptotiquement normal, *i.e* $\sqrt{k_n}(\hat{\xi}^{P}_{k_n} - \xi) \sim N(0, \frac{\xi^2(2^{2\xi+1}+1)}{4(log(2))^2(2^{\xi}-1)^2})$

Application

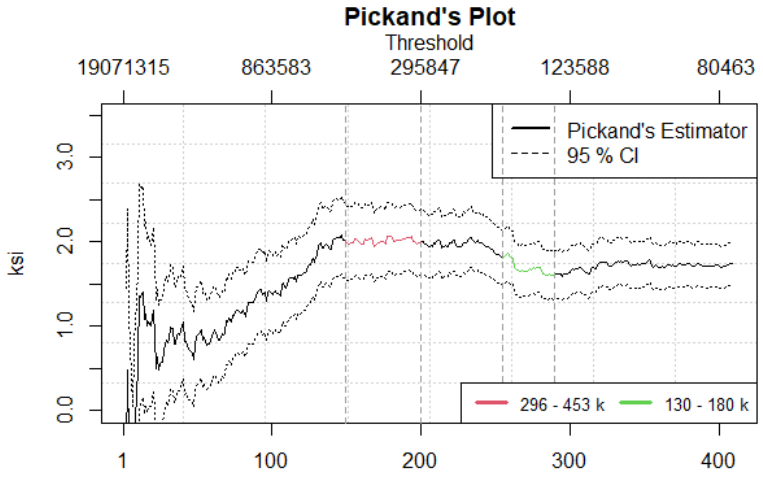

FIGURE 27 – pickand's plot

En observant le graphe, plusieurs seuils peut être candidats, car il y a différentes zones où la courbe semble être stable notamment entre 296 000 et 463 000 mais aussi entre 130 000 et 18 0000. De ce fait, deux personnes différentes pourraient choisir deux zones diférentes par conséquent deux seuils différents. Dans notre cas, pour des raisons de volume de données au-delà du seuil, nous optons à choisir notre seuil dans l'intervalle [130 000,180 000] .

6.2.5 Méthode de Gerstengarbe

La méthode de Gerstengarbe est une procédure qui permet de déterminer un seuil en se basant sur la statistique de Mann-Kendall. Elle consiste à tracer deux séries et à identifier un point de changement dans le comportement des écarts consécutifs entre les coûts. Ce changement de comportement indique le point de départ de la zone extrême. Cette méthode permet à la fois de déterminer le point de départ de la région extrême, mais aussi de donner une estimation du seuil optimal. Pour $i = 1, ..., n-1$ on construit les deux séries de différence comme suit :

$$H_k = \frac{\sum_{i=1}^{k} n_i \ - \ a_k}{b_k} \quad ; \quad \tilde{H_k} = \frac{\sum_{i=1}^{k} \tilde{n_i} \ - \ a_k}{b_k}$$

où

$$n_k = \sum_{i=1}^{k-1} 1_{\Delta_i < \Delta_k} \quad ; \quad \tilde{n_k} = \sum_{i=1}^{k-1} 1_{\tilde{\Delta_i} < \tilde{\Delta_k}}$$

$$a_k = \frac{k(k-1)}{4} \quad ; \quad b_k = \sqrt{\frac{k(k-1)(2k+5)}{72}}$$

$$\Delta_i \ = \ X_{(i+1)} \ - \ X_{(i)} \quad ; \quad \tilde{\Delta}_i \ = \ X_{(n-i)} \ - \ X_{(n-i+1)}$$

Pour trouver le seuil, on trace H_k et $-\tilde{H}_k$ puis on choisi la valeur correspondant au k associé au point d'intersection des deux courbes. Pour obtenir un seuil adéquat, il est recommandé de répéter cette procédure plusieurs fois, car la méthode de Gerstengarbe n'est pas très stable.

Application

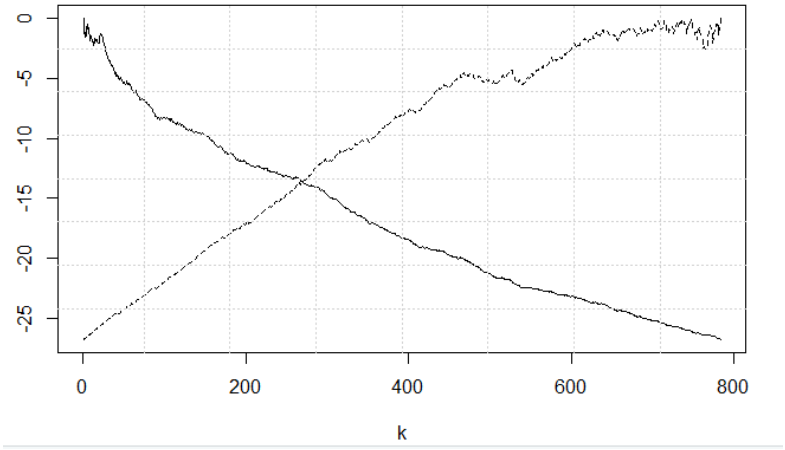

FIGURE 28 – Gerstengarbe plot

Ce graphe correspond au Gerstengarbe plot pour trois itérations sur la charge. Le k associé au point d'intersection des deux courbes est 265, ce qui correspond à un seuil de 150 000.

6.3 Choix définitif du seuil de sinistre grave

Sur l'ensemble des quatre méthodes utilisées, les seuils candidats obtenus sont les suivantes :

Méthodes	seuils
Mean excess plot	150 000
Hill plot	[100 000,155 000]
Pickand's plot	[130 000,180 000]
Gerstengarbe plot	150 000

On sait que la convergence de la queue de distribution vers une GPD se fait pour un seuil u assez grand. Cependant, le choix du seuil ne doit pas aussi être trop élevé afin de garder un nombre d'éléments suffisants permettant de faire la modélisation. Cela étant dit, on a constaté que le seuil de **150 000** est candidat sur l'ensemble

des quatre méthodes. De ce fait, il a été choisi comme seuil de grave pour la suite du travail.

6.3.1 Adéquation d'une GPD au-delà du seuil

Étant donné que le choix d'un seuil ne se repose que sur des méthodes purement graphique, ce seuil n'est pas unique. De ce fait, pour conforter notre choix, nous allons vérifier l'adéquation des données au-delà du seuil à une GPD. Ainsi, nous comparons la fonction de répartition empirique de nos observations et la fonction de répartition théorique d'une GPD, mais aussi nous représenterons le graphe quantile-quantile.

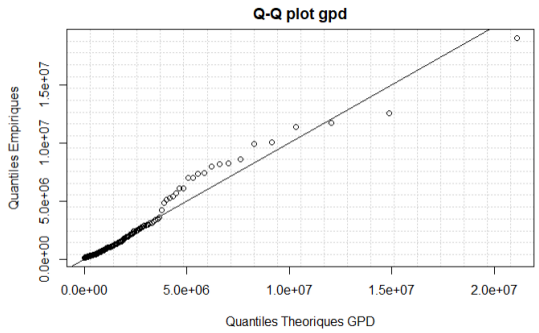

FIGURE 29 – QQ-plot au delà du seuil

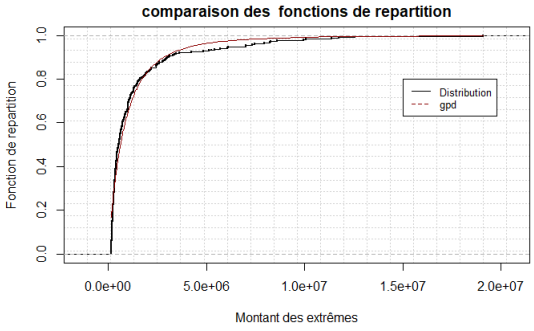

FIGURE 30 – Comparaison des Fonctions de répartition au-delà du seuil

Au regard de ces graphes, les sinistres au-delà du seuil **u=150 000** peuvent être approximés à une Pareto généralisé.

6.4 Synthèse

dans cette section, l'utilisation des méthodes de la théorie des valeurs extrêmes (Mean excess plot, Hill, Pickands, Gerstengarbe) ont permis de déterminer notre seuil de grave. En effet, le Mean excess plot et Gerstengarbe plot ont donné le même seuil $u = 150\ 000$ et les deux autres nous ont fourni une zone de recherche de seuil contenant la valeur de 150 000. Étant donné que cette valeur correspond au seuil fixé par Generali en plus d'être candidat sur l'ensemble des méthodes appliquées. De plus, l'adéquation au-delà du seuil à une GPD semble être vérifié, donc **la valeur de 150 000** a été choisie comme le seuil séparant les sinistres attritionnels et graves. Toutefois, il est important de noter que ce choix est arbitraire et que notre modélisation sera sensible par rapport au seuil.

7 Outils mathématiques pour la modélisation

7.1 Modèle linéaire généralisé ou GLM

7.1.1 Présentation de la théorie des GLMs

Le modèle linéaire généralisé est une extension de la régression linéaire qui relâche l'hypothèse de normalité sur les résidus. Rappelons que le principe de la régression est de modéliser $E[Y|X]$ comme une fonction g des variables explicatives X, soit

$$E[Y|X] \ = \ g(X)$$

La variable aléatoire d'interêt Y s'écrit alors

$$Y = g(X) + \epsilon$$

où ϵ est un bruit aléatoire qui représente l'erreur commise lorsqu'on modélise Y par son espérance conditionnelle.
En régression linéaire, nous considérerons les hypothèses fortes suivantes :

- $E[Y|X] \ = \ X\beta$
- la matrice de X est de plein rang
- Les variables aléatoires Y et ϵ sont indépendantes
- $Y|X \sim \mathcal{N}(X\beta, \sigma)$ et $\epsilon|X \sim \mathcal{N}(0, \sigma)$

Définition

Un modèle est un modèle linéaire généralisé s'il vérifie les hypothèses suivantes :

1. $Y|X \sim P_{\theta,\phi}$ appartient à une famille exponnentielle
2. $g(\mu(X)) \ = \ g(E[Y|X]) \ = \ X\beta$, pour une certaine fonction g bijective appelée fonction de lien

L'avantage des GLMs est qu'ils conservent la simplicité des modèles linéaires en élargissant les distributions acceptées à l'ensemble de la Famille exponentielle parmi laquelle nous pouvons citer la loi normale, exponentielle, gamma, Poisson, Bernoulli, etc.

7.1.2 Familles exponentielles

Un modèle statistique $(\Omega, \mathcal{F}, (P_{\theta,\phi})_{\theta\in\Theta,\phi>0})$ est de famille exponentielle si les probabilités $P_{\theta,\phi}$ admettent une densité f par rapport à la mesure dominante $\mathcal{F}$ avec

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

où

— θ est appelé paramètre canonique
— ϕ est appelé paramètre de dispersion
— $a(\theta)$ est de classe $\mathcal{C}^2$ et convexe
— $c_\phi(y)$ ne dépend pas de θ

Par exemple, considerons la variable $X \sim exp(\lambda)$ alors sa fonction densité est donnée comme suit :

$$f(x) = \lambda exp(-\lambda x)$$

Nous pouvons réécrire cette densité de la manière suivante :

$$f(x) = exp(\frac{-\lambda x - (-log(\lambda))}{1})$$

donc nous retrouvons les paramétres :

— $a(\theta) = -log(\lambda) = log(\frac{1}{\lambda})$
— $\theta = \frac{1}{\lambda}$
— $\phi = 1$
— $c_\phi(y) = 1$

7.1.3 Fonction de lien

Le principe de l'utilisation d'une fonction de lien en régression linéaire généralisée est d'appliquer une transformation de sorte que l'espérance conditionnelle de Y appliqué à la fonction soit linéaire en X, soit $g(E[Y|X]) = X\beta$. Nous citons ci-dessous quelques exemples classiques de fonction de lien

Distribution	Support	Nom du lien	Fonction de lien
Normale	R	identité	$g(x) = x$
Gamma	R_+^*	inverse	$g(x) = \frac{1}{x}$
Poisson	N	logarithme	$g(x) = log(x)$
Binomiale	$\{0, 1\}$	logit	$g(x) = \frac{x}{1-x}$

En pratique les fonctions identité et logarithme sont privilégiées bien que toute fonction bijective soit candidate. Ce choix est fait pour des raisons d'interprétabilités. En effet, le modèle peut se réécrire de la façon suivante :

$$E[Y|X] = g^{-1}(X\beta)$$

Ainsi, chaque coefficient de β dépend de la fonction de lien g choisie. La fonction de lien identité nous donne un modèle « additif », et la fonction logarithmique un modèle « multiplicatif ».

7.1.4 Prise en compte de l'exposition

L'exposition peut-être vue comme une variable de normalisation qui peut jouer un rôle très important dans certains cas. Dans la modélisation des montants des sinistres, pondéré par les engagements de l'assureur permet de prendre en compte la taille du risque. En effet, notons Y_i le montant du sinistre d'un assuré au cours de l'année i, X_i ses caractéristiques et ω_i l'engagement de l'assureur. Décrivons Y_i pondéré à ω en fonction de X_i avec un glm Gamma. en considérant l'inverse comme la fonction de lien canonique, nous obtenons le modèle suivant :

$$\frac{\omega_i}{E[Y|X]} = X_i\beta$$

qui peut se réécrire comme suit :

$$\frac{1}{E[Y|X]} = \frac{\omega_i}{X_i}\beta$$

on voit que plus ω_i est grand plus Y_i a des chances de l'être aussi.
Ceci est pareil pour la modélisation aussi en considérant ω_i comme l'exposition du contrat dans l'année et Y_i le nombre de sinistre. en ajustant un glm poisson on obtient :

$$log(E[Y_i|X_i]) = X_i\beta + \underbrace{\log(\omega_i)}_{offset}$$

7.1.5 Estimation des paramètres

Les principaux paramètres à estimer sont les coefficients $\beta_1, \beta_1, ..., \beta_p$. Ces derniers sont estimés par maximum de vraisemblance. Ainsi, en supposant que les Y_i soient indépendants et en notant :

$$\begin{cases} \eta_i &= X_i\beta \\ \mu_i &= E[Y_i|X_i] = g^{-1}(X_i\beta) = a^{'}(\theta_i) = g^{-1}(\eta_i) \\ \theta_i &= (a^{'})^{-1}(\mu_i) = (a^{'})^{-1}(g^{-1}(X_i\beta)) = (a^{'})^{-1}(g^{-1}(\eta_i)) \end{cases}$$

La log-vraisemblance s'écrit comme suit :

$$l(\beta) = \sum_{i=1}^{n} log(f(Y_i, \beta, \phi)) = \sum_{i=1}^{n} \underbrace{\left\{ logC_\phi(Y_i) + \frac{Y_i\theta_i - a(\theta_i)}{\phi} \right\}}_{:=l_i(\theta_i)}$$

Les solutions de ce problème d'optimisation sont obtenues par des méthodes numériques comme par exemple la méthode de Newton-Raphson.

7.1.6 Sélection de modèle

Le critère de la déviance

Le principe du critère de la déviance est de comparer notre modèle $\mathcal{M}$ au modèle dite saturé à partir de leur vraisemblance. Le modèle saturé est un modèle de GLM avec la même distribution et la même fonction de lien que le modèle $\mathcal{M}$ mais qui a autant de paramètre que de variables réponses. Par exemple, si on suppose que $rang(X) = n$ et qu'on a $Y = (Y_1, Y_2, ..., Y_n)$, alors le modèle saturé est décrit avec n paramètres. Une manière naturelle de comparer ces deux modèles à partir de leur vraisemblance est d'effectuer un test de rapport de vraisemblance.

Notons $\hat{\mathcal{L}}$ la vraisemblance du modèle ajusté et $\tilde{\mathcal{L}}$ la vraisemblance du modèle saturé. Alors, la déviance est définie par la quantité suivante :

$$\Delta = 2log(\frac{\tilde{\mathcal{L}}}{\hat{\mathcal{L}}}) = 2(\tilde{l} - \hat{l})$$

avec

— $\tilde{l}$ la log-vraisemblance du modèle saturé
— $\hat{l}$ la log-vraisemblance du modèle ajusté

en remplaçant $\tilde{l}$ et $\hat{l}$ par leurs expressions, la déviance peut se reécrire de la façon suivante :

$$\Delta = \sum_{i=1}^{n} \underbrace{2\left\{ \frac{Y_i(\tilde{\theta}_i - \hat{\theta}_i) + a(\hat{\theta}_i) - a(\tilde{\theta}_i)}{\phi} \right\}}_{:=\delta_i^2}$$

L'objectif est de trouver le modèle qui a le moins de covariable possible et qui minimise cette quantité afin d'avoir un modèle robuste qui décrit bien la variable

d'intérêt. En pratique, grâce au théorème de Wilks qui nous dit que la déviance converge en loi vers une χ^2_{n-p} , comparé l'estimé au saturé est équivaut à comparer la déviance à une loi de χ^2 au bon nombre de degrés de liberté. En plus comme la déviance est égale à la somme des contributions des observations sur la déviance, soit $\Delta = \sum_{i=1}^{n} \delta_i^2$, donc une valeur de δ_i grand indique que l'observation i contribue au mauvais ajustement. En effectuant une Anova, nous pourrons éliminer les variables qui ne sont pas significatives dans le modèle, autrement dit qui n'apportent pas d'information au modèle.

Le critère de sélection pas-à-pas

La méthode de sélection par étape consiste à déterminer le « meilleur » sous ensemble de variables explicatives d'un modèle. Il y a trois façons de faire une sélection de variable pas-à-pas :

1. L'approche *forward* part du modèle ne contenant que l'intercepte puis ajoute progressivement les autres variables jusqu'à ce que la nouvelle variable ne soit pas significative ou que l'AIC (Akaïke Information Criterion) ne baisse plus. L'AIC permet d'évaluer la perte d'information engendrée par l'utilisation d'un modèle pour décrire le processus qui génère les données. Cependant, il ne nous dit rien sur la qualité absolue du modèle. Il est défini par la formule suivante :

   $$AIC = -2l + 2p$$

   où p est le nombre de paramètres à estimer du modèle et l est la fonction de log-vraisemblance maximisée du modèle.
2. L'approche *backward* part du modèle contenant l'ensemble des variables explicatives puis les retire progressivement en fonction de celle qui est la moins pertinente dans le modèle jusqu'à ce que l'AIC ne diminue plus.
3. L'approche stepwise combine les deux définies ci-dessus. Partant du modèle nul, une variable est ajoutée dans le modèle à chaque itération puis on vérifie si l'ajout de la variable améliore le modèle. Si certaines variables deviennent non significatives, alors on retire la moins significative. Cette procédure est réitérée jusqu'à ce qu'aucune variable ne puisse être ajoutée ou retirée du modèle.

Toutefois, notons qu'une variable qui était très significative à une étape, peut redevenir à une étape ultérieure non-significative. De ce fait, il est important d'inspecter et de valider en fonction de la connaissance de la branche les variables retenues à l'issue de ces techniques.

7.2 Modèle CART

7.2.1 Présentation de l'algorithme

La méthode CART (classification and regression trees) est un algorithme d'apprentissage supervisé qui consiste à élaborer une séquence de nœud. À chaque étape, l'algorithme cherche un **critère de division** et une **variable** afin de diminuer l'impurité du nœud en effectuant un découpage du dernier en deux. Ainsi, cette procédure est réitérée jusqu'a ce qu'on obtient l'arbre maximal. Le premier nœud de l'algorithme, appelé **racine**, comprend l'ensemble de l'échantillon de départ. Il correspond au sommet de l'arbre. Les nœuds terminaux, appelés **feuilles**, contiennent les sous-espaces homogènes crées.

Un arbre très profond n'améliore pas forcément le modèle et peut entraîner un sur-apprentissage (peu de biais, mais beaucoup de variances). Pour éviter cela, nous effectuerons de **l'élagage** (pruning en anglais) pour ensuite choisir l'arbre qui a la plus petite erreur de prédiction calculée par validation croisée.

Critère de découpage

L'objectif de l'algorithme est de construire des nœuds de sorte que l'hétérogénéité soit maximale entre elles et minimale au sein d'un nœud. En régression, la mesure utilisée pour évaluer l'impurité d'un nœud $\mathcal{N}$ est la variance :

$$\mathcal{I}(\mathcal{N}) = \frac{1}{|\mathcal{N}|} \sum_{i:X_i \in \mathcal{N}} (Y_i - \bar{Y}_{\mathcal{N}})^2$$

où

— $\bar{Y}_{\mathcal{N}}$ désigne la moyenne des Y_i dans $\mathcal{N}$,
— $|\mathcal{N}|$ désigne l'effectif du noeud $\mathcal{N}$

Dans le cas d'un problème de classification, le critère de Gini sera utilisé pour mesurer l'impurité d'un nœud $\mathcal{N}$. L'impurité d'un Gini est définie par :

$$\mathcal{I}(\mathcal{N}) = \sum_{k=1}^{m} p_k(\mathcal{N})(1 - p_k(\mathcal{N}))$$

où $p_k(\mathcal{N})$ désigne la proportion de la classe k dans le noeud $\mathcal{N}$ avec $k = 1, ..., m$.

Dans la pratique, la procédure permettant de construire l'arbre optimal se fait en trois étapes que sont :

1. **étape1 :** Construction de l'arbre maximale T_{max} qui consiste à déterminer de mannière récursive un ensemble de « règles » $R_j(x)$ pour diviser les données, dans le but d'optimiser une fonction objective (également appelée critère de découpage). Chaque règle $R_j(x)$ à l'étape k génère deux règles $R_{j1}(x)$ et $R_{j2}(x)$ à l'étape k + 1. Pour chaque valeur possible des covariables x ; $\sum_j R_j(x) = 1$ et pour tout $j \neq j'$ $R_j(x)R'_j(x) = 0$. L'arbre maximale T_{max} est celui associé à la partition la plus fine.

2. **étape2 :** Elagage de T_{max}. Cette étape consiste à extraire une sous-suite de l'arbre T_{max} précédemment obtenu par minimisation, pour $\alpha > 0$, du critère pénalisé $C_\alpha(T)$.

$$C_\alpha(T) = \Gamma(\hat{s}_T) + \alpha|T|$$

   où $\Gamma(\hat{s}_T)$ mesure l'erreur d'ajustement de l'arbre T et $|T|$ désigne le nombre de feuille de l'arbre T.

3. **étape3 :** Sélection finale. On choisi l'arbre qui a la plus petite erreur de prédiction calculée par validation croisée.

7.3 Modèle Random Forest

7.3.1 Présentation de l'algorithme

L'algorithme du Random Forest ou Forêts aléatoires est un algorithme de type **baggin** introduit par Breiman. il consiste à agréger des arbres construits sur des échantillons bootstrap (ré-échantillonnage avec remise). Étant donné que les paramètres à calibrer pour un modèle d'apprentissage sont le biais et la variance, l'avantage des méthodes de bagging est qu'elles ne modifient pas le biais, mais diminuent la variance ; par conséquent, elles améliorent la performance du modèle.
Le prédicteur des forêts aléatoires $\hat{T}_B(x)$ est donné par la formule suivante :
Soit $T_k(x),\ k = 1, ..., B$ des prédicteurs par arbre avec $T_k\ :\ R^p \to R$ où p est le nombre de variable initiale

$$\hat{T}_B(x)\ =\ \frac{1}{B}\sum_{k=1}^{B} T_k(x)$$

Lors de la séparation d'un nœud en deux, Breiman suggère de sélectionner la « meilleure » parmi les m variables choisies aléatoirement dans les p variables initiales afin de diminuer la corrélation entre les arbres que l'on agrège.

Algorithme

Soient (X, Y) un couple aléatoire à valeur dans R^p x R , $\mathcal{D}_n = (X_1, Y_1), ..., (X_n, Y_n)$ un n-échantillon *iid* de même loi que (X, Y) et B le nombre d'arbres.
Pour k=1,...,B :

1. Tirage d'un échantillon bootsrap dans $\mathcal{D}_n$
2. Construction d'un arbre T_{θ_k}(x,$\mathcal{D}_n$) sur l'échantillon bootstrap
3. Retourne $\hat{T}_B(\text{x}) = \frac{1}{B}\sum_{k=1}^{B} T_{\theta_k}(\text{x},\mathcal{D}_n)$

La construction de l'arbre se fait de sorte que pour chaque segmentation, on ait le « meilleur » nœud à partir des m variables choisies aléatoirement dans les p variables initiales.

7.3.2 Importance des variables

Contrairement aux modèles paramétriques où il est facile de trouver l'influence des variables dans le modèle à travers des coefficients explicites, les forêts aléatoires présentent un aspect « boite noire ». Cependant, il existe des moyens permettant de mesurer l'importance des variables dans le modèle. Parmi ces moyens, nous avons l'indicateur d'erreur OOB (Out Of Bag). Ce dernier se base sur le fait qu'on n'utilise qu'un certain nombre d'observation pour construire les arbres de la forêt. Pour chaque (X_i, Y_i) de $\mathcal{D}_n$ on note $\mathcal{I}_B$ l'ensemble des arbres de la forêt qui ne contiennent pas l'observation (X_i, Y_i) dans leur échantillon bootstrap. La prévision $\hat{Y}_i$ de Y au point X_i est donnée par :

$$\hat{Y}_i = \frac{1}{\mathcal{I}_B} \sum_{k \in \mathcal{I}_B} T_{\theta_k}(X_i, \mathcal{D}_n)$$

Les estimateurs de OOB sont les suivants :

$$Erreur_{prédiction} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

$$Probabilite_{erreur} = \frac{1}{n}\sum_{i=1}^{n} 1_{\hat{Y}_i \neq Y_i}$$

Définition

Soient OOB_k^j l'échantillon OOB_k (l'échantillon de OOB associé au $k^{ème}$ arbre) dans lequel on a perturbé aléatoirement les valeurs de la variable j et $E_{OOB_k^j}$ l'erreur de prédiction de l'arbre k mesurée sur cet échantillon et définie de la façon suivante :

$$E^j_{OOB_k} = \frac{1}{|OOB_k^j|} \sum_{i \in OOB_k^j} (T_{\theta_k}(X_i^j, \mathcal{D}_n) - Y_i)^2$$

L'importance de la $j^{ième}$ variable est définie par :

$$Imp(X_j) = \frac{1}{B} \sum_{k=1}^{B} (E^j_{OOB_k} - E_{OOB_k})$$

où

$$E_{OOB_k} = \frac{1}{|OOB_k|} \sum_{i \in OOB_k} (T_{\theta_k}(X_i, \mathcal{D}_n) - Y_i)^2$$

7.4 Modèle Gradient Boosting

7.4.1 Présentation de l'algorithme

Le principe des méthodes de type boosting consiste à créer successivement des arbres d'apprentissage appelés des "weaks learners" qui ne sont pas indépendants entre eux. Chaque "weak learner" est entraîné pour corriger les erreurs des "weaks learners" précédents. Ce processus nous permettra d'obtenir à la fin un "strong learner". Le gradient boosting est une technique de boosting dont l'objectif est de trouver dans une famille de règle $\mathcal{G}$ (par exemple l'ensemble des arbres binaires) celle qui minimise une fonction de perte (fonction permettant de mesurer l'erreur de prédiction) en utilisant un algorithme de descente de gradient.

Algorithme

Soient $(x_1, y_1), ..., (x_n, y_n)$ un échantillon de couple de variable aléatoire (X, Y), λ un paramètre de régularisation tel que $0 < \lambda \leq 1$ et $\mathbf{M}$ le nombre d'itération

1. Initialisation avec $f_0 = \underset{\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \gamma)$ où l designe la fonction de perte.

2. pour m allant de 1 à $\mathbf{M}$

   - Pour i allant de 1 à n on calcule les pseudos-résidus :

   $$\left[r_{im} = \frac{\partial}{\partial f(x_i)} l(y_i, f(x_i)) \right]_{f(x_i)=f_{m-1}(x_i)}$$

   - On ajuste un classifieur faible h_m sur l'echantillon $(x_1, r_{1m}), ..., (x_n, r_{nm})$ puis on détermine le poids associé à ce classifieur

   $$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_{m-1}(x_i) + \gamma h_m(x_i))$$

   - On met à jour le modèle

   $$f_m(x) + f_{m-1}(x) + \lambda \gamma_m h_m(x)$$

3. Sortie : suite de régles $(f_m(x))_m$ qui n'est qu'une combinaison d'arbre

Comme pour la plupart des méthodes de machine learning, l'algorithme du gradient boosting nécessite le calibrage des hyperparamètres afin d'éviter le surapprentissage, mais aussi d'optimiser la performance du modèle. Parmi ces hyperparamètres nous avons le nombre d'arbres dans le modèle $\mathbf{M}$, le coefficient γ pour régulariser les profondeurs des arbres et le taux d'apprentissage λ appelé "learning rate". Un taux élevé pourrait entraîner la non-convergence du modèle vers un optimal et inversement un taux très faible entraînera une convergence très lente vers l'optimal. Ainsi, ces hyperparamètres seront optimisés par validation croisée avec une recherche par grille.

7.5 Indicateurs de performance

Dans cette partie, nous allons définir plusieurs indicateurs permettant d'évaluer la performance de nos différents modèles qui vont être mis en place.
Définissons les notations suivantes :

- n : le nombre d'observations,
- y_i : la valeur de la $i^{ème}$ observation,
- $\hat{y_i}$: la valeur éstimée pour la $i^{ème}$ observation,
- $\bar{y}$: la moyenne des observations

7.5.1 Erreur moyenne absolue : MAE

L'erreur moyenne absolue notée MAE (Mean Absolute Error) est la moyenne arithmétique des valeurs absolues des écarts entre la valeur prédite et la valeur observée. Elle est définie par la formule suivante :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

7.5.2 Erreur quadratique moyenne : RMSE

L'erreur quadratique moyenne notée RMSE (Root Mean Square Error) est la racine carrée de la moyenne du carré des erreurs de prédiction. Elle est définie comme suit :

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

Notons toutefois que cet indicateur est sensible aux grandes valeurs : en regardant le carré des erreurs, nous attribuons un poids très important aux erreurs.

7.5.3 Le pourcentage des erreurs absolues moyennes : MAPE

Le pourcentage des erreurs absolues moyennes noté MAPE(Mean Absolute Percent Error) est défini de la manière suivante :

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y_i}}{yi}|$$

7.5.4 L'erreur globale de la modélisation

L'erreur globale de la modélisation notée ξ_{global} est définie comme suit

$$\xi_{global} = 1 - \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{y}_i}$$

Plus ξ_{global} est proche de 0 meilleur est le modèle. Une valeur de $\xi_{global} > 0$ indiquera que notre modèle sur-estime les valeurs et inversement une valeur de $\xi_{global} < 0$ indiquera que le modèle sous-estime les valeurs.

7.5.5 L'indice de Gini

L'indice de Gini ou coefficient de Gini est un indicateur permettant de mesurer la répartition d'une variable au sein d'une population. En modélisation, il est utilisé pour évaluer la capacité du modèle à bien classer les risques, soit la capacité de segmentation du modèle.
Il est compris entre 0 et 1, 0 étant la situation d'égalité parfaite (par exemple, tous les assurés du portefeuille ont la même fréquence de sinistre.) et 1 étant la situation d'inégalité totale (seul un assuré qui représente toute la sinistralité du portefeuille).

Cet indicateur a été développé par le statisticien italien Corrado Gini et son calcul se base sur la courbe de Lorenz qui peut être définie comme étant la représentation graphique du pourcentage cumulé d'une variable y en fonction du pourcentage cumulé d'une variable x.

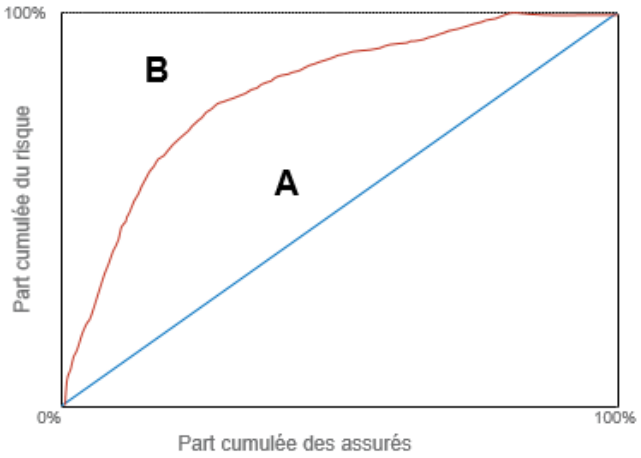

FIGURE 31 – courbe de lorenz

En considérant la figure ci-dessus, on définit l'indice de gini par la formule suivante :

$$IG = \frac{Aire\ de\ A}{Aire\ de\ A + Aire\ de\ B}$$

7.5.6 L'AUC

L'AUC permet de mesurer la qualité d'un modèle de classification. Il correspond à la surface sous la courbe de ROC. Cette courbe représente le taux de vrais positifs (VP) appelé sensibilité en fonction du taux de faux positifs (FP) appelé anti-spécificité.
Notons FN = faux négatifs et VN = vrais négatifs, on a :

$$sensibilité = Rappel = \frac{VP}{FN + VP}$$

$$spécificité = \frac{VN}{VN + FP}$$

$$anti - spécificité = 1 - spécificité$$

l'objectif est de maximiser l'AUC car plus il est grand, meilleur est le modèle.

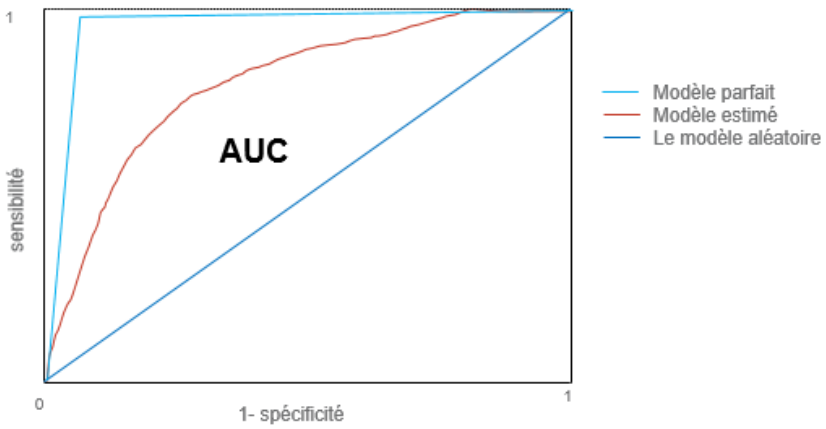

FIGURE 32 – courbe de ROC

7.6 Critère de validation croisée

Le principe de la validation croisée ou cross-validation permet d'évaluer la robustesse et la capacité du modèle à être généralisé sur de nouvelles données. Ce critère permet de lutter contre le sur-apprentissage. En effet, lors de la modélisation, si nous construisons un modèle sur la base d'apprentissage puis calculer une erreur sur la base de test ; on risque d'avoir une mauvaise idée sur la robustesse du modèle à cause d'une parfaite séparation de nos données d'entraînement et de contrôle.

Avec la validation croisée, les données d'apprentissage sont divisées en k sous-échantillons. Ainsi, le modèle est entraîné puis un indicateur de performance (erreur du modèle par exemple) est calculé sur la base test k fois. À chaque étape, il est entraîné sur k-1 échantillons pour être testé sur l'échantillon restant. De ce fait, l'erreur du modèle est obtenue en calculant la moyenne des erreurs commises à chaque itération. Le graphe ci-dessous schématise la procédure de cette méthode pour k=5 :

FIGURE 33 – illustration du principe de la validation croisée

8 Pré-sélection des variables explicatives

La sélection de variables servant à construire les modèles est une étape très importante, car la suite de l'étude dépendra de ces variables pour l'explication des résultats. Notre base de données d'étude contient plusieurs variables, nous allons, dans cette section, sélectionner les variables les plus discriminantes pour l'ensemble des modèles que nous souhaitons mettre en œuvre. De ce fait, un **Random Forest** est entraîné pour chaque modèle afin de sélectionner les vingt variables les plus importantes comme facteurs explicatifs pour la suite. Pour les glms, un stepwise sera effectué pour choisir le modèle définitif.

Application

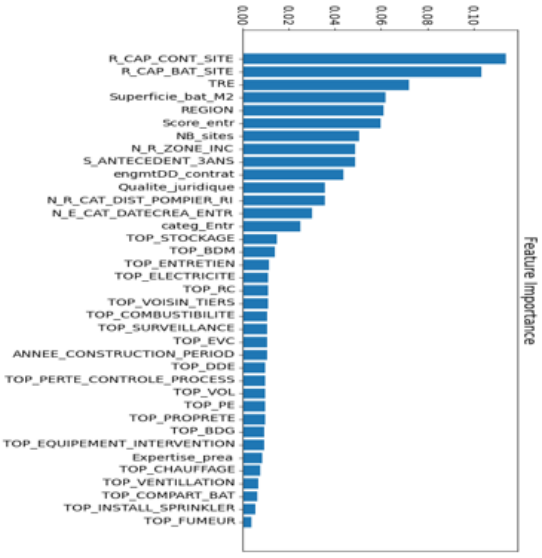

FIGURE 34 – Variables d'importance pour la propension

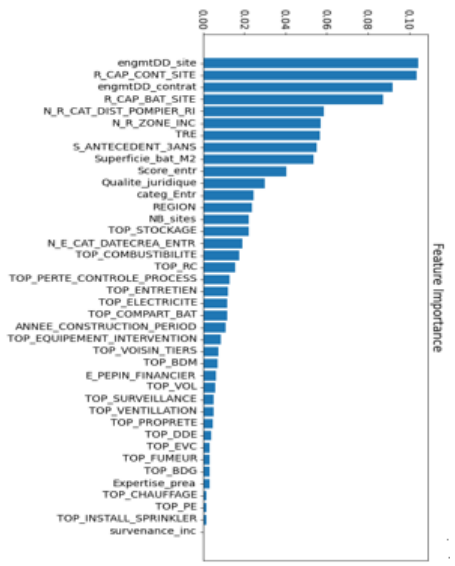

FIGURE 35 – Variables d'importance pour la sévérité

FIGURE 36 – Variables d'importance pour la fréquence

Nous sélectionnons par la suite les vingt variables les plus importantes pour l'élaboration des modèles. Ainsi, à chaque fois, une étude de corrélation sera effectuée sur l'ensemble des variables pré-sélectionnées afin de voir s'il y a une présence de dépendance entre elles. À cet effet, si deux variables sont corrélées, alors nous retenons soit la plus pertinente dans le sens opérationnel ou nous effectuerons notre étude en prenant en compte les deux variables séparément pour ensuite choisir celle qui permet d'obtenir le modèle le plus performant.

8.1 Étude des corrélations

8.1.1 Corrélation de Pearson

Le coefficient de corrélation de Pearson permet d'évaluer une relation linéaire entre deux variables continues. Il est défini comme suit :

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

où

- $cov(X, Y)$ est la covariance entre X et Y
- σ_X, σ_Y sont respectivement l'écart type de X et de Y.

Ce coefficient est compris entre -1 et 1. Si les deux variables sont indépendantes, alors $r_{XY} = 0$ sinon $r_{XY} \approx 1$ si elle évolue dans le même sens ou $r_{XY} \approx -1$ si elle évolue dans le sens opposé.

8.1.2 Corrélation de Spearman

Le coefficient de corrélation de Spearman permet de mesurer une relation entre deux variables en utilisant leur rang. En notant $rang(X)$ et $rang(Y)$ comme étant respectivement le rang de X et de Y, la corrélation de Spearman est définie comme suit :

$$\rho_{XY} = \frac{cov(rang(X), rang(Y))}{\sigma_X \sigma_Y}$$

Sa valeur varie entre -1 et 1 où 0 représente une relation nulle entre les deux variables, une valeur positive reflète une comonotonie entre les deux variables et une valeur négative représente une anti-comonotonie.

8.1.3 Tau de Kendall

Le tau de Kendall est une mesure de corrélation de rang qui est comprise entre -1 et 1 s'interprète de la même façon que les précédents. Il est défini par :

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

où

- n_c est le nombre de paire $\{(x_i, y_i), (x_j, y_j)\}$ concordante
- n_d est le nombre de paire $(x_i, y_i), (x_j, y_j)$ discordante
- n est la taille de l'echantillon.

Définition

Soient $(x_1, y_1), ..., (x_n, y_n)$ des copies des variables aléatoires jointes (X, Y). On dit que (x_i, y_i) et (x_j, y_j) sont concordants si $x_i > y_i$ et $x_j > y_j$ ou si $x_i y_i$ et $x_j y_j$

8.1.4 V de Crammer

Le V de Crammer permet de mesurer la liaison entre deux variables qualitatives qui se base sur le test de statistique de khi2. Il varie entre -1 et 1 et s'interprète de la même manniere que les précendents. Il est défini par :

$$V = \sqrt{\frac{\chi^2}{n[min(l, c) - 1]}}$$

où

- n est le nombre total d'observation
- l est le nombre de modalité de X
- c est le nombre de modalité de Y

Application

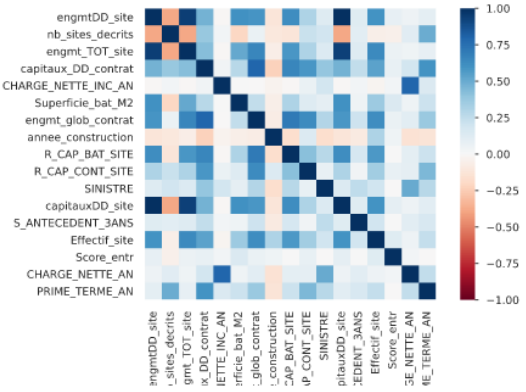

FIGURE 37 – corrélation de pearson

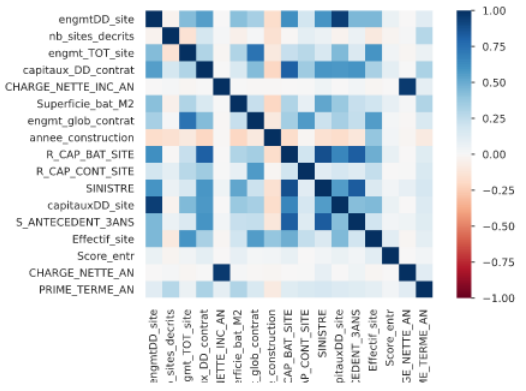

FIGURE 38 – corrélation de spearman

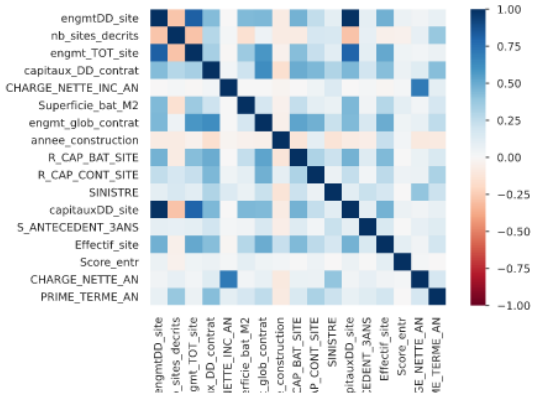

FIGURE 39 – Taux de kendall

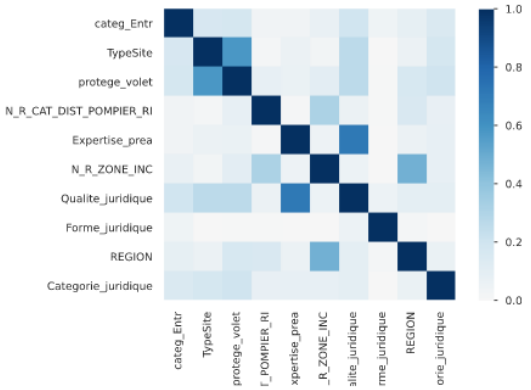

FIGURE 40 – Vcrammer

Les graphes ci-dessus représentent les matrices de corrélations entre nos variables explicatives présentes dans notre jeu de données. Ils s'interprètent de la façon suivante : plus la couleur est foncée plus les deux variables sont corrélées. Ainsi, au regard des matrices de corrélation, on voit qu'il y a peu de variables fortement corrélées dans notre base d'étude.

9 Pré-traitement des variables

9.1 Regroupement de Modalité

Ce travail consiste à regrouper des modalités de certaines variables ayant un grand nombre de niveaux dans notre jeu de donnée afin d'améliorer la performance du modèle, mais aussi d'éviter d'avoir des modalités qui soient présentes que sur la base d'entraînement ou sur la base de test. Cela permettra aussi d'éviter le sur-apprentissage et la présence d'un biais qui peut être entraînée par les modalités les moins représentées dans la base d'étude.
Ainsi, nous avons choisi de regrouper les modalités dont le nombre est inférieur à cinquante. Par exemple pour la variable code TRE, toutes les activités d'un fascicule dont le nombre est inférieur à cinquante dans la base sont regroupées ensembles.

9.2 Discrétisation des données quantitatives

Comme le regroupement des modalités, la discrétisation des variables quantitatives aidera à améliorer la performance du modèle. Ainsi, il existe plusieurs

techniques permettant de faire ce travail notamment des méthodes de machine learning (les k plus proches voisins,...). Dans ce mémoire, nous avons opté de ne pas utiliser des méthodes de machine learning pour la discrétisation de nos variables quantitatives, mais plutôt à la main de la façon suivante :

1. D'abord, nous créons des classes de faibles amplitudes,
2. Ensuite, nous lançons notre modèle et vérifions celles qui ne sont pas significatives pour les regrouper à la classe la plus proche,
3. Nous relançons ensuite le modèle pour vérifier si on améliore la performance,
4. Ainsi de suite, nous répétons cette procédure jusqu'à ce que toutes les classes soient significatives ou que le modèle ne s'améliore plus.

À présent, nous allons passer à la construction des modèles permettant de mettre en place la nouvelle méthode de tarification adaptée aux risques volatiles. Pour chaque étape, un modèle linéaire généralisé sera ajusté et quelques méthodes d'apprentissage supervisée afin de voir si ces dernières améliorent la qualité du modèle de glm.

10 Construction du modèle de propension de grave

10.1 Application de la régression logistique

Notre étude commence par la modélisation de la survenance des sinistres graves, c'est-à-dire ceux dont le montant dépasse 150 000€, soit 265 sinistres. L'objectif de cette étude est d'essayer de trouver les facteurs qui pourraient expliquer la survenance (ou pas) de ces sinistres extrêmes et les probabilités qu'un assuré ait un sinistre grave. Ainsi, en notant C comme étant le montant du sinistre, la variable à expliquée Y a été construite de la manière suivante :

$$Y = \begin{cases} 1 & si\ C \geq 150000 \\ \\ 0 & sinon \end{cases}$$

On se retrouve devant un problème de classification binaire. De ce fait, un modèle de régression logistique sera ajusté. Ce dernier est un modèle de régression binomiale avec comme fonction de lien la fonction *logit* définie par :

$$\begin{aligned} g : [0,1] &\longrightarrow R \\ x &\longmapsto log(\tfrac{x}{1-x}) \end{aligned}$$

Et il s'écrit de la façon suivante :

$$log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = X\beta$$

où X est la matrice des variables explicatives et β le vecteur des coefficients.

Étant donné que toutes les bases du modèle de logistique ont été posées, nous avons procéder à la construction du modèle sur **R** avec les vingt variables présélectionnées à la section 8. Ces variables sont pré-traitée en utilisant la procédure présentée à la section 9. Ainsi, le modèle définitif est obtenu par la méthode de *stepwise* en utilisant le critère d'AIC. Ce travail nous mène à retenir les variables explicatives suivantes : la superficie du bâtiment, le nombre de sites, l'antécédent sinistre, le TOP_STOCKAGE, la catégorie d'entreprise et l'activité. Ainsi la performance du modèle retenu sur la base d'apprentissage et sur la base test est donnée ci-dessous.

10.1.1 Analyse de la qualité du modèle

base d'apprentissage

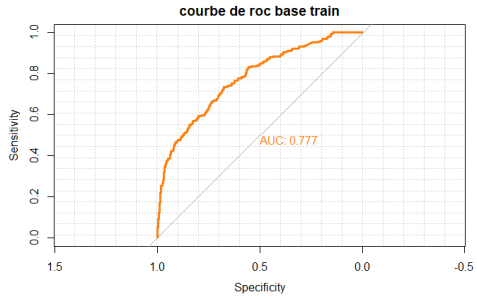

FIGURE 41 – courbe de roc sur la base d'entrainement

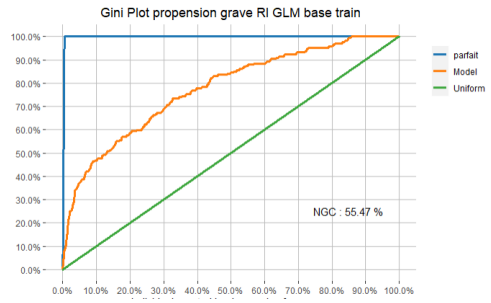

FIGURE 42 – courbe de Lorenze sur la base d'entrainement

base de test

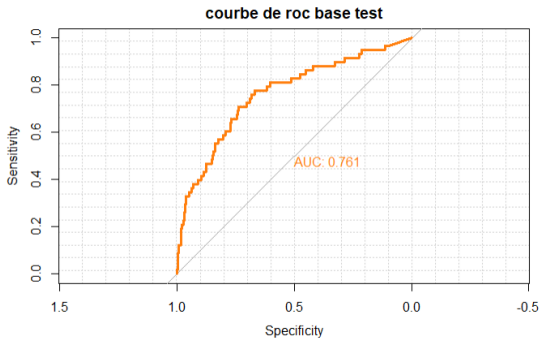

FIGURE 43 – courbe de roc sur la base test

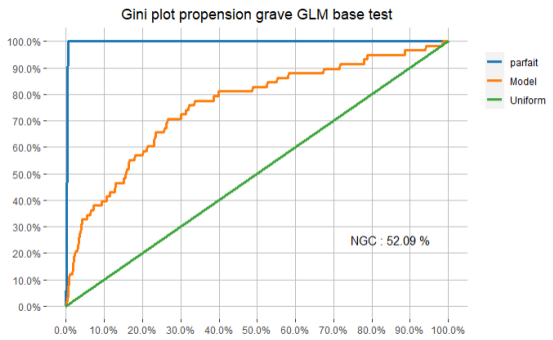

FIGURE 44 – courbe de Lorenze sur la base test

Comme nous somme face à un problème de classification, il est pertinent de vérifier que le modèle classe bien nos individus. Pour cela, vu qu'on est dans une situation où les classes sont très déséquilibrées (99% de 0 et 1% de 1), regarder la F_1-mesure est plus pertinent pour l'évaluation du modèle. Ce dernier peut-être défini comme étant une moyenne harmonisée de la précision et du rappel. Ce choix est motivé du fait que dans les situations où une classe prédomine sur l'autre, la précision n'est pas un indicateur fiable. En effet, dans notre cas de figure, même si le modèle ne prédit que des 0, nous aurons une précision de 99%. Donc on doit prendre en compte le rappel. la F_1-mesure est obtenu en faisant le calcul suivant :

$$F_1 - mesure = 2\frac{Précision * Rappel}{Précision + Rappel}$$

indicateur	Base train	Base test
F_1	0.06	0.04

cet indicateur a retenu toute notre attention car il nous dit que notre modèle n'est pas de bonne qualité contrairement à ce que les indicateurs d'AUC et le Gini nous révèlent. En effet, ce qui se passe est que le nombre de 1 prédit par notre modèle est très très faible. De ce fait on doit essayer de corriger cette anomalie soit en faisant du ré-échantillonnage pour rééquilibrer nos classes ou travailler directement sur notre seuil d'affectation.

Technique de ré-échantillonnage

Le principe de ré-échantillonnage consiste à équilibrer les classes soit en faisant du sur-échantillonnage, qui consiste à ajouter des 1 jusqu'à ce qu'on ait autant de 0 que de 1 dans notre échantillon, ou en faisant du sous-échantillonnage qui consiste à faire le phénomène inverse. Il est important de rappeler que ce travail sera fait sur notre base d'apprentissage.
Dans notre étude, nous avons fait du sur-échantillonnage et nous avons constaté une amélioration de la mesure F_1 qui est égale à 11% , par contre notre AUC et le Gini se sont un peu dégradés.

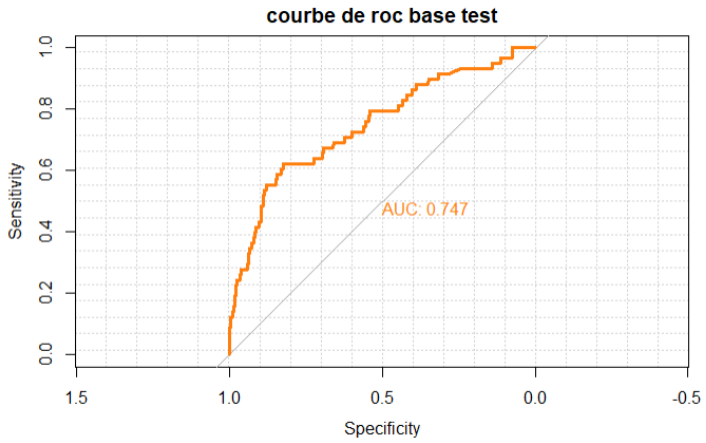

FIGURE 45 – courbe de ROC aprés ré-échantillonnage

Étant donné que la technique de re-échantillonnage n'améliore que la F_1 score grâce au rappel, nous allons essayer de voir si nous pouvons améliorer cet indicateur de manière explicite sur notre premier modèle en travaillant sur le seuil d'affectation.

Manipulation du seuil d'affectation

En pratique, le seuil d'affectation classique retenu est de 0.5. Il est obtenu grâce à l'affectation de Bayes qui nous dit que :

$$Y = \begin{cases} 1 & si\ P(Y=1|X) \geq P(Y=0|X) \\ \\ 0 & sinon \end{cases}$$

Or, on sait que $P(Y=1|X) + P(Y=0|X) = 1$ donc on obtient :

$$Y = \begin{cases} 1 & si\ P(Y=1|X) \geq 0.5 \\ \\ 0 & sinon \end{cases}$$

Dans cette partie, un travail est fait sur ce seuil en effectuant la chose suivante : sur l'échantillon d'apprentissage, on varie le seuil et pour chaque seuil, on calcule la F_1 mesure afin de trouver le seuil qui maximise la F_1 mesure. Ainsi, ce seuil optimal obtenu sera appliqué dans la prédiction sur l'échantillon test comme seuil d'affectation de la probabilité d'affectation.
Le graphe ci-dessous nous montre quel est le seuil qui maximise la F_1 mesure

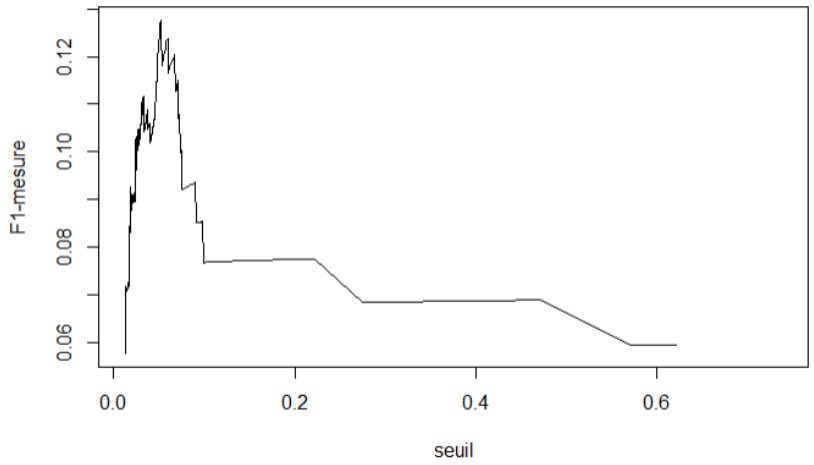

FIGURE 46 – graphe F_1 mesure

Ainsi en appliquant le seuil optimal sur notre échantillon de test, on obtient une F_1 mesure égale à 12%. Ce dernier est un peu amélioré par rapport à la méthode de ré-équilibrage de données.

10.1.2 Conclusion

Au final, le premier modèle où on a modifié le seuil d'affectation a été retenu. L'avantage de cette méthode est qu'on travaille directement sur les vraies données en optimisant un critère.

10.2 Application des arbres CART et Random forest

Nous avons essayé d'appliquer les méthodes de CART et des Forêts aléatoires pour la modélisation de la survenance des sinistres graves afin de voir si ces dernières améliorent la performance du modèle de glm.

10.2.1 Analyse de la qualité des modèles

Après avoir calibré nos hyperparamètres par validation croisée pour contrôler le processus d'apprentissage, nous obtenons les résultats suivants :

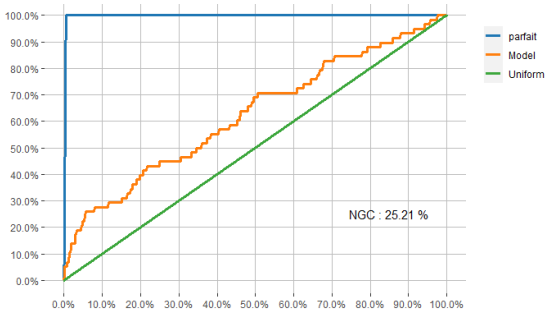

FIGURE 47 – Courbe de Lorenze méthode cart

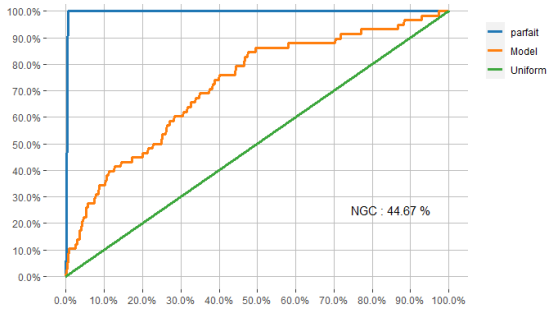

FIGURE 48 – Courbe de Lorenze méthode random forest

On constate que la méthode de CART n'est pas bien adaptée sur nos données car elles sont très déséquilibrées. Raison pour laquelle son Gini est très faible. Par contre, les forêts aléatoires qui sont une méthode provenant de la méthode CART par bagging a eu un score beaucoup plus intéressant que la méthode de CART. Cependant, toutes les deux méthodes ne font pas mieux que la régression logistique.

10.2.2 Conclusion

Les méthodes de CART et de random forest n'améliorent pas la performance du modèle de régression logistique pour la modélisation de la propension de grave au regard du critère de Gini. Donc nous conservons pour le moment la régression

logistique comme méthode de modélisation de la propension et nous allons ajuster le modèle de gradient boosting afin de voir si ce dernier parvient à améliorer la qualité de prédiction.

10.3 Application du modèle de Gradient Boosting

Toujours dans le but d'améliorer la performance de la modélisation de la survenance des sinistres graves par le glm binomial, nous avons calibré un modèle de xgboost.

10.3.1 Analyse de la qualité du modèle

Après avoir optimisé nos hyperparamètres par validation croisée avec une recherche par grille, nous obtenons un xgboost avec 500 arbres, une profondeur maximale de 10 et un coefficient de régularisation égale à 0.1. Ainsi, nous avons évalué notre modèle sur la base test et voici les résultats obtenus :

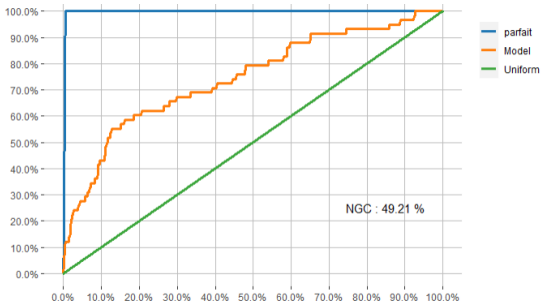

FIGURE 49 – Courbe de Lorenze méthode xgboost

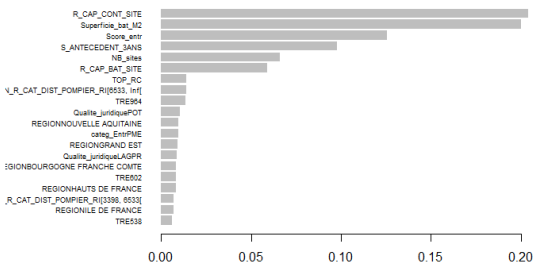

FIGURE 50 – les variables significatives

On constate que le xgboost a un Gini plus élevé que les méthodes CART et random Forest. Ce qui signifie qu'il classe mieux nos risques par rapport à ces dernières. Cependant, il est moins performant que la régression logistique. Toutefois, notons que les variables les plus significatives des deux modèles sont quasiment identiques.

10.3.2 Conclusion

Comme les méthodes CART et random forest, le xgboost est moins performant que le modèle de régression logistique construit à la section 10.1. Donc ce dernier sera retenu pour la modélisation de propension de grave.

11 Construction du modèle de Fréquence

11.1 Application de la régression de Poisson

Dans cette partie, nous procédons à la modélisation de la fréquence de sinistre. De ce fait, nous avons ajusté une régression linéaire généralisée de Poisson définie comme suit :

$$log(E[Y|X]) = X_i\beta + \underbrace{\log(\omega)}_{offset}$$

où

- X est la matrice des variables explicatives,
- β le vecteur des coefficients,
- Y le vecteur du nombre de sinistre annuel,
- ω le vecteur de l'exposition annuelle.

Ainsi, de façon analogique avec le modèle logistique, un pré-traitement de nos variables pré-sélectionnées a été effectué puis le modèle définitif est obtenu par le critère de *stepwise*. Par conséquent, les variables explicatives qui ont été retenues sont les suivantes : l'antécédent de sinistre, qualité de l'occupant, l'engagement, le nombre de sites, l'activité.

11.1.1 Analyse de la qualité du modèle

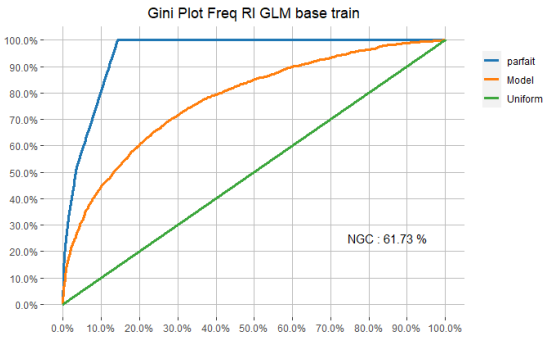

FIGURE 51 – courbe de Lorenze sur la base train

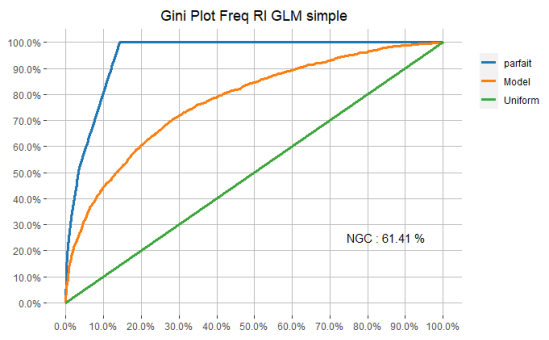

FIGURE 52 – courbe de Lorenze sur la base test

indicateur	Base train	Base test
$RMSE$	0.53	0.55
ξ_{global}	-2,57%	-3%

Les indices de Gini obtenus sur la base d'apprentissage et la base test sont très proches. Ceci indique que notre modèle ne fait pas de sur-apprentissage. En plus, le pouvoir discriminant du modèle ajusté est de bonne qualité, car son Gini est élevé. En regardant par la suite les indicateurs d'erreurs de prédiction, on observe que le $RMSE$ et l'erreur totale du modèle sont faibles et stables. Donc notre modèle est considéré comme étant robuste.

Comparaison de la fréquence prédite et observée

Dans cette partie, quelques variables significatives du modèle sont choisies afin de comparer pour chaque modalité la moyenne prédite et observée sur un même graphique.

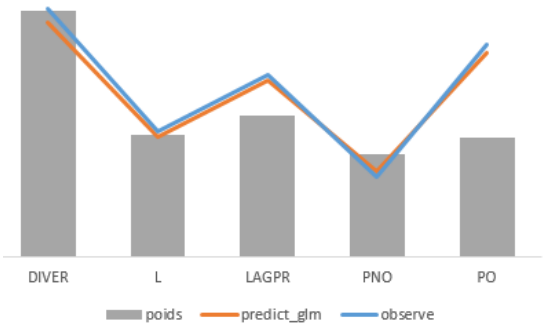

FIGURE 53 – graphe de comparaison des moyennes prédites et observées pour la qualité de l'occupant

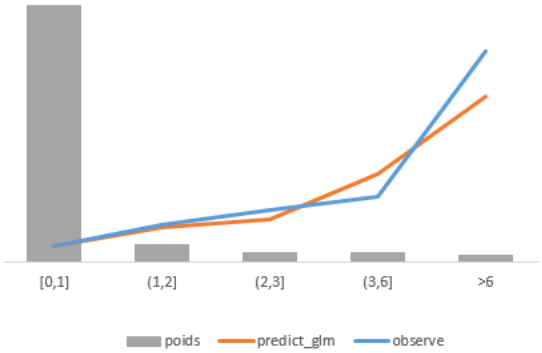

FIGURE 54 – graphe de comparaison des moyennes prédites et observées par antécédent de sinistre

Nous constatons que la fréquence prédite est proche de celle observée. Pour la qualité de l'occupant, on remarque que les propriétaires occupants (PO) et les locataires agissant pour le compte du propriétaire (LAGPR) ont une fréquence plus élevée par rapport aux locataires (L) et aux propriétaires non-occupant (PNO). De plus, on voit que la fréquence augmente en fonction de l'antécédent de sinistres. Le schéma ci-dessous montre aussi que la fréquence augmente en fonction de l'engagement.

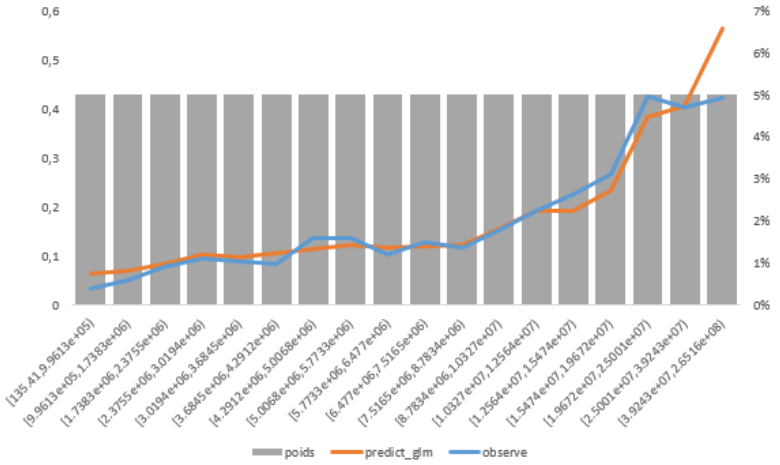

FIGURE 55 – graphe de comparaison des moyennes prédites et observées par tranche d'engagement

Dans le but d'améliorer la performance de notre modèle, un GLM de Poisson pénalisé par la méthode du Lasso est ajusté. Cette méthode consiste à maximiser la vraisemblance pénalisée par la norme 1 en rendant nuls certains coefficients de β donnée par :

$$\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0 + x_i^T\beta) - exp(\beta_0 + x_i^T\beta) + \lambda||\beta||_1\right\}$$

où $\lambda \in R^+$ est un paramètre de lissage. Une valeur élevée de ce paramètre augmente le nombre de coefficients nuls.

Performance du modèle

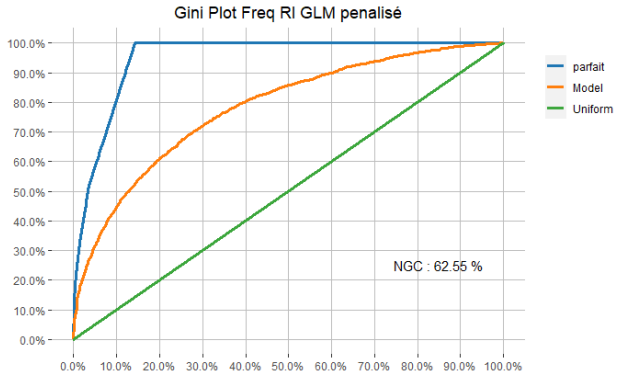

FIGURE 56 – Courbe de Lorenze glm pénalisé

On constate une légère amélioration de l'indice de gini par rapport au glm non pénalisé. Nous allons comparer les fréquences moyennes prédites et observées pour les variables antécédent de sinistres et engagement.

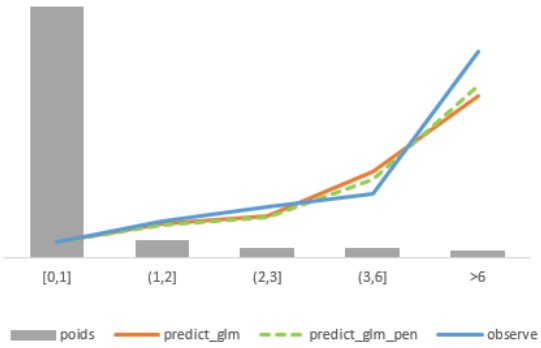

FIGURE 57 – graphe de comparaison des moyennes prédites et observées par antécédent de sinistres

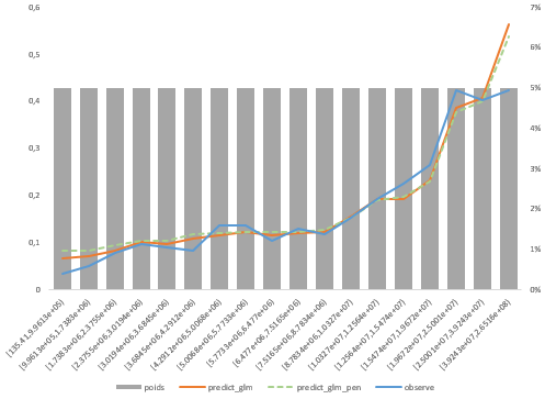

FIGURE 58 – graphe de comparaison des moyennes prédites et observées par tranche d'engagement

Nous constatons que la prédiction des deux modèles a la même tendance que l'observé et augmente tous en fonction de l'engagement et l'antécédent de sinistres. En plus le modèle de glm pénalisé n'améliore pas beaucoup la qualité de la prédiction moyenne du glm Poisson simple. En regardant les indicateurs d'erreur de prédiction, nous obtenons les résultats suivants :

indicateur	Base train	Base test
$RMSE$	0.53	0.54
ξ_{global}	-1.77%	-1.9%

11.1.2 Conclusion

Au regard des indicateurs de performance, on voit que la pénalisation diminue un peu l'erreur de prédiction et que l'indice de Gini s'est amélioré de 1 point.

11.2 Application des arbres CART et Random forest

De même que la survenance de sinistre grave, les modèles d'arbre de régression de CART et random forest ont été ajustés pour décrire la fréquence de sinistre du portefeuille.

11.2.1 Analyse de la qualité des modèles

Après avoir calibré nos hyperparamètres par validation croisée pour contrôler le processus d'apprentissage, les résultats obtenus sur la base test sont présentés ci-dessous :

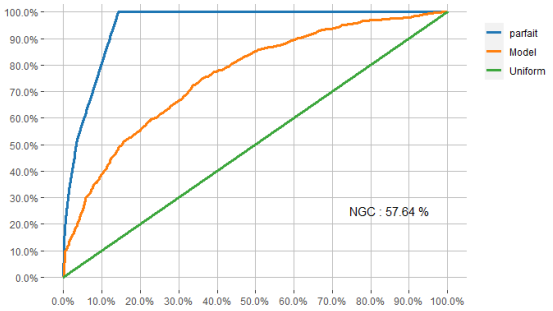

FIGURE 59 – Courbe de Lorenze méthode cart

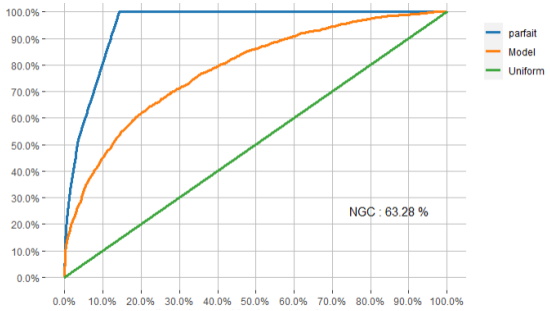

FIGURE 60 – Courbe de Lorenze méthode random forest

indicateur	random forest	CART
$RMSE$	0.49	0.59
ξ_{global}	1.25%	3.9%

On constate que, contrairement à la méthode de CART, les forêts aléatoires sont clairement meilleurs que la régression de poisson en termes de pouvoir prédictif. En effet, son Gini est assez élevé et l'erreur de prédiction est faible.

Comparaison de la fréquence prédite et observée

Les graphes ci-dessous nous montrent les fréquences moyennes prédites et observées par tranche d'engagement et par antécédents sinistres sur l'ensemble des modèles testés jusqu'à présent.

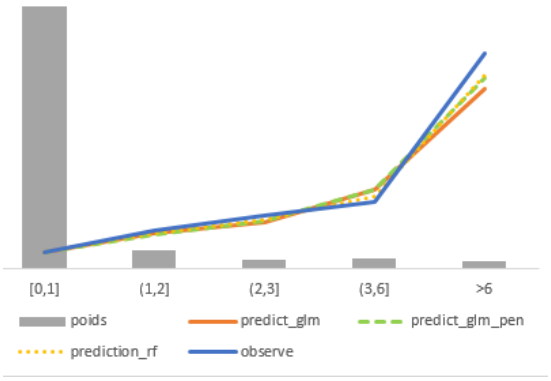

FIGURE 61 – graphe de comparaison des moyennes prédites et observées par antécédents sinistre

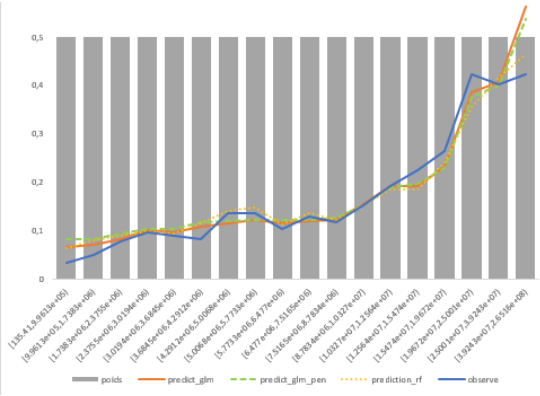

FIGURE 62 – graphe de comparaison des moyennes prédites et observées par tranche d'engagement

Au regard de ces graphiques, on voit que la méthode random forest améliore légèrement la qualité de prédiction des modèles précédement ajustés.

11.2.2 Conclusion

Contrairement à la méthodes de CART, le modèle de random forest a amélioré la performance du glm de Poisson. Donc nous le conservons pour le moment comme méthode de modélisation de la fréquence et nous allons ajuster le modèle de gradient boosting afin de voir si ce dernier parvient à améliorer la qualité de prédiction.

11.3 Application du modèle de Gradient Boosting

De la même façon avec le modèle de propension, les hyperparamètres du modèle de fréquence ont été calibrés par validation croisée avec une recherche par grille. Ainsi, nous obtenons un modèle avec 500 arbres, une profondeur maximale de 8 et un coefficient de régularisation de 0.02.

11.3.1 Analyse de la qualité du modèle

L'évaluation du modèle final obtenu, après calibration des hyperparamètres, sur la base test nous donne les résultats suivants :

indicateur	Base train	Base test
$RMSE$	0.45	0.49
ξ_{global}	0%	-1%

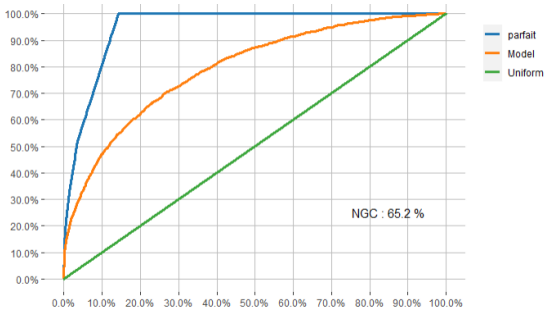

FIGURE 63 – Courbe de Lorenze base test

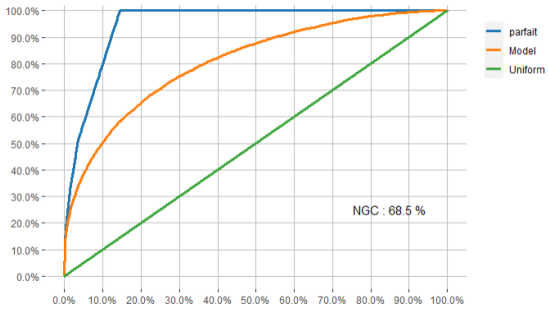

FIGURE 64 – Courbe de Lorenze base train

Au vu des résultats obtenus, il apparaît clairement que la modélisation de la fréquence par la méthode xgboost est meilleure que les autres méthodes utilisées précédemment avec un Gini de 65% et une erreur globale de -1% sur la base test. De plus, on a vérifié que le modèle ne sur-apprend pas, car les indicateurs de performance sur la base train et test sont proches. Ce qui signifie que la calibration des hyperparamètres est de bonne qualité. Les variables les plus significatives du modèle sont les suivantes :

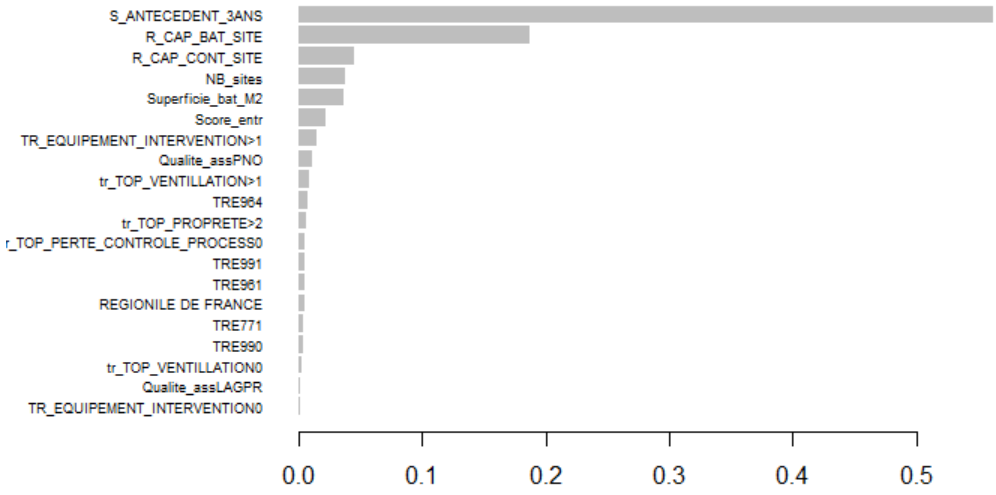

FIGURE 65 – variables significatives

On remarque que la variable antécédents de sinistres est la plus significative sur l'ensemble des modèles élaborés. De plus, les variables les plus discriminantes des modèles sont presque identiques.

Comparaison de la fréquence prédite et observée

Les graphes ci-dessous nous montrent les fréquences moyennes prédites et observées par tranche d'engagement et par antécédents sinistres sur l'ensemble des modèles mis en place.

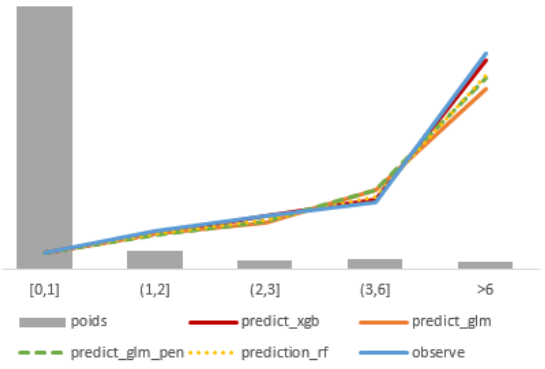

FIGURE 66 – graphe de comparaison des moyennes prédites et observées par antécédent de sinistres

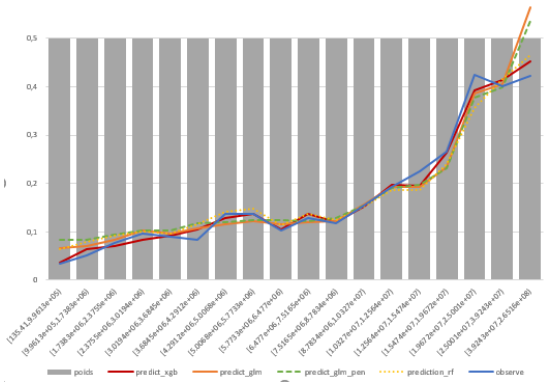

FIGURE 67 – graphe de comparaison des moyennes prédites et observées par tranche engagement

On voit nettement que la fréquence moyenne prédite par la méthode xgboost est très proche de la fréquence moyenne observée. Ce constat confirme la performance du modèle constatée sur les indicateurs.

11.3.2 Conclusion

Après avoir analysé la qualité du modèle xgboost, nous constatons qu'il est le plus performant parmis les modèles testés. Par conséquent, il sera retenu pour la modélisation de la fréquence du portefeuille.

12 Construction du modèle de sévérité non grave

12.1 Application du GLM

L'objectif de cette partie est de modéliser le coût moyen des sinistres attritionnels. Pour ce faire, il y a deux distributions classiques permettant de le modéliser que sont : la loi gamma ou log-normal. Pour cette raison, un choix de la loi la plus adaptée pour décrire le montant de nos sinistres sera effectué en premier lieu et ensuite les résultats du modèle final seront présentés.

12.1.1 Choix de la loi de modélisation

Le graphe des QQ-plots

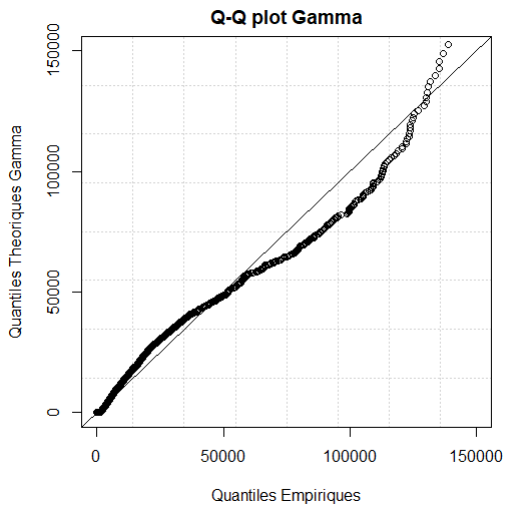

FIGURE 68 – QQ-plot Gamma et distribution empirique des montants des sinistres

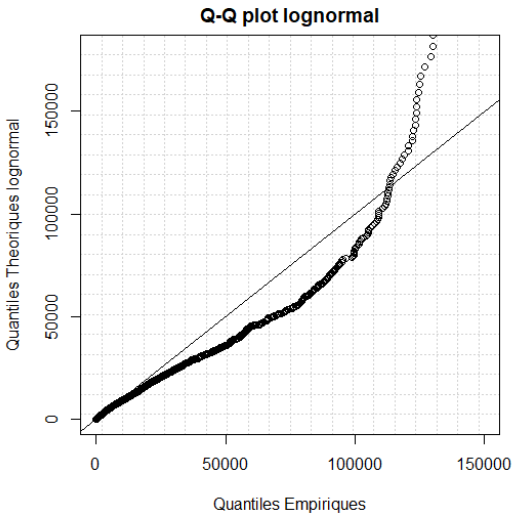

FIGURE 69 – QQ-plot Lognormal et distribution empirique des montants des sinistres

Au vu des graphes des qq-plots, la loi gamma est celle qui est la plus adaptée pour modéliser le coût des sinistres car elle a les points plus proches de la première bissectrice. De ce fait, pour conforter notre remarque graphique, nous allons faire appel à d'autres critères notamment le MSE (l'erreur quadratique moyenne) et la comparaison graphique entre les fonctions de répartition empiriques et théoriques. Voici les résultats obtenus :

indicateur	Gamma	Log-normal
$RMSE$	0.36%	0.42%

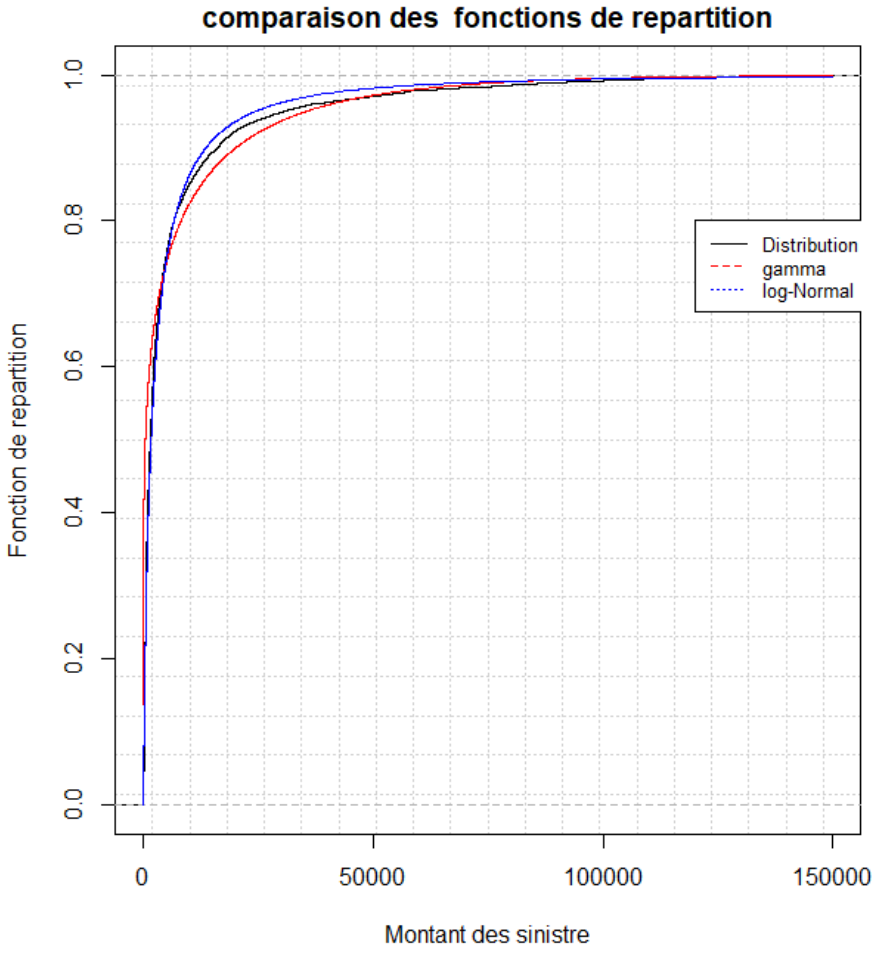

FIGURE 70 – comparaison des fonctions de répartition

Vu que le coût est l'aire au dessus de la fonction de répartition, on constate qu'avec la loi log-normal on a tendance à sous-évaluer le coût du sinistre. Cependant pour les sinistres à coût moyen, elle les évalue parfaitement. Par contre, la loi gamma a tendance à surévaluer un peu le montant des sinistres. Ainsi, compte tenu de la forte volatilité de la branche, la loi gamma a été choisie pour la modélisation des sinistres attritionnels pour plus de prudence.

Comme la loi pour la modélisation est choisie, nous procédons à l'application et à l'analyse des résultats. Le modèle définitif est obtenu en réitérant la même démarche faite pour la propension ou la fréquence. Ceci nous amène à retenir pour le modèle final, les variables explicatives suivantes : l'activité, la catégorie d'entreprise, la qualité de l'occupant, la garantie RC souscrite et l'engagement.

12.1.2 Analyse de la qualité du modèle

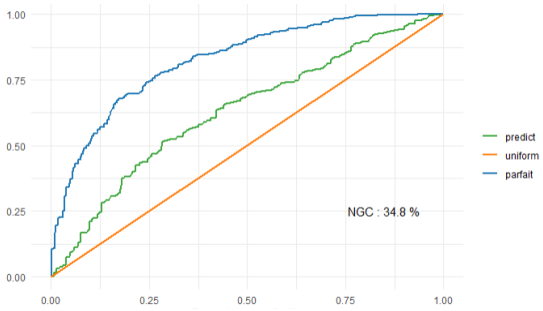

FIGURE 71 – Courbe de Lorenze sur la base train

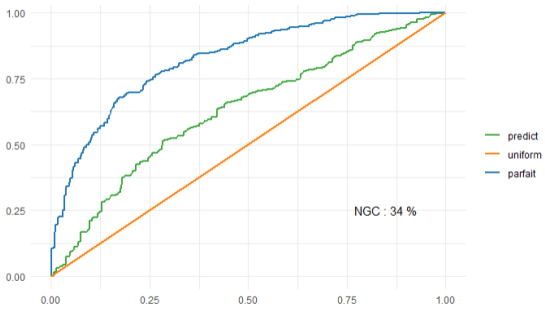

FIGURE 72 – Courbe de Lorenze sur la base test

On voit que l'indice de gini du modèle n'est pas élevé. Ceci peut être expliqué du fait qu'on a une forte variabilité des montants des sinistres dans notre base. Cependant, on aperçoit que le Gini de l'apprentissage est quasiment égale à celui du test, donc il n'y a pas eu de sur-apprentissage. Le tableau ci-dessous nous donne les indicateurs d'erreur de prédiction

indicateur	base train	Base test
MAE	7610	7748.13
ξ_{global}	4%	8%

Comparaison coût moyen Prédit et observé

Les graphes ci-dessous nous montrent les coût moyens prédits et observés par tranche d'engagement et par qualité de l'occupant.

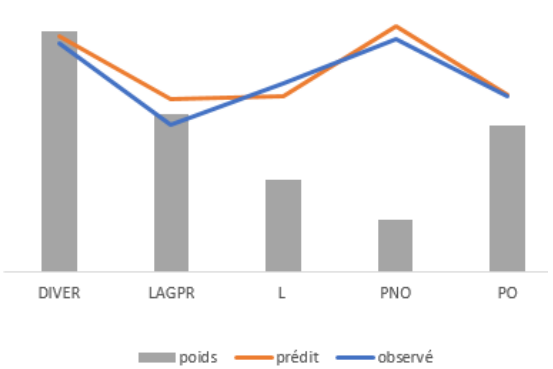

FIGURE 73 – Coût moyen prédit VS observé par qualité de l'occupant

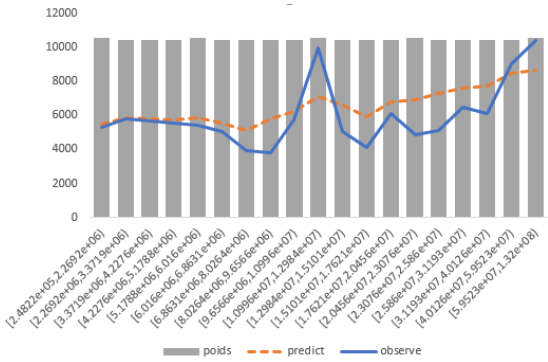

FIGURE 74 – Coût moyen prédit VS observé par engagement

On voit que sur la qualité de l'occupant, les coûts moyens prédits et observés ont la même tendance et sont assez proches. Concernant les engagements, les contrats qui ont des engagements moins élevés sont bien prédits en moyenne. Par contre ceux avec des engagements très élevés, le modèle suit à peu près la même tendance, mais plus lisse. Ainsi, de manière générale, le modèle n'est pas de mauvaise qualité. Dans le but d'améliorer notre qualité de prédiction, on a tenté d'ajuster un glm pénalisé. Cependant, en regardant nos indicateurs de performance, on s'aperçoit qu'il ne fait pas mieux que le glm simple. Voici les résultats du modèle :

indicateur	base train	Base test
MAE	7560	7700.13
ξ_{global}	45%	48%

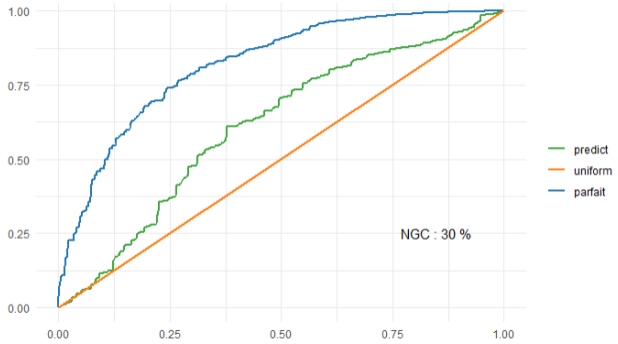

FIGURE 75 – Courbe de Lorenze sur la base test

12.1.3 Conclusion

Au final, le modèle de glm simple a été retenu car au global son erreur de prédiction n'est pas très importante et que la pénalisation n'améliore pas la qualité de prédiction.

12.2 Application des arbres CART et Random forest

12.2.1 Analyse de la qualité des modèles

Les résultats obtenus sur la base test après avoir calibré nos hyperparamètres par validation croisée pour contrôler le processus d'apprentissage sont donnés ci-dessous :

indicateur	random forest	CART
MAE	7800	8010
ξ_{global}	7.9%	8.9%

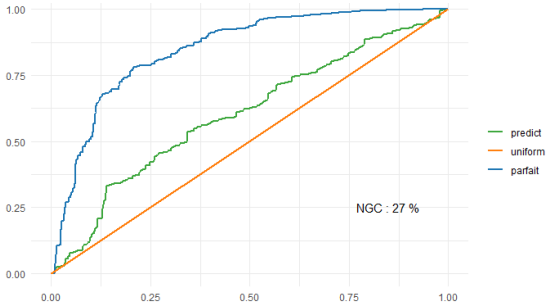

FIGURE 76 – Courbe de lorenze méthode CART

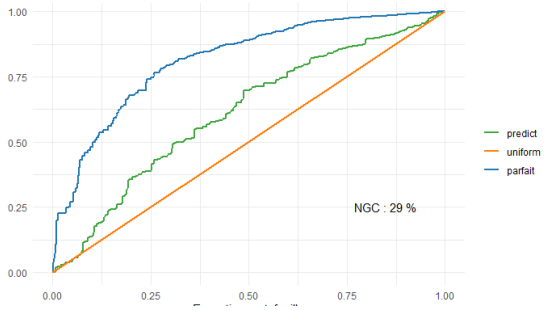

FIGURE 77 – Courbe de Lorenze sméthode random forest

Il apparaît clairement que les arbres de régression (CART et Forêts aléatoires) sont moins adaptés que la régression linéaire généralisée gamma pour la modélisation des coûts des sinistres dans notre cas. Notre glm gamma a cinq points de Gini plus que les arbres de régression. De plus ils ont une erreur de prédiction beaucoup plus importante que le glm notamment l'erreur absolue moyenne et le pourcentage d'erreur totale.

12.2.2 Conclusion

les arbres de régression (CART et Forêts aléatoires) sont moins performants que la régression linéaire généralisée gamma. Donc cette dernière sera conservé, pour le moment, pour la modélisation des coûts des sinistres attritionnels.

12.3 Application du modèle de Gradient Boosting

La calibration des hyperparamètres du modèle nous a permis de retenir un modèle final avec 400 arbres, une profondeur maximale de 5 et un coefficient de régularisation de 0.01.

12.3.1 Analyse de la qualité du modèle

L'évaluation du modèle final obtenu, après calibration des hyperparamètres, sur la base test nous donne les résultats suivants :

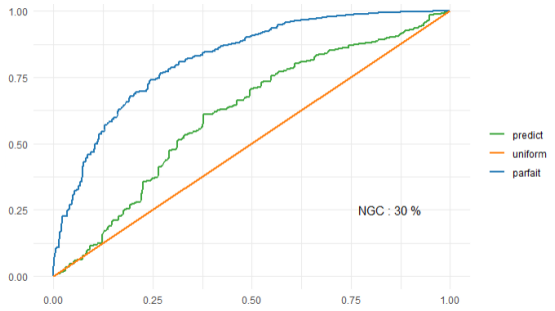

FIGURE 78 – courbe de Lorenze base test

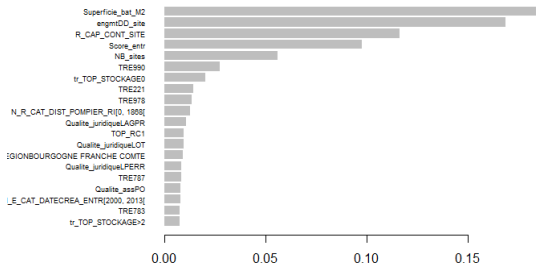

FIGURE 79 – variables significatives

indicateur	base train	Base test
MAE	6878.1	7110.3
ξ_{global}	5%	8.5%

Au regard des indicateurs de performance, on voit que le xgboost n'est pas meilleur que la régression linéaire généralisée. sur l'indice de gini, cette dernière à quatre points plus que la méthode xgboost. Sur ce graphe ci-dessous, nous comparons les coûts moyens prédits et observés sur l'ensemble des modèles de prédiction du coût moyen élaborés dans ce mémoire :

Comparaison de la fréquence prédite et observée

Les graphes ci-dessous nous montrent les coûts moyens prédits et observés par tranche d'engagement et par qualité de l'occupant sur l'ensemble des modèles mis en place.

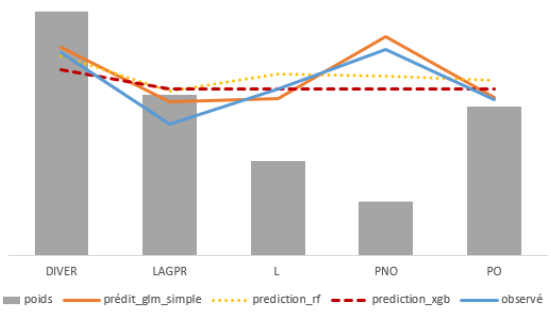

FIGURE 80 – graphe de comparaison des moyennes prédites et observées par Qualité de l'occupant

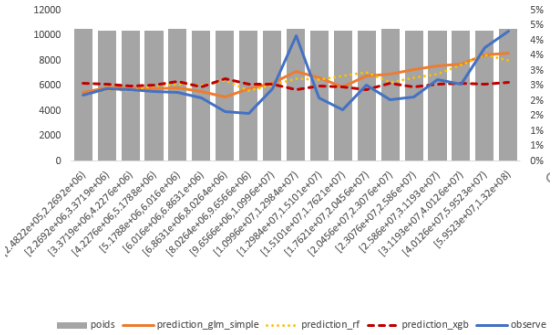

FIGURE 81 – graphe de comparaison des moyennes prédites et observées par tranche engagement

On constate que le glm de gamma à tendance de suivre la même allure que l'observé. C'est pour cela qu'il a un gini plus élevé que les autres méthodes. Il discrimine mieux la sévérité de nos sinistres. Par conséquent, nous gardons la régression linéaire généralisée comme notre modèle de sévérité attritionnelle pour la suite.

12.3.2 Conclusion

Après avoir analysé la qualité du modèle xgboost, nous constatons qu'il n'améliore pas le glm gamma. Par conséquent, ce dernier est retenu pour la modélisation de la sévérité attritionnelle du portefeuille.

12.3.3 Analyse du modèle de sévérité retenu

L'analyse des résultats des modèles de sévérité attritionnelle nous conduit à retenir la régression linéaire généralisée en tant que méthode d'estimation du coût moyen des sinistres hors graves. Toutefois, ce modèle n'est pas très performant au regard des indicateurs de performance et le graphe comparant les coûts moyens prédits et observés. De ce fait, il serait important de vérifier si l'augmentation de la taille de l'échantillon améliorera la capacité prédictive du modèle. Autrement dit s'il est nécessaire de chercher plus de données pour mieux calibrer nos coûts

de sinistres. Ainsi, cette étude sera faite grâce à la courbe d'apprentissage appelé **learning curve** en anglais.

courbe d'apprentissage

La courbe d'apprentissage est une méthode graphique qui permet d'analyser comment l'augmentation des observations améliore les performances d'un modèle en traçant les performances mesurées sur les données d'apprentissage et de test pour chaque sous-échantillon. Sur le schéma, l'axe des abscisses représente la taille du sous-échantillon et l'axe des ordonnées la métrique de performance.
Une grande augmentation des observations peut diminuer la variance du modèle, mais augmentera le biais, inversement. En effet, plus le biais est important, plus la variance est faible. Le meilleur modèle est celui qui arrive à trouver un bon compromis entre le bais et la variance. Les schémas ci-dessous illustrent bien le principe général du courbe d'apprentissage.

FIGURE 82 – courbe d'apprentissage meilleur modèle

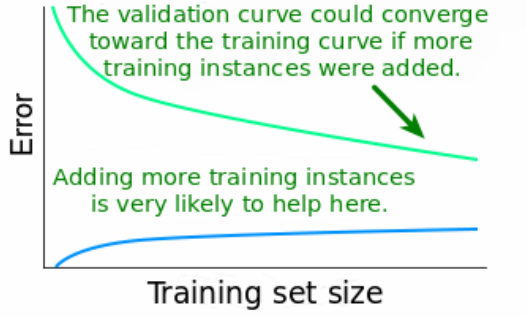

FIGURE 83 – courbe d'apprentissage modèle à améliorer

Sur la figure à gauche, on voit que les courbes d'apprentissage et de test ont convergé vers l'erreur optimale ; donc ajouter plus de lignes aux données d'apprentissage n'aidera vraiment pas à améliorer la qualité du modèle. Par contre sur le graphique à droite, on constate que la courbe de l'échantillon de test pourrait converger vers la courbe d'apprentissage. De ce fait, ajouter plus d'instances (lignes) pourrait améliorer la performance du modèle.
Cela étant dit, nous allons tenter de vérifier sur notre modèle de sévérité attritionnel, comment évolue notre courbe d'apprentissage.

Application au modèle de sévérité

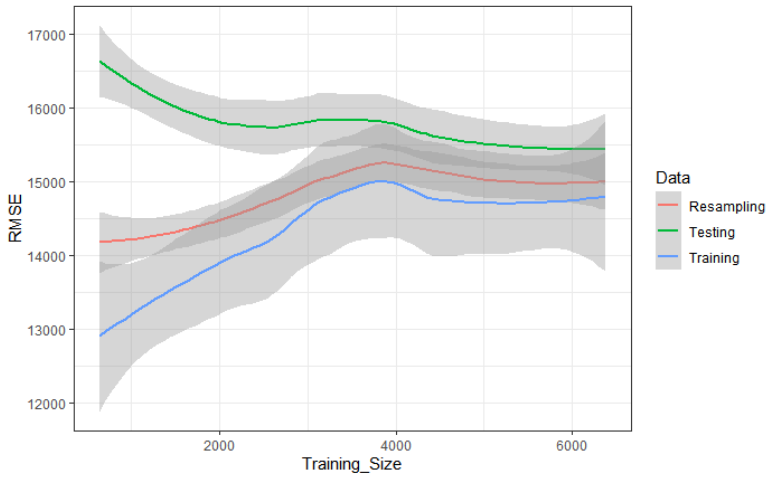

FIGURE 84 – Courbe d'apprentissage du modèle de coût moyen

Sur ce graphe, on constate que la courbe de l'échantillon de test baisse au fur et à mesure qu'on ajoute de données sur notre base d'apprentissage et tend vers la courbe servant à entraîner le modèle. Cependant, cette baisse commence à être très faible à partir de 5000 points d'apprentissage, ce qui implique que l'ajout de données supplémentaires peut ne pas beaucoup améliorer notre modèle.

En effet, le risque industriel est une branche très hétérogène donc la calibration de la sévérité des sinistres n'est pas chose facile. Donc aller collecter plus de données pourrait augmenter le biais du modèle, par conséquence la qualité du modèle pourrait se dégrader.

13 Synthèse

Après avoir rappelé les points essentiels de la théorie des modèles linéaires généralisés puis les arbres de CART et ses extensions (random forest, xgboost), nous avons construit un modèle de survenance de grave, un modèle de fréquence et un modèle de sévérité attritionnel en testant plusieurs méthodes. Ainsi pour chaque modèle nous nous sommes assuré de ne pas faire du sur-apprentissage en comparant nos indicateurs sur la base test et train (validation croisée). Ensuite, on a analysé la qualité des modèles finaux obtenus. D'après les résultats obtenus seule la fréquence sera modélisée par la méthode de xgboost et le reste sera modélisé par un glm.

14 Modélisation de la sévérité Grave

14.1 Arbres de régression de Pareto généralisés

La théorie des valeurs extrêmes nous dit que si X est une variable aléatoire de fonction de répartion F et u le seuil de grave. Si F appartient au domaine d'attraction de Fréchet alors la variable aleatoire $Y = X - u|X > u$ suit une distribution de Paréto généralisé. Par conséquent, il existe ξ et $\sigma(u)$ tel que :

$$P(X - u > y|X > u) = (1 + \xi \frac{y}{\sigma(u)})^{-\frac{1}{\xi}}$$

où

- $\sigma(u) > 0$ est le paramètre d'échelle du GPD,
- $\xi > 0$ est le paramètre de forme qui reflète la lourdeur de la queue de distribution

Comme la fonction de répartition est : $F_Y(y) = 1 - P(X - u > y|X > u)$ donc la fonction densité de $Y = X - u|X > u$ est donnée par :

$$f(y) = \frac{1}{\sigma(u)}(1 + \xi \frac{y}{\sigma(u)})^{-\frac{1}{\xi}-1}$$

L'espérence de $Y = X - u > y|X > u$ définie si $\xi \in]0, 1[$ est donnée par :

$$E[X - u > y|X > u] \quad = \quad \frac{\sigma(u)}{1-\xi}$$

Si $\xi > 1$ alors l'espérance est infinie donc elle n'est pas définie. De ce fait, nous perdons la notion d'assurabilité. Face à cette situation, nous avons deux choix : soit on exclut le risque, soit nous fixons une limite d'indemnisation à la police.

Étant donné que la sévérité des sinistres en risque industriel est très hétérogène même au niveau de la queue de distribution, ce qui peut amener à des pertes très différentes. Par conséquent, il est nécessaire de séparer les situations dans le but de ne pas mettre tous les cas de figure dans le même lot. Ceci permettra de voir s'il y a des risques non "assurables" (qui n'admettent pas d'espérance) dans notre portefeuille, mais aussi parmi les assurables quels sont ceux susceptibles d'avoir des

sinistres très graves. En effet, si nous mettons toutes les situations dans le même lot, alors le paramètre de forme ξ qui reflète la lourdeur de la queue de distribution qui sera estimé va être proche du pire des cas. Par exemple si nous mélangeons des risques "assurables" ($\xi_1 < 1$) et des risques non "assurables" ($\xi_2 > 1$) alors le ξ qui sera estimé va être proche de ξ_2 .

Ainsi, les arbres de régression de Pareto généralisés nous permettront de spécifier des typologies de classe dont la sévérité des sinistres est homogène. Cette méthode utilise les arbres de régression CART en modifiant le critère permettant de définir les divisions optimales de l'arbre par la perte quadratique de la log-vraissemblance du paréto généralisé. Pour toute observation $(Y_i - u, X_i)$ tel que $Y_i > u$ et X_i un facteur de risque, la log-vraissemblance du Paréto généralisé est définie comme suit :

$$\Phi(y, m(x)) = -log(\sigma(x)) - (1 + \frac{1}{\xi(x)})log(1 + \xi(x)\frac{y}{\sigma(x)})$$

où $m(x) = (\sigma(x), \xi(x))$ les paramètres du paréto généralisé à estimer.

Étant donné que l'objectif des arbres de régression, c'est de construire des nœuds de sorte que l'hétérogénéité soit maximale entre elles et minimale au sein d'un nœud. Nous obtenons au final, après élagage, les feuilles de l'arbre qui identifient des classes d'assuré, chacune correspondant à des comportements de queue différents c'est-à-dire avec des valeurs différentes de $m(x) = (\sigma(x), \xi(x))$.

14.2 Application

La théorie des valeurs extrêmes élaborée à la section 6 nous a permis de déterminer le seuil u au-dessus duquel l'approximation de la distribution de Paréto généralisé semble raisonnable. Le seuil de 150 000 retenu nous mène à garder environ 2,84% des sinistres survenus entre 2013 et 2021 pour la garantie incendie. Ainsi, l'application de la méthode d'arbres de régression de Pareto généralisés donne le résultat ci-dessous :

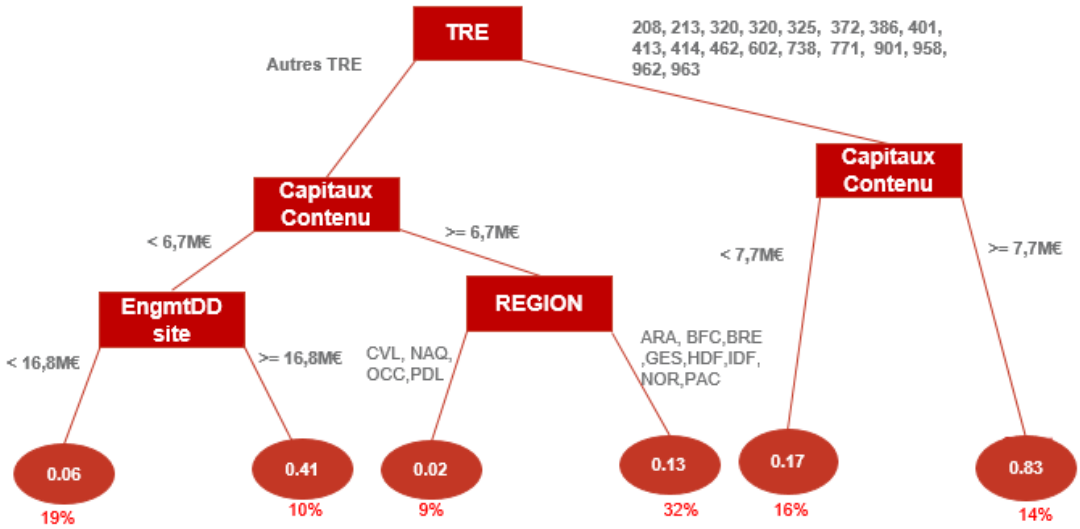

FIGURE 85 – Arbres de régression Pareto généralisés

Cet arbre shématise une classification des types d'assurés faite au niveau de la queue de distribution. Dans chacune des feuilles de l'arbre, le paramètre ξ indiquant la lourdeur de la queue et le paramètre d'échelle σ sont estimés en fonction des facteurs de risque (variables explicatives). De ce fait, on constate que l'ensemble des paramètres de forme estimés sont tous inférieurs à 1, ce qui nous conduit à dire que chaque classe de risque a une espérance finie. Cependant, certaines activités ayant des capitaux contenus supérieurs à 7,7M€ ont un indice de queue de distribution égale à 0,83 ($> 0,5$) donc leur variance est infinie. Cette classe d'assuré ne représente que 14% des événements extrêmes de notre jeu de données. Pour cette raison, il est tout à fait logique d'appliquer plus de restriction à ces groupes d'assuré, par exemple redéfinir la LCI (Limite contractuelle d'indemnisation) autrement dit le montant maximum indemnisé par l'assureur en cas de sinistre, ou éviter de beaucoup souscrire ces groupes d'assurés. Néanmoins, plus de 75% des événements extrêmes de la garantie incendie du risque industriel ont un paramètre d'échelle $\xi < 0.5$, ce qui est rassurant pour la santé financière de la branche.

Dans le but de vérifier que si cette méthode nous a permis de détecter certains groupes d'assurés moins contraignants que d'autres parmi les événements extrêmes de notre portefeuille, nous avons ajusté une régression de Pareto généralisé sur l'ensemble des observations au-delà du seuil. De ce fait, nous avons obtenu une estimation du paramètre d'échelle $\xi = 0.82$. Par conséquent, on remarque que si tous les événements extrêmes avait été mis dans le même groupe, on allait surévaluer la sévérité des sinistres sur certains types d'assuré. Cela mènera au risque d'anti-sélection, c'est-à-dire voir partir les risques les moins contraignants à gérer

et à garder ceux plus contraignants à gérer. Les paramètres σ et ξ estimés par la régression de Pareto généralisée au-delà du seuil u et les intervalles de confiance à 95% sont donnés sur le tableau ci-dessous :

feuille	ξ	$\sigma 10^{-}2$
feuille1	0.41 [0.01 ;0.8]	6578,74
feuille2	0.06 [0.002 ;0.32]	9942,6
feuille3	0.02 [0.001 ;0.3]	4622,04
feuille4	0.13 [0.02 ;0.3]	6905,09
feuille5	0.17 [0.01 ;0.52]	23978,33
feuille6	0.83 [0.4 ;1.3]	7113,03

À partir de ces paramètres estimés, nous pouvons calculer l'espérance pour chaque feuille de l'arbre, c'est-à-dire le coût moyen des sinistres pour chaque groupe d'assuré à partir de la formule donnée à la section 14.1. Ci-dessous, les variables les plus significatives du modèle :

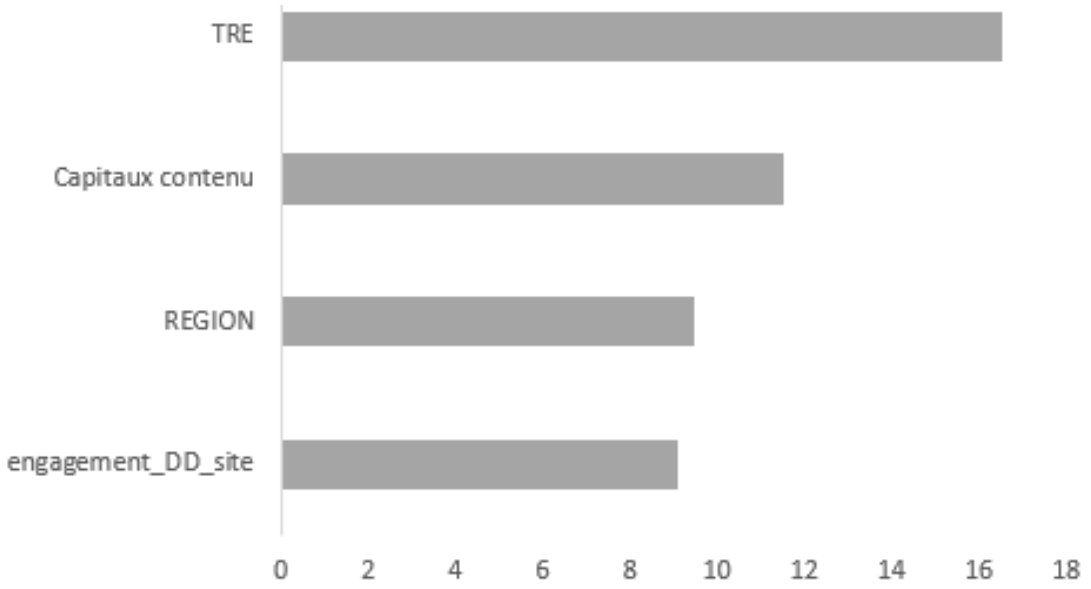

FIGURE 86 – Variables d'importance

15 Validation de la Méthode calibrées

Cette partie consistera à challenger la nouvelle méthode de tarification mise en place à celle existante. Rappelons que l'approche mis en place dans ce mémoire a pour objectif de mieux segmenter les assurés afin de faire payer pour chacun une prime correspondant à son niveau de risque. Étant donné que c'est la première fois qu'on élabore un tarif technique en risque industriel, de ce fait, la comparaison entre le tarif modélisé et celui qui est appliqué actuellement a été faite entre la prime commerciale et la prime pure chargée (prime pure modélisée ajouté des frais de gestion, de commission et de réassurance).

Pour ce faire, les contrats ayant vécu au moins un jour dans le portefeuille en 2021 ont été sélectionnés comme échantillon afin d'évaluer la capacité de segmentation du modèle en fonction de la taille du risque notamment l'engagement. Les indices de Gini calculés à partir de la courbe de Lorenze sont représentés sur les graphes ci-dessous :

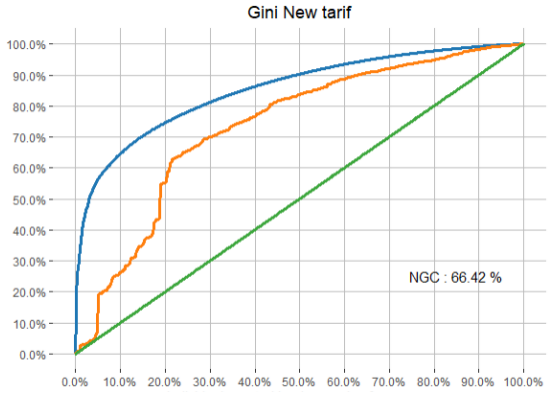

FIGURE 87 – courbe de Lorenze nouveau tarif technique

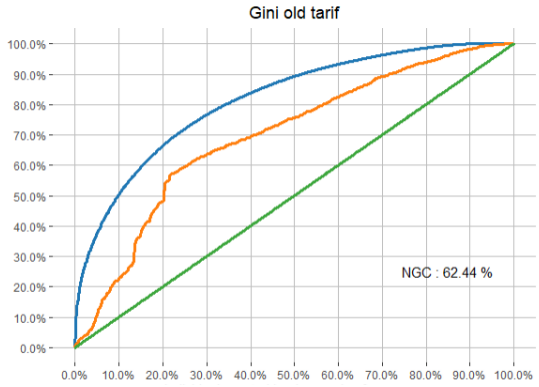

FIGURE 88 – courbe de Lorenze tarif actuel

Ces graphiques mesurent la répartition des primes en fonction des engagements. Le nouveau tarif technique mis en place a un Gini de 66,42% contre 62,44% le tarif actuel. De ce fait, la nouvelle méthode de tarification apporte une meilleure segmentation des risques. Or, tous les modèles servant à mettre en place cette méthode de tarification ont été robuste, par conséquent, la méthode calibrée peut être validée.

16 Conclusion

La garantie incendie représente plus de la moitié de la charge sur les neuf dernières années et la part des sinistres graves est de 88% (moins de 3% en nombre). De ce fait, l'actuaire doit consacrer une étude spécifique de ces événements extrêmes pour la mise en place d'un tarif technique permettant de contrôler les sinistres d'intensité. Ainsi, l'objectif de cette étude a été de mettre en relief une nouvelle méthode de tarification du risque industriel pour la garantie incendie permettant de prendre en compte les sinistres graves. L'étude s'est déroulée en trois grands axes :

- Le premier axe consistait à la construction et aux retraitements de la base de données permettant de réaliser notre étude,
- Le deuxième axe consistait de trouver le montant au-delà duquel on considère qu'un sinistre est grave. Cette étude a été faite grâce à la théorie des valeurs extrêmes et qui nous a conduit à retenir le montant de 150 000 comme seuil de grave.
- Et le dernier axe consistait à la modélisation permettant de calculer la prime pure.

Dans le troisième axe, nous avons mis en place un modèle de propension de grave, un modèle de fréquence, un modèle de sévérité attritionnel et un modèle de sévérité grave. Sur les trois premiers modèles, nous avons ajusté à chaque fois une régression linéaire généralisée et quelques méthodes d'apprentissage supervisées. L'analyse de nos résultats à travers des indicateurs de performance nous a mené à retenir la régression linéaire généralisée pour l'estimation de la propension et du coût moyen des sinistres puis la méthode xgbbost pour l'estimation de la fréquence.
Pour la modélisation de la sévérité grave, nous avons utilisé la méthode d'arbre de régression de Paréto généralisé qui estime le paramètre ξ indiquant la lourdeur de la queue et le paramètre d'échelle σ en fonction des caractéristiques de l'assuré. Cette approche permet de mieux comprendre les facteurs de risques impactant la sévérité des sinistres extrêmes et de les gérer de façon différente.

Étant donné que le risque industriel est caractérisé par une forte volatilité induit par les graves, cette approche de tarification permettant de contrôler les sinistres d'intensité semble être plus adaptée pour la branche. En plus de proposer une méthode de tarification, ce mémoire pourrait servir également comme outil permettant d'améliorer la politique de souscription et la gestion du portefeuille. En effet, l'arbre de régression de Pareto généralisé a construit des classes d'assurés en fonction de la sévérité des sinistres extrêmes du portefeuille. De plus, la modélisation de la propension de grave estime une probabilité d'avoir un sinistre

grave pour chaque assuré. En combinant ces deux études, nous pouvons créer des classes de risque et de déterminer les risques sur lesquels nous allons nous pencher pour souscrire beaucoup de contrats (les risques cibles) et ceux sur lesquels nous allons nous débarrasser ou diminuer dans le portefeuille.

La prochaine étape du travail consistera à industrialiser toute cette étude développée dans ce mémoire et de comparer le tarif obtenu à celui du marché dans le but de mesurer la cohérence de nos résultats par rapport à ce que propose le marché. Ensuite de généraliser les résultats sur l'ensemble des garanties soit par un taux de passage de la garantie principale (l'incendie) ou bien refaire la démarche sur les autres garanties d'intensité.

Certaines activités sont sous-représentées dans notre portefeuille notamment les fascicules 0, 1 et 8. De ce fait, cette insuffisance d'information de ces groupes d'activité pourrait biaiser les résultats leur concernant. Une solution à ce fléau est d'utiliser les résultats du marché pour les ajuster.
Une autre limite de cette étude concerne le seuil de grave utilisé sur lequel on s'est basé pour modéliser les probabilités de grave et la sévérité extrême. Par conséquent, les résultats obtenus sont sensibles à ce seuil. Pour cette raison, il serait intéressant d'étudier la sensibilité de l'étude par rapport à ce choix, c'est-à-dire comment varient les résultats lorsque nous choisissons un autre seuil candidat.
Et en fin, la non identification du site sinistré est aussi une limite de l'étude réalisée.

Références

[1] Arthur CHARPENTIER et Christophe DUTANG : L'Actuariat avec R.

[2] Maude THOMAs Econométrie de l'assurance non-vie : Cours ISUP

[3] Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance. Article from FARKAS, LOPEZ, THOMAS (2021) https://hal.archives-ouvertes.fr/hal-02118080v2/document

[4] Marie KRATZ (2022) Extreme Value Theory : Cours ISUP

[5] Théorie des valeurs extrêmes : Application au calcul de risque (thèse) https://www.researchgate.net/publication/334468946_Theorie_des_Valeurs_Extremes_Application_au_Calcul_de_Risques

[6] Terry M. THERNEAU Elizabeth J. ATKINSON Mayo FOUNDATION. "An Introduction to Recursive Partitioning Using the RPART Routines". In : (2015). https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

[7] Terry Therneau et Mayo Clinic "User written splitting functions for RPART". https://cran.r-project.org/web/packages/rpart/vignettes/usercode.pdf

[8] Laurent ROUVIERE https://lrouviere.github.io/machine_learning/cours_article.pdf

[9] Christine Malot-Tuleau : méthode CART https://math.unice.fr/~malot/CART.pdf

[10] Mémoire d'actuariat Fatima-zahra NAJI Nouvelle modélisation du risque extrême dans la tarification de la garantie incendie en assurance multirisques habitation.

[11] Fédération Française de l'Assurance : chiffres clés. https://www.franceassureurs.fr/nos-chiffres-cles/lassurance-de-dommages-et-responsabilite/

[12] Tutoriel : courbe d'apprentissage pour l'apprentissage automatique https://www.dataquest.io/blog/learning-curves-machine-learning/

Annexe

Distribution au-delà du seuil

Ici on cherche à décrire une variable aléatoire X dépassant un seuil u élevé. Ainsi, le théorème de Fisher-Tippet-Gnedenko nous dit que :

$$F^n(x) \approx \exp(-[1+\xi(\frac{x}{\sigma})]^{\frac{-1}{\xi}})$$

Donc

$$nlog(F(x)) \approx -[1+\xi(\frac{x}{\sigma})]^{\frac{-1}{\xi}}$$

Etant donné que nous nous intéressons aux valeurs de la queue de distribustion, on peut supposer que $F(x) \approx 1$. De ce fait, en utilisant le developpement de taylor d'ordre 1 de la fonction logarithme on a :

$$log(F(x)) = log(1-(1-F(x)) \approx -(1-F(x)) \; au \; voisinage \; de \; 0$$

Donc nous obtenons :

$$nlog(F(x)) \approx -n(1-F(x)) \approx -[1+\xi(\frac{x}{\sigma})]^{\frac{-1}{\xi}}$$

D'où

$$(1-F(x)) \approx \frac{1}{n}[1+\xi(\frac{x}{\sigma})]^{\frac{-1}{\xi}}$$

Or comme

$$P(X-u>y|X>u) = \frac{1-F(y+u)}{1-F(u)}$$

on a finalement

$$P(X-u>y|X>u) = (1+\xi\frac{y}{\sigma+\xi u})^{-\frac{1}{\xi}} = (1+\xi\frac{y}{\sigma(u)})^{-\frac{1}{\xi}}$$

Le terme de doite correspond à la fonction de survie de la loi de Paréto généralisée. D'ou

$$Y = X-u>y|X>u \sim GPD(\sigma(u),\xi)$$

Espérence mathématique au-delà du seuil

L'espérence de $Y = X - u > y | X > u$ définie si $\xi \in]0, 1[$ est donnée par :

$$E[X - u > y | X > u] = \frac{\sigma(u)}{1-\xi}$$

En effet,

$$\begin{aligned}
E[Y] &= \int_0^{+\infty} \frac{y}{\sigma(u)}(1 + \xi\frac{y}{\sigma(u)})^{-\frac{1}{\xi}-1} dy, \text{ on pose } z = \frac{y}{\sigma(u)} \\
&= \int_0^{+\infty} z(1+\xi z)^{-\frac{1}{\xi}-1} dz \\
&= \left[-z(1+\xi z)^{-\frac{1}{\xi}}\right]_0^{+\infty} + \sigma(u) \int_0^{+\infty} (1+)^{-\frac{1}{\xi}} dz \\
&= \sigma(u) \left[\frac{\xi}{\xi-1} \frac{1}{\xi(1+\xi z)^{\frac{1}{\xi}-1}}\right]_0^{+\infty} \\
&= \frac{\sigma(u)}{1-\xi}
\end{aligned}$$

d'où

$$E[X - u > y | X > u] = \frac{\sigma(u)}{1-\xi}$$

Estimation des paramètres GPD

1.Méthode du maximum de vraisemblance

Soit $Y_1, ..., Y_{N_u}$ un échantillon aléatoire des exès tel que $N_u = Card\{i : X_i > u\}$. L'estimateur par la méthode de maximum de vraisemblance consiste à maximiser la vraisemblance :

$$L(\sigma, \xi) = \prod_{i=1}^{N_u} f_{\sigma,\xi}(y_i)$$

où

$$f_{\sigma,\xi}(y_i) = \frac{1}{\sigma}(1 + \xi\frac{y_i}{\sigma})^{-\frac{1}{\xi}-1}$$

Etant donné que maximiser la vraisemblance revient à maximiser la log-vraisemblance, donc nous avons à résoudre les deux systèmes suivants :

$$\begin{cases} \frac{\partial log(L(y_1,...,y_{N_u},\sigma,\xi)}{\partial\sigma} &= 0 \\ \frac{\partial log(L(y_1,...,y_{N_u},\sigma,\xi)}{\partial\xi} &= 0 \end{cases}$$

$$\frac{\partial log(L(y_1,...,y_{N_u},\sigma,\xi)}{\partial\sigma} = 0 \Leftrightarrow \frac{\partial \sum_{i=1}^{N_u} -log(\sigma)-(\frac{1}{\xi}+1)log(1+\xi\frac{y_i}{\sigma})}{\partial\sigma} = 0$$

$$\frac{\partial log(L(y_1,...,y_{N_u},\sigma,\xi)}{\partial\xi} = 0 \Leftrightarrow \frac{\partial \sum_{i=1}^{N_u} -log(\sigma)-(\frac{1}{\xi}+1)log(1+\xi\frac{y_i}{\sigma})}{\partial\xi} = 0$$

Ainsi, nous obtenons le système d'équation suivant :

$$\begin{cases} \frac{\partial(-N_u log(\sigma)-(\frac{1}{\xi}+1)\sum_{i=1}^{N_u} log(1+\xi\frac{y_i}{\sigma}))}{\partial\sigma} = 0 \\ \frac{\partial(-N_u log(\sigma)-(\frac{1}{\xi}+1)\sum_{i=1}^{N_u} log(1+\xi\frac{y_i}{\sigma}))}{\partial\xi} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \frac{-N_u}{\sigma} - (\frac{1}{\xi}+1)\sum_{i=1}^{N_u}(\frac{-\xi y_i}{\sigma(\sigma+\xi y_i)}) = 0 \\ \frac{1}{\xi^2}\sum_{i=1}^{N_u} log(1+\xi\frac{y_i}{\sigma}) - (\frac{1}{\xi}+1)\sum_{i=1}^{N_u}(\frac{y_i}{\sigma+\xi y_i}) = 0 \end{cases}$$

Pour trouver la solution explicite de ce système, nous allons faire appel à des méthodes numérique, par exemple l'algorithma de Newton-Raphson.

2.Méthode des moments

Soit Y une variable aléatoire distribuée selon la loi $GPD(\sigma, \xi)$ avec $0 < \xi < 1$. Comme on sait que $E[Y] = \frac{\sigma}{1-\xi}$ et $V[Y] = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$ alors on a :

$$\begin{aligned} \frac{V[X]}{E[Y]^2} &= \frac{1}{1-2\xi} \\ \Leftrightarrow E[Y]^2 &= V[X](1-2\xi) \\ \Leftrightarrow \xi &= \frac{1}{2}\left[1 - \frac{E[Y]^2}{V[X]}\right] \end{aligned}$$

d'où

$$\hat{\xi} = \frac{1}{2}\left[1 - \frac{\bar{Y}^2}{\bar{Y^2} - \bar{Y}^2}\right]$$

et

$$\hat{\sigma} = \bar{Y}\left[\frac{1}{2} + \frac{\bar{Y}^2}{\bar{Y^2} - \bar{Y}^2}\right]$$

avec

$$\begin{cases} \bar{Y} = \frac{1}{N_u}\sum_{i=1}^{N_u} Y_i \\ V[Y] = \frac{1}{N_u - 1}\sum_{i=1}^{N_u}(Y_i - \bar{Y})^2 \end{cases}$$