Mémoire présenté le :

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

Par : ORNELLIA DJOFFON

Titre Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel

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

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

Membre présents du jury de l'Institut des Actuaires

signature

Entreprise :

Nom : Dorothée PAGES

Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :

Nom :

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Secrétariat

Signature du candidat

Bibliothèque :

Résumé

Mots clés : CART, bagging, boosting, forêt aléatoire, SCR, Modèle Interne, Formule standard, Solvabilité II, déséquilibre des classes, Arbre de décisions, Régression logistique

Pour modéliser un évènement dichotomique la méthode la plus courante est la régression logistique. Mais avec le développement des techniques d'apprentissage, les arbres de décisions sont de plus en plus utilisés. En assurance, lorsque les données de sinistralité présentent un déséquilibre des classes, les méthodes classiques de modélisation de la survenance de sinistres ne sont pas toujours adéquates.

Le déséquilibre des classes est une situation dans laquelle l'une des classes de la variable dichotomique est minoritaire. Dans ce cas de figure trouver un bon modèle pour modéliser cette variable n'est pas toujours aisé. Pour traiter le déséquilibre de classes observées sur les données de sinistralité, nous avons choisi de considérer ces deux approches. D'une part les techniques utilisées pour adapter le modèle de régression logistique à des données présentant un déséquilibre des classes et d'autre part les méthodes utilisées en apprentissage. Pour chaque approche, le meilleur modèle en termes de performances d'ajustement et de prédiction est retenu pour modéliser la survenance du sinistre.

L'étude de ces méthodes s'inscrit dans le cadre de la recherche d'un bon modèle de survenance de sinistre, ce qui s'avère primordial dans une démarche d'évaluation des risques de souscription par une approche modèle interne. En s'appuyant sur les modèles de survenance adaptés au problème d'asymétrie des classes, nous avons pu modéliser le risque de prime. Les autres risques de souscription sont également abordés au travers d'une approche modèle interne. Nous présentons le modèle interne partiel construit pour l'évaluation du capital de solvabilité requis pour les risques de souscription puis les résultats obtenus pour ce modèle sont comparés aux résultats en formule standard.

Abstract

Keywords: CART, bagging, boosting, random forest, SCR, Internal Model, Standard Formula, Solvency II, Class Imbalance, Decision Tree, Logistic Regression

To model a dichotomous event, the most common method is the logistic regression. With the development of learning techniques, decision trees are though increasingly used. In assurance, when claims data presents a class imbalance problem, the traditional methods used for modeling the occurrence of claims are not always adequate.

Class imbalance is a situation in which one of the classes of the dichotomous variable is a minority. In this case the total number of one class of the variable is far less than the total number of the other class. It's then complicated to find a good method to model this variable.

To deal with the imbalance of classes observed on the claims data, we have chosen to consider two approaches. On one hand we studied techniques used to adapt the logistic regression model to data presenting a class imbalance problem and on the other hand methods used in learning to deal with unbalanced data. For each approach, the model with the best predictive performance and adjustment quality is selected to model the occurrence of claims.

In the context of modeling underwriting risk through an internal model approach, the search for a good model for the occurrence of claim is central. With the model selected to correct the class imbalance problem, we have implemented the calculation of the premium risk. The other underwriting risks are also modeled by an internal model approach. Ultimately, we are proposing a partial internal model for the valuation of the solvency capital required for underwriting risks. The results obtained for this partial internal model are then compared with the results obtained with standard formula.

Synthèse

Sous la directive Solvabilité II, les compagnies d'assurance se doivent de disposer d'un niveau donné de fonds propres en représentation de leurs engagements. Pour déterminer le montant de capital requis, elles ont deux possibilités : utiliser la formule standard ou construire un modèle interne. En raison de sa capacité à prendre en compte les particularités de leur profil de risque, plusieurs assureurs ont opté pour la mise en place d'un modèle interne.

C'est dans ce contexte que notre assureur souhaite mettre en place un modèle interne partiel couvrant le périmètre des risques de souscription. Il s'avère que les données de sinistralité disponible sur le périmètre d'étude, présentent une asymétrie des classes. Afin de disposer d'un modèle représentatif des spécificités de la compagnie, cette particularité doit être considérée. C'est dans cette démarche que s'inscrivent les travaux présentés dans ce mémoire.

L'objectif est d'étudier les modèles permettant de traiter le déséquilibre des classes, d'appliquer ces modèles à nos données, de retenir les modèles les plus performants et de proposer une méthodologie d'évaluation des risques de souscription intégrant les modèles jugés les plus pertinents pour la modélisation de la survenance du sinistre.

Le déséquilibre des classes

Il est commun que les classes d'une variable binaires soient déséquilibrées. L'asymétrie des classes ne devient un problème que lorsque l'écart de proportion entre les classes est très important. Dans ce cas de figure, l'une des classes est très peu représentée et il devient alors délicat de bien modéliser la variable binaire. Dans la littérature, certains auteurs s'accordent à dire que lorsque la classe minoritaire représente moins de 5% des observations on est dans une situation de déséquilibre. Pour d'autres de seuil s'établi à partir de 10%. Toutefois, il est important de faire la distinction entre la situation que nous décrivons et une situation d'absolue rareté des observations de la classe minoritaire car ces deux problèmes sont traités différemment.

En ce qui concerne le déséquilibre des classes, plusieurs méthodes de correction sont proposées dans la littérature. Il s'agit en réalité de techniques ou d'algorithmes permettant d'adapter les modèles classiques de modélisation d'une variable binaire à une situation de déséquilibre des classes.

L'une des méthodes les plus utilisée pour décrire une variable binaire est la régression logistique. Mais avec l'essor des techniques d'apprentissage, les arbres de décisions en particulier l'arbre CART *(Classification and Regression Tree)* sont de plus en plus utilisés pour prédire les réalisations d'une variable binaire. Nous avons donc considéré ces deux approches et étudié les méthodes permettant de les rendre performants lorsque l'échantillon d'étude présente une asymétrie des classes.

En régression logistique

Plusieurs chercheurs se sont interrogés sur la légitimité d'utiliser une régression logistique classique dans le cas d'une base présentant un déséquilibre des classes. En effet, l'avantage principal lorsqu'on utilise le maximum de vraisemblance pour estimer les coefficients d'une régression est que l'on obtient des estimateurs possédant asymptotiquement des propriétés de convergence. Il s'avère qu'en cas de déséquilibre des classes les estimateurs peuvent être biaisés, leur variance très sensible aux observations de la classe minoritaire et les propriétés de convergence ne sont pas toujours vérifiées. En solution à ces problèmes, on retient les travaux de KING et ZENG qui proposent deux méthodes de correction de biais après rééchantillonnage à savoir l'ajustement préalable et la méthode par pondération. Pour l'ajustement préalable, la base de données initiale est rééquilibrée dans un premier temps. Ensuite, un modèle de régression logistique classique est appliqué à la nouvelle base obtenue. Les coefficients des variables explicatives obtenues sont conservés et le paramètre de constante estimé par la régression logistique, est ajusté par une formule donnée.

La seconde approche consiste à rééquilibrer la base de données initiale dans un premier temps puis à pondérer les données de façon à réduire l'impact de la différence observée entre la proportion de sinistres dans la base initiale et la proportion de sinistres dans la nouvelle base. Ensuite, la vraisemblance pondérée est maximisée pour estimer les coefficients de la régression logistique

En apprentissage

Lorsqu'on construit un classificateur sur une base présentant un déséquilibre des classes, les modèles manquent de précision et il peut arriver que le classificateur échoue à classer les observations. L'une des plus importantes contributions pour remédier au déséquilibre des classes est celle de WEISS. Dans ses travaux, il décrit un certain nombre de méthodes pour traiter l'asymétrie des classes notamment le rééchantillonnage, le boosting des algorithmes d'apprentissage et l'utilisation de métriques plus pertinentes pour évaluer les modèles. Parmi les méthodes de corrections existantes nous avons retenu deux approches permettant de pallier à la dissymétrie de nos données : l'approche par la base et l'approche par algorithme.

La première est la plus couramment utilisée. Elle consiste à rééquilibrer la base en jouant sur les proportions de positifs et de négatifs dans la base. Cette technique est indépendante du classificateur utilisé et de ce fait est plus souple et flexible. Il est donc possible d'équilibrer la base et d'y appliquer le classificateur de son choix.

La deuxième méthode appelée approche interne consiste à jouer sur les algorithmes de classification. L'idée est d'essayer de modifier ou d'adapter certaines caractéristiques des classificateurs de façon à améliorer leurs performances. Il s'agit des méthodes d'agrégation des classificateurs comme le *boosting*, le *bagging* et les forêts aléatoires.

Le point commun des méthodes de correction en régression logistique et en apprentissage est le rééchantillonnage de la base initiale.

Rééchantillonnage

Nous nous sommes intéressés à cinq méthodes de rééchantillonnage stratifiés selon la variable réponse. On distingue les méthodes classiques et les méthodes hybrides de rééchantillonnage. Les premières consistent à réduire ou à augmenter par sélections aléatoires le nombre d'observations d'une classe donnée. Il s'agit de l'*undersampling* (réduction des observations) et de l'*oversampling* (augmentation des observations). Ces deux méthodes conduisent à diminuer de façon considérable le déséquilibre entre les classes avec pour finalité, l'obtention de bases où la classe minoritaire est mieux représentée. L'inconvénient de ces méthodes peut être la perte d'information ou un excès d'information par rapport à la base initiale. Il existe des approches dites hybrides qui permettent de corriger ces méthodes et de tirer parti de leurs avantages respectifs. Parmi ces techniques, la méthode *Both sampling*, l'algorithme SMOTE *(Synthetic Minority Oversampling Technique)* et l'algorithme ROSE *(Random Oversampling Examples)* ont été étudiées. Le *both sampling* est une combinaison des approches *undersampling et oversampling.* Les algorithmes ROSE et SMOTE visent à construire un échantillon équilibré, en générant des données artificielles.

En nous basant sur ces techniques de rééchantillonnage nous avons construit des modèles de régression logistique et d'apprentissage. Dans un premier temps nous avons séparé la base initiale en base d'apprentissage et base de test. Les modèles sont construits sur la base d'apprentissage puis testés sur la base de test. La méthodologie adoptée est la suivante :

1. Rééchantillonnage de la base d'apprentissage avec les 5 techniques retenues à savoir l'*undersampling*, l'*oversampling*, le *Both sampling*, l'algorithme SMOTE et l'algorithme ROSE
2. En régression logistique, construction des deux modèles retenus à savoir l'ajustement préalable et la méthode par pondération sur chacune des 5 bases d'apprentissage. On obtient donc 10 modèles.
3. En apprentissage, construction de 4 types de modèles à savoir arbres CART, *bagging* d'arbres CART, *boosting* d'arbres CART, forêt aléatoire d'arbre CART sur chacune des 5 bases d'apprentissage. On obtient donc 20 modèles.
4. Evaluation de la qualité d'ajustement et de prédiction des modèles à partir de la base test avec les métriques appropriées à savoir l'indice de GINI, la courbe ROC, la précision, le rappel, la spécificité l'indicateur MCC, l'indicateur F-mesure, l'indicateur G-means.

Après évaluation de la qualité des modèles, nous avons retenu un modèle par approche.

En régression logistique le modèle retenu est le *rose pondéré*. Il est construit par un rééchantillonnage de la base par l'algorithme ROSE puis une application de la méthode de pondération de la régression logistique. En apprentissage le modèle retenu est le *bagging*

smote. Il est construit par un rééchantillonnage de la base avec l'algorithme SMOTE puis une application d'un modèle de *bagging* des arbres CART.

Forts de ces modèles qui prennent en compte les particularités des données, nous avons proposé un modèle interne partiel pour le calcul du SCR de souscription.

Les risques de souscription

Conformément à Solvabilité II, on peut identifier 4 types de risques de souscription sur le segment d'activité de la caution où opère notre assureur : le risque de prime, le risque de réserve, le risque catastrophe et le risque de rachat. Le risque de prime peut être défini comme le risque pour l'assureur que les primes soient inférieures aux coûts des futurs sinistres. Le risque de réserve est le risque que les provisions calculées soient insuffisantes en raison d'une mauvaise estimation des montants de paiement des sinistres, du niveau des recouvrements ou de leurs cadences. Le risque catastrophe est composé du risque catastrophe individuel et du risque catastrophe récession. Le risque catastrophe individuel est un risque de sévérité lié au fait que les gros clients aux encours les plus importants fassent défaut. Le risque catastrophe de récession est le risque de défaillance d'un grand nombre de clients à cause d'une situation de récession économique caractérisée par un ralentissement de l'activité immobilière. Le risque de rachat, peut être défini comme le risque lié à une dérive des taux de rachat sur les contrats offrant une possibilité de rachat anticipé avant la fin de la période de couverture ou de reconduction annuelle et unilatérale. Compte tenu des particularités du marché, nous avons modélisés 3 risques à savoir le risque de prime, le risque catastrophe individuel et le risque de réserve.

- Risque de prime

Le modèle construit est basé sur une modélisation de la survenance du sinistre par le modèle *rose pondéré* ou le modèle *bagging smote.* Il consiste en une simulation des entrées et sorties du portefeuille par des lois calibrées sur l'historique disponible pour construire le portefeuille d'assurance, une simulation de la sinistralité et une simulation des taux d'exposition pour déterminer les montants de charge générés par les sinistres futurs modélisés. Ces montants sont projetés selon une cadence de provisionnement puis les flux obtenus sont actualisés à l'aide la courbe des taux en vigueur.

- Risque catastrophe individuel

Sa modélisation a nécessité la définition d'un seuil de sinistre atypique. Nous avons utilisé le graphique des dépassements moyens en théorie des valeurs extrêmes pour fixer ce seuil. L'approche retenue pour la modélisation de ce risque est une approche du type $cout\ moyen * fréquence$.

Pour modéliser le nombre de sinistres atypiques, et les montants de sinistres atypiques, nous avons testé plusieurs lois sur l'historique de sinistralité. Nous retenons la loi

binomiale négative pour le nombre de sinistre et une loi gamma pour les montants de sinistres. A partir de ces lois, nous réalisons une simulation de la charge totale de sinistres de type catastrophe individuel

- Risque de réserve

Le modèle de provisionnement utilisé s'appuie en déterministe sur une méthode de type Chain-Ladder pondéré pour l'évaluation de la charge ultime et des cadences de règlement. En stochastique, la charge ultime est déterminée à partir d'un modèle basé sur une réplication par Bootstrap à 1 an. Ce modèle reprend les hypothèses de Mack. Il consiste à réaliser des tirages aléatoires avec remise, sur les résidus associés aux facteurs de développement, pour créer des pseudo-données. Afin d'obtenir une volatilité à un an, conformément à Solvabilité II, les simulations sont réalisées de telle sorte que la volatilité des distributions des paiements cumulés à un an contienne l'erreur de processus et l'erreur d'estimation.

Pour évaluer le niveau de capital requis, la valorisation économique des engagements de la CEGC est effectuée en scénario central en t=0 puis en t=1 sous forme de distribution après introduction d'une nouvelle année de production.

- A la date d'évaluation t=0, la valeur économique des engagements est déterminée comme l'espérance de la distribution de la somme actualisée espérée des flux futurs de trésorerie relatifs à ces engagements.
- La meilleure estimation des engagements en date t=1 est obtenu par la simulation des facteurs de risque et des flux de trésorerie associés sur la 1ère année de projection. Pour chaque simulation de 1ère année, on fait la projection des flux de paiements futurs et des frais de gestion conditionnellement aux réalisations de 1ère année et enfin on actualise ces flux futurs en t=1 pour l'évaluation de la valeur économique des engagements en fin de première période.

Les SCR primes, réserve et catastrophe individuel sont ensuite agrégés en utilisant la matrice de corrélation de la formule standard. L'application de la formule standard sur les données de l'assureur conduit aux résultats suivants :

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
6,82 M€	2,99 M€	119,23 M€	14,83 M€	-	120,98 M€

Pour l'approche par modèle interne, le risque de prime a été modélisé deux fois car nous avons considéré les deux modèles construits pour la survenance du sinistre attritionel. La mise en œuvre du modèle interne avec le modèle *rose pondéré* pour le calcul des probabilités de survenance aboutit aux résultats résumés dans le tableau suivant avec en dernière ligne la variation par rapport à la formule standard.

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
11,22M€	0,87 M€	40,56 M€	-	-	44,96 M€
65%	-71%	-66%	-100%	-	-63%

La mise en œuvre du modèle interne avec le modèle *bagging Smote* pour le calcul des probabilités de survenance aboutit aux résultats suivants :

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
25,45 M€	0,87 M€	40,56 M€	-	-	53,30 M€
273%	-71%	-66%	-100%		-56%

Le SCR de prime obtenu pour le *bagging smote* est plus du double du SCR prime obtenu pour le Rose pondéré. Le modèle de prime basé sur le *bagging smote* est encore plus coûteux pour la compagnie. Cela est dû au fait que les probabilités construites par les modèles en apprentissage sont plus élevées. Sur ce périmètre du risque de prime, la modélisation interne ne permet pas à la compagnie de réaliser de gains en capital par rapport à la formule standard mais lui permet d'avoir une vision plus précise du coût de ses engagements. En revanche, sur le risque catastrophe individuel et le risque de réserve on observe un gain important par rapport à la formule standard. Au global le SCR souscription en modèle interne est inférieur au SCR souscription en formule standard. La compagnie aurait donc tout à gagner en mettant en place un modèle interne que ce soit sur la maîtrise en interne de ses risques que sur le coût en capital de ses engagements.

Synthesis

Under the Solvency II Directive, insurance companies must have a given level of own funds to represent their commitments. To determine the amount of required capital, they have two options: use the standard formula or build an internal model. Because of its ability to take into account the specificities of its risk profile, several insurers have opted for the implementation of an internal model.

It is in this context that our insurer wants to set up a partial internal model covering the scope of underwriting risks. It turns out that the loss data available on the perimeter of this study shows a class imbalance problem. To have a model representative of the specificities of the company, this particularity must be considered. It is in this process that the work presented in this thesis takes place.

The aim is to study the models used to deal with class imbalance problem, to apply these models to our data, to retain the best performing models and to propose a methodology for the evaluation of underwriting risks, integrating the models judged more relevant for modeling the occurrence of claims.

What's class imbalance problem

It is a common problem in machine learning where the total number of a class of data is far less than the total number of another class of data. It is common for the classes of a binary variable to be unbalanced. The asymmetry of classes becomes a problem only when the difference in the proportion between classes is very large. In this case, one of the classes is not well represented and it then becomes difficult to model the binary variable. In the literature, some authors agree that when the minority class represents less than 5% of the observations, we are in a situation of imbalanced data problem. For others from 10% onwards, we are already in situation of imbalanced class problem. However, it is important to make the distinction between the situation we describe and a situation of absolute rarity in the minority class, as these two problems are treated differently.

Several methods are proposed in the literature to correct the class imbalance problem. It is generally techniques or algorithms to adapt the classical modeling method for a binary variable to a situation of imbalanced class.

One of the most commonly used methods for describing a binary variable is logistic regression. Though, with the development of learning techniques, decision trees, in particular the CART tree (Classification and Regression Tree), are increasingly used to predict the value of a binary variable. We have therefore considered these two approaches and studied methods to make them perform better when the study sample presents a class imbalance problem.

Logistic regression approach

Several researchers have questioned the legitimacy of using a classical logistic regression in the case of imbalanced data. Indeed, the main advantage when using the maximum likelihood to estimate a regression's coefficients is the asymptotical convergence properties of the estimators. It turns out that in the case of class imbalance the estimators may be biased, their variance are very sensitive to the observations of the minority class and convergence properties are not always verified. In order to solve these problems, KING and ZENG propose two methods of bias correction after resampling, namely the prior correction and the weighting method.

For the prior correction, the initial database is rebalanced at first. Then a classical logistic regression model is applied to the new basis obtained. The coefficients of the explanatory variables obtained are retained and the constant parameter estimated by the logistic regression is adjusted by a given formula.

The second method is to rebalance the initial database in a first step and then to weigh the data in order to reduce the impact of the difference observed between the proportion of claims in the initial base and the proportion of claims in the new database. Then the weighted likelihood is maximized to estimate the logistic regression coefficients.

Learning approach

When constructing a classifier on an unbalanced data, models lack sensitivity and the classifier may fail to categorize the observations. One of the most important contributions to correct class imbalance problem was introduced by WEISS. In his work, he describes a number of methods to deal with class asymmetry including resampling, boosting learning algorithms and using more relevant metrics to evaluate models. Among the existing corrections methods we have chosen two approaches to solve the problem of class imbalance: the approach by the base and the approach by algorithm.

The first is the most commonly used. It consists in rebalancing the base by playing on proportions of positives and negatives in the base. This technique is independent of the classifier used and is therefore more flexible and adaptable.

The second method also called internal approach is to play on the classification algorithms. The idea is to try to modify or adapt certain characteristics of classifiers in order to improve their performance. It is for example methods of aggregating classifiers such as boosting, bagging, and random forests.

The common point of methods of correction in logistic regression and in learning is the resampling of the initial base.

Resampling

We considered five methods of response based sampling. There are two categories of resampling method: conventional methods and hybrid methods. The first consists in reducing or increasing by random selections the number of observations of a given class. These include under sampling (reduction of observations) and oversampling (increase of observations). These two methods lead to a considerable reduction in the imbalance between the classes, with the aim of obtaining bases where the minority class is better represented. The disadvantage of these methods can be the loss of information or an excess of information compared to the initial base. Hybrid methods try to correct this inconvenient. These techniques include the "both" sampling method, the SMOTE (Synthetic Minority Oversampling Technique) algorithm and the ROSE (Random Oversampling Examples) algorithm. Both sampling is a combination of undersampling and oversampling methods. The ROSE and SMOTE algorithms aim to build a balanced sample, by generating artificial data.

Using these resampling techniques we have implemented logistic regression and learning models. In a first step, we have separated the initial data into training data and test data. The models are built on the training data and then tested on the test data. The methodology used is as follows:

1. Resampling of the training data with the five techniques chosen: under sampling, oversampling, Both sampling, the SMOTE algorithm and the ROSE algorithm
2. For logistic regression, application of the two models chosen (the prior correction and the weighting method) on each of the five learning bases. We obtain 10 models.
3. For learning, application of four models (CART trees, CART tree bagging, CART tree boosting and CART tree random forest) on each of the five training data. We obtain 20 models.
4. On the data test, assessment of each model adjustment quality and prediction performance using appropriate metrics such as GINI index, ROC curve, accuracy, recall, MCC indicator, specificity, F indicator -measure, G-means indicator.

After evaluating each model's quality, we have selected one model per approach. In logistic regression, the model chosen is the *weighted rose*. It is constructed by resampling the data using the ROSE algorithm and then applying the method of weighted logistic regression. In learning, the chosen model is the *bagging smote*. It is constructed by resampling the data with the SMOTE algorithm and then applying a bagging model of the CART trees.

Using these models, which took into account the specificities of the data, we proposed a partial internal model for the underwriting SCR calculation.

Underwriting risks

In accordance with Solvency II, 4 types of underwriting risks can be identified in the business segment of surety: premium risk, reserve risk, catastrophe risk and lapse risk. Premium risk can be defined as the risk to the insurer that the premiums are less than the costs of future claims. The reserve risk is the risk that the calculated provisions are insufficient due to an incorrect estimate of the claims payment amounts, the level of recoveries or their rates. Catastrophe risk is composed of individual catastrophe risk and recession catastrophe risk. Individual catastrophe risk is a severity risk due to the fact that large customers with the largest exposure amounts are defaulting. The recession catastrophe risk is the risk of failure of a large number of customers due to a situation of economic recession characterized by a slowdown of the real estate activity. Lapse risk may be defined as the risk associated with a drift in the redemption rates on early redemption contracts before the end of the unilateral or annual renewal period. Given the particularities of the market, we have modeled three risks: premium risk, individual catastrophe risk and reserve risk.

• Premium risk

The model implemented is based on the claims occurrence modeling using the *weighted rose* or the *bagging smote* model. It consists of the portfolio's inflows and outflows simulation following a distribution calibrated on the available history, a simulation of the loss experience and a simulation of the exposure rates to determine the amount of expenses generated by future claims modeled. These amounts are projected according to a provisioning rate and the cash flows are discounted using zero coupon rate.

- Individual catastrophe risk

Its modeling requires the definition of an atypical loss threshold. We use the mean excess plot in extreme value theory to set this threshold. The approach used for modeling this risk is an $average\ cost * frequency$ approach. To model the number of atypical claims and the amounts of atypical claims, we tested several laws on the historical claims. We retain the negative binomial distribution for the number of claims and a gamma distribution for the amounts of claims. Using these distributions, we perform a simulation of the total gross claims for the individual catastrophe risk.

- Reserve risk

The deterministic provisioning model is based on the weighted Chain-Ladder method for the evaluation of the ultimate gross claims and settlement rates. Stochastically the ultimate load is determined with a model based on 1-year Bootstrap replication. This model use Mack's assumptions. It consists in performing random data drawings with replacement, on the residues associated with the development factors, in order to create pseudo-data. In order to achieve one year volatility, in accordance with Solvency II, the

simulations are performed in such a way that the volatility of the cumulative one-year payment distributions contains the process error and the estimation error.

In order to evaluate the level of capital required, the economic valuation of the company's commitments is carried out in a central scenario in t = 0 and then in t = 1 in the form of a distribution after the introduction of a new year of production.

- At the valuation date t = 0, the economic value of the commitments is determined as the expectation of the distribution of the discounted sum of the future cash flows relating to these commitments.
- At the valuation date t = 1, the best estimate of commitments is obtained by simulating risk factors and associated cash flows over the first year of projection. For each first year simulation, future payments and management costs are projected on the basis of the fulfilment of the first-year assumptions, and the future cash flows are discounted at t = 1 to calculate the economic value of commitments at the end of the first period.

The reserve SCR, the premium SCR and individual catastrophe SCR are then aggregated using the correlation matrix of the standard formula. The application of the standard formula on the insurer data leads to the following results:

Premium SCR	Reserve SCR	Individual catastrophe SCR	Recession catastrophe SCR	Lapse SCR	Underwritting SCR
6,82 M€	2,99 M€	119,23 M€	14,83 M€	-	120,98 M€

For the internal model approach, the premium risk was modeled twice because we considered the two models selected for the occurrence of the attritional loss. The implementation of the internal model with the *weighted rose* model for the calculation of the occurrence probabilities leads to the results summarized in the following table. The last row of the table shows the variation compared to the results obtained with the standard formula.

Premium SCR	Reserve SCR	Individual catastrophe SCR	Recession catastrophe SCR	Lapse SCR	Underwritting SCR
11,22M€	0,87 M€	40,56 M€	-	-	44,96 M€
65%	-71%	-66%	-100%	-	-63%

The implementation of the internal model with the *smote bagging* model for the calculation of probabilities of occurrence leads to the following results:

Premium SCR	Reserve SCR	Individual catastrophe SCR	Recession catastrophe SCR	Lapse SCR	Underwritting SCR
25,45 M€	0,87 M€	40,56 M€	-	-	53,30 M€
273%	-71%	-66%	-100%		-56%

The premium SCR obtained for *bagging smote* is more than twice the SCR premium obtained for the *weigthed rose*. The premium model based on *bagging smote* is even more expensive for the company. This is because the probabilities constructed by the learning models are higher. Within this perimeter of risk, internal modeling does not allow the company to realize gains regarding solvency capital compared to the standard formula but offers significant insight on risk exposure. On the other hand, for individual catastrophe risk and the reserve risk, there is a significant gain compared to the standard formula. This compensates the loss in premium fields so that the underwriting SCR calculated in internal model is less than the underwriting SCR in standard formula. So the company would make gain by setting up an internal model both in terms of solvency capital requirement and in the knowledge and management of its own risks.

Remerciements

Je voudrais remercier Mr Mansour SOW pour m'avoir offert l'opportunité d'effectuer mon alternance au sein de CEGC.

Un merci particulier à ma tutrice Mme Dorothée PAGES pour avoir encadré ce mémoire. Sa disponibilité, ses conseils avisés, ses nombreuses explications et toutes ses connaissances qu'elle n'a pas hésité à partager ont rendu ce mémoire possible.

Je voudrais exprimer toute ma gratitude à l'équipe du modèle interne pour leur gentillesse, leur écoute, pour m'avoir permis de travailler dès le départ dans des conditions remarquable où la rigueur et le sérieux requis par la profession, n'empêchent pas une certaine convivialité.

Je remercie également ma tutrice académique Mme Ying JIAO pour ses conseils.

Pour finir, mes pensées vont à ceux qui m'ont tendu la main.

SOMMAIRE

Introduction

Dans le cadre de son activité de cautionnement et de garantie aux administrateurs de biens et aux agents immobiliers, la compagnie européenne de garantie et caution est soumise à un risque de défaut qu'elle se doit de maîtriser et d'anticiper. L'EIOPA[1] à travers la directive Solvabilité II impose à la compagnie d'avoir des fonds propres à la hauteur de ses engagements pour pouvoir faire face en cas de sinistralité. Pour calculer le niveau de capital requis en représentation des risques de souscription, la compagnie a le choix entre une approche par la formule standard et une approche par modèle interne. La compagnie souhaite opter pour un modèle interne car cette approche permettrait de tenir compte des particularités de l'activité et de son propre profil de risque.

C'est dans ce cadre général que s'inscrivent nos travaux. Dans le cas particulier du risque de prime, l'un des facteurs de risque principaux dont doit tenir compte la compagnie est la survenance du sinistre. Il s'avère que les données de sinistralité disponibles sur ce périmètre d'activité souffrent d'un problème de déséquilibre des classes. Dans ce cas, la régression logistique qui est la méthode classique utilisée pour modéliser un évènement dichotomique est inadaptée. Il en est de même pour les arbres binaires de classification CART de plus en plus utilisés pour prédire la réalisation d'une variable binaire. Toutefois, il existe des méthodes alternatives et des techniques de modélisation utilisées pour traiter le problème du déséquilibre des classes. Deux approches ont été étudiées, l'une basée sur une adaptation de la régression logistique et l'autre basée sur une adaptation et une amélioration des performances des arbres de décisions CART. L'enjeu étant de construire un modèle adéquat pour la survenance du sinistre, ces méthodes seront étudiées, appliquées aux données puis leurs performances seront comparées.

A terme l'objet est le développement d'un modèle interne en vue du calcul du capital de solvabilité requis. Certains travaux avaient déjà été réalisés en ce sens pour le calcul du SCR[2] de réserve. Nous avons repris ces travaux et nous les avons complétés par une modélisation des autres risques de souscription à savoir le risque de prime et le risque catastrophe individuel. Le but étant de challenger la formule standard, le niveau de capital requis par le modèle interne partiel construit est comparé aux montants de SCR en formule standard. Le mémoire sera structuré comme suit : dans une première partie, nous commencerons par présenter le contexte des travaux à savoir la directive Solvabilité II, l'activité de la compagnie, et les données utilisées pour nos travaux. Dans une seconde partie, nous présenterons le problème du déséquilibre des classes et les méthodes utilisées pour traiter ce problème. A partir des résultats obtenus sur nos données nous retiendrons un modèle par approche et ces modèles seront intégrés à la modélisation du risque de prime. Dans une troisième partie, nous présenterons le modèle interne partiel et les résultats obtenus par cette approche seront comparés à ceux de la formule standard.

[1] European Insurance and Occupational Authority

[2] Solvency Capital Requirement (capital de solvabilité requis)

PARTIE 1: CONTEXTE GENERAL

Chapitre 1: Les exigences réglementaires de la directive solvabilité II

Afin d'éviter une crise comme celle observée par le passé dans certaines compagnies d'assurance, les autorités européennes ont mis en place un certain nombre de règles prudentielles visant à s'assurer de la solvabilité des compagnies d'assurance. L'ensemble de ces règles prudentielles établies par l'EIOPA[3] ont été regroupées sous le nom de Solvabilité II.

En France, la mise en œuvre de ces règles est régie par l'Autorité de Contrôle et de Résolution (ACPR).

1. Présentation de la directive

1.1 Enjeux de la directive

La directive Solvabilité II est entrée en vigueur le 01/01/2016 et s'impose à toutes les compagnies d'assurance. Elle remplace la directive Solvabilité I dont les règles prudentielles étaient jugées insuffisantes par les autorités européennes.

En effet, sous Solvabilité I, le profil de risque de l'entité n'est pas pris en compte et les postes du bilan sont évalués à partir de leur valeur nette comptable. Dans la directive Solvabilité II le profil de risque de l'entreprise est intégré dans les différentes exigences réglementaires et les postes du bilan sont valorisés de façon économique. C'est-à-dire que les actifs sont valorisés suivant leur valeur de marché et les engagements selon leur meilleure estimation future. Le but de cette directive est de renforcer la protection des assurés en exigeant des sociétés d'assurances d'avoir des fonds propres à la hauteur de leurs engagements.

Ainsi Solvabilité II préconise d'identifier les risques, de les quantifier et de dégager le montant de fonds propres nécessaire pour faire face à ces risques. Le respect de ces exigences passe par le calcul de deux niveaux de fonds propres : le capital de solvabilité requis (SCR) et le minimum de capital requis (MCR[4])

1.2 Architecture de la directive Solvabilité II

La directive Solvabilité II, comprend plusieurs règles prudentielles regroupées sous trois piliers.

Le premier pilier est dit quantitatif. Il concerne les exigences quantitatives, notamment en matière de fonds propres et de calculs des provisions techniques. Il s'agit essentiellement du calcul du :

[3] European Insurance and Occupational Authority
[4] Minimum capital requirement

- **SCR** que l'on peut définir comme le niveau de fonds propres requis pour éviter à la compagnie une ruine dans un an due au choc provoqué par une sinistralité exceptionnelle avec une probabilité de 99,5%.

- **MCR** que l'on définit comme le niveau de fonds propres en dessous duquel l'autorité de contrôle intervient automatiquement.

Le second pilier est qualitatif. Il concerne les exigences en matière d'organisation et de gouvernance des organismes. Il a pour but de fixer des normes qualitatives de suivi des risques en interne aux sociétés et comment l'autorité de contrôle doit exercer ses pouvoirs de surveillance dans ce contexte.

Le troisième pilier est informationnel. Il s'agit de déterminer d'une part l'ensemble des informations auxquelles aura accès le public et d'autre part l'information destinée aux autorités de contrôle.

2. Évaluation des postes du bilan économique

Comme évoqué dans le paragraphe précédent, sous Solvabilité II, les actifs en représentation des engagements de l'entreprise sont évalués selon leur valeur de marché et les passifs sont calculés par une approche meilleure estimation. Dans le bilan économique, ils sont scindés en deux grands postes : les fonds propres et les provisions techniques. Ces dernières sont constituées de la marge pour risque et de la meilleure estimation des engagements de l'entreprise. Ainsi, le bilan économique sous Solvabilité II, peut être représenté par le graphique suivant :

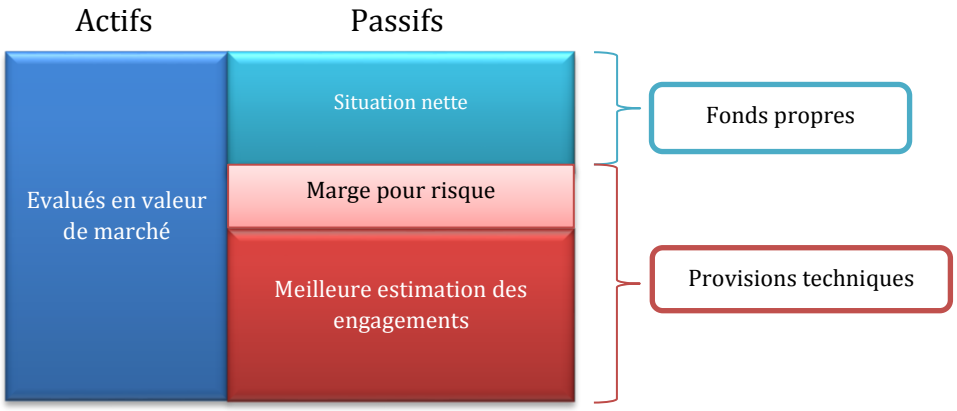


Figure 1: Le bilan économique de solvabilité II

2.1 La meilleure estimation des passifs d'assurance

Dans les spécifications techniques, la directive européenne Solvabilité II défini, la meilleure estimation comme suit :

« *La meilleure estimation est égale à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle probable des flux de trésorerie futurs), déterminée à partir de la courbe des taux sans risque pertinente* »

Concrètement, c'est la valeur actuelle probable attendue de tous les flux de trésorerie futurs en tenant compte des informations dont on dispose concernant le portefeuille d'assurance et de réassurance. Encore appelée Best Estimate, elle se note généralement *BE*.

Le calcul de la meilleure estimation des passifs d'assurance est une étape importante dans la modélisation des risques de souscription et en particulier pour le calcul du capital économique. Le dispositif Solvabilité II exige le calcul des provisions selon une vision « meilleure estimation ». On cherche donc, à évaluer au plus juste, les engagements de la compagnie et pour cela il est nécessaire d'identifier et de modéliser au préalable plusieurs facteurs de risques.

2.2 La marge pour risque

La marge pour risque fait aussi partie des provisions techniques dans le bilan Solvabilité II. Autrement dit, c'est le montant complémentaire au Best Estimate, pour que le montant de provisions technique global soit en adéquation avec les engagements de la compagnie. Elle fait intervenir la notion de coût de capital. C'est le capital jugé nécessaire pour qu'un assureur accepte de reprendre le portefeuille tel qu'il est avec les engagements qu'il implique en supposant qu'il existe un marché parallèle d'échange des portefeuilles d'assurances. On fait l'hypothèse que l'assureur qui détient le portefeuille n'as pas suffisamment de fonds en représentation de ses engagements et va donc essayer de céder ce portefeuille à une autre entité. La marge pour risque se calcule à partir du SCR par la formule suivante :

$$Risk\ Margin = TC * \sum_{t\geq 0} \frac{SCR_t}{(1+r_{t+1})^{t+1}},$$

Avec :

- TC : *taux lié au coût du capital présumé égal* à 6%
- SCR_t : *capital de solvabilité requis pour la date t*
- r_t: *taux sans risque correspondant* à *la date t*

2.3 Les fonds propres

Les fonds propres constituent la troisième composante au passif du bilan économique. Encore appelés Net Asset Value(NAV), ils se définissent comme l'écart entre la valeur de l'actif et les provisions techniques.

3. Le capital de solvabilité requis du module souscription en assurance non-vie

L'une des exigences quantitatives de la directive est le calcul du capital de solvabilité requis. Encore appelé SCR, il peut être défini comme le montant de fonds propres nécessaire à l'assureur pour faire face aux pertes potentielles liées aux évènements imprévus à un an avec une probabilité de 99,5%.

Conformément à la directive, le calcul du SCR peut se faire par deux approches différentes : la formule standard ou le modèle interne.

3.1 La formule standard

La formule standard est une technique de calcul du capital de solvabilité requis (SCR) proposé par l'EIOPA. Elle vise à prendre en compte les risques importants et quantifiables auxquels les compagnies d'assurance sont soumises. Le SCR couvre les risques quantifiables liés aux contrats en cours mais aussi aux contrats dont la souscription est anticipée dans les douze prochains mois. La formule standard se calcule à partir des volumes de primes et de réserves, catastrophe et de rachat auxquels on applique des coefficients fixés en fonction du type de risque à évaluer. Ces SCR sont ensuite agrégés selon une matrice de corrélation dont les coefficients ont été fixés. Une formule de ce type est pratique et facile à mettre en place mais par construction elle ne prend pas en compte le profil de risque et les spécificités de la compagnie qui la met en œuvre.

3.2 Le modèle interne

En modèle interne, deux types d'approches existent pour le calcul du capital de solvabilité requis: le modèle interne intégral où tous les risques sont modélisés par le moteur de calcul de la compagnie et le modèle interne partiel pour lequel seule une partie des risques est modélisée par le moteur de calcul avec plusieurs simulations et le reste des risques est calculé avec la formule standard. Cette méthode implique la construction d'une distribution de Fonds Propres à un an puis l'évaluation d'un quantile à partir de cette distribution. Le management de la compagnie souhaite mettre en place sur le marché des ADBAI, un modèle interne car elle dispose de suffisamment d'historique pour construire un modèle de calcul du SCR qui prend en compte les particularités de son activité. Pour évaluer ce niveau de capital requis, la valorisation économique des engagements de la CEGC est effectuée en scénario central en t=0 puis en t=1 sous forme de distribution après introduction d'une nouvelle année de production.

- A la date d'évaluation t=0, la valeur économique des engagements est déterminée comme l'espérance de la distribution de la somme actualisée espérée des flux futurs de trésorerie relatifs à ces engagements.
- Le BE des engagements en date t=1 est obtenu par la simulation des facteurs de risque et des flux de trésorerie associés sur la 1ère année de projection. Pour chaque simulation de 1ère année, on fait la projection des flux de paiements futurs et des frais de gestion conditionnellement aux réalisations de 1ere année et enfin on actualise ces flux futurs en t=1 pour l'évaluation de la valeur économique des engagements en fin de première période.

Chapitre 2: Impact de solvabilité II et problématique

1. L'entreprise et son activité

La Compagnie Européenne de Garantie et Caution est spécialisée dans les métiers de la garantie et de la caution. Du fait de son activité, elle est soumise aux exigences règlementaires de Solvabilité 2. C'est l'une des 15 branches de l'assurance visées par la directive et elle concerne en particulier le segment d'activité « credit and surety ».

1.1 L'activité de cautionnement

L'activité de cautionnement ou de garantie financière peut être définie comme un acte par lequel le garant s'engage à honorer l'exécution ou le respect des termes d'un contrat en cas de défaillance du débiteur (personne cautionnée) vis-à-vis du créancier (bénéficiaire de la caution). Elle a pour principal objectif de fournir aux créanciers une garantie de paiement d'une dette et de ce fait instaure la confiance dans les relations commerciales et sécurise les transactions. Elle fait intervenir 3 principaux acteurs : le garant de la caution, le cautionné et le bénéficiaire de la caution.

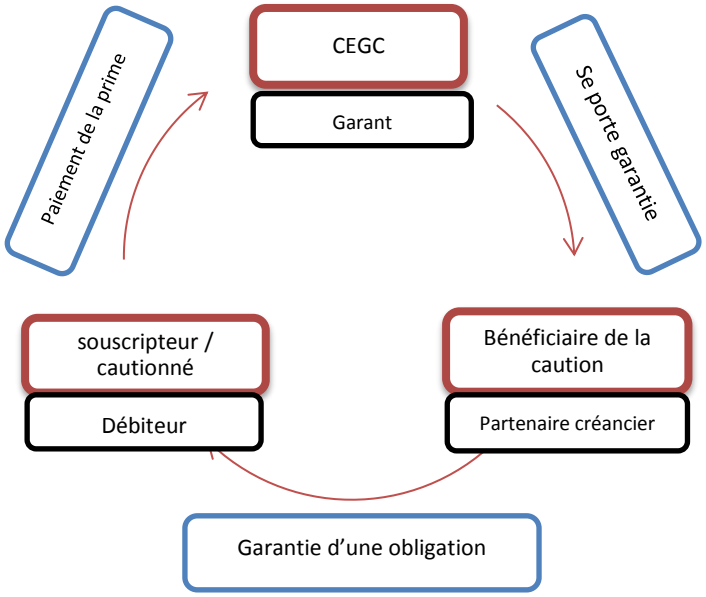


Figure 2: L'activité de cautionnement

Le marché de la caution est aujourd'hui principalement réservé à des sociétés spécialisées telles que les banques, sociétés de cautions mutuelles et compagnies d'assurance. La compagnie propose une offre assez diversifiée qui s'adresse à différents acteurs de l'économie allant du cautionnement des crédits immobiliers pour les particuliers à la garantie aux activités des administrateurs de biens et agents immobiliers.

Compte tenu de la diversité de l'offre de garantie proposée, le calcul des exigences de capital s'effectue par périmètre d'activité. Les travaux réalisés dans le cadre de ce mémoire portent sur le marché des administrateurs de biens et agents immobiliers (ADBAI)

1.2 La garantie sur le marché des administrateurs de biens et agents immobiliers

La garantie sur le marché des ADBAI est destinée à deux types de souscripteurs : les administrateurs de biens et les agents immobiliers. Ce sont tous les deux des professionnels de l'immobilier qui sont contraints par la loi de fournir des garanties financières dans le cadre de leur activité professionnelle. En effet depuis la loi HOGUET du 02/01/1970, les administrateurs de biens et les agents immobiliers sont obligés pour l'exercice de leur activité et l'obtention d'une carte professionnelle de fournir une garantie financière.

La CEGC propose deux types de garanties financières spécialisées. Il s'agit de la garantie de gestion pour les administrateurs de biens et de la garantie de transaction pour les agents immobiliers. A travers ses deux garanties spécialisées, la CEGC s'engage sur le montant maximum des fonds dont le professionnel de l'immobilier est redevable envers ses clients (en général il s'agit du montant le plus élevé sur l'exercice précédent, appelé pic de trésorerie). En contrepartie d'une certaine prime réglée annuellement, la compagnie s'engage sur une année renouvelable, à payer les mandants, de l'agent immobilier ou de l'administrateur de biens en cas de défaut de ceux-ci.

2. Problématique du mémoire

La garantie aux activités des administrateurs de biens et agents immobilier, engendre un aléa lié aux défauts de ceux-ci. La particularité du marché des ADBAI fait que la fréquence de sinistre observé est faible. En effet le sinistre survient en cas d'impossibilité pour l'ADBAI de restituer à ses mandants les fonds détenus pour leur compte. Il s'agit donc d'un risque de détournement de fonds de tiers. Dans le cadre de leurs activités, les administrateurs de biens ou agents immobiliers possèdent à la fois leur propre compte de gestion et le compte de gestion des fonds appartenant à leurs créanciers. Cette précaution permet de réduire considérablement le risque de sinistre. Et quand bien même l'ADBAI serait en défaut, on observe dans la majorité des cas que celui-ci ne va pas puiser de ressources dans les fonds des tiers. Sur un historique de plus de 16 ans, le taux de sinistralité observé est de 0,203%.

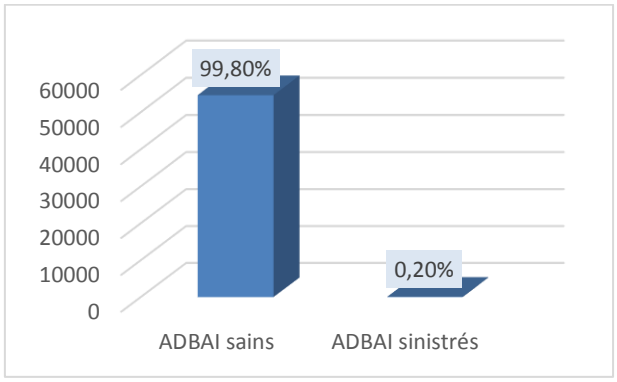


Figure 3: Proportion de sinistres sur l'historique

Face à une aussi faible proportion de sinistre, l'enjeu est de trouver un modèle adéquat pour modéliser la variable de sinistralité.

2.1 Modélisation de la survenance du sinistre en cas de déséquilibre des classes

Prédire la survenance d'un sinistre revient à modéliser un évènement possédant deux modalités. La méthode la plus couramment utilisée est la régression logistique. Avec l'essor des méthodes d'apprentissage, les arbres de décisions binaires sont de plus en plus utilisés pour prédire les évènements dichotomiques. Ces deux méthodes permettent de prévoir l'appartenance d'un individu à l'une ou l'autre des classes d'une variable dite cible en se basant sur les réalisations des variables explicatives. Mais ces méthodes sont sensibles en cas de déséquilibre des classes de la variable cible.

Le déséquilibre des classes correspond à une situation dans laquelle l'une des classes de la variable cible est minoritaire par rapport à l'autre modalité. Dans ce cas, les méthodes classiques n'offrent pas toujours des résultats satisfaisant. Pour y remédier, plusieurs auteurs ont proposé des méthodes dites de correction. Il s'agit en fait de techniques qui viennent en complément des méthodes classiques de façon à les adapter à des données présentant un déséquilibre des classes. Dans le cadre de ce mémoire, ces techniques seront présentées, testées et comparées en considérant une application sur la base de données des ADBAI, dont la variable de sinistralité présente clairement un déséquilibre des classes.

2.2 Application dans le cadre de la modélisation interne du SCR de souscription

Le deuxième objectif de ce mémoire est de s'appuyer sur les techniques de modélisation de la survenance du sinistre pour proposer un modèle interne partiel du risque de souscription permettant de challenger la formule standard. En effet, pour quantifier le besoin en fonds propres d'une société, il est nécessaire de prendre en compte et d'étudier les différents risques auxquels elle pourrait être confrontée. Les risques de souscription sont les risques spécifiques qui résultent du contrat d'assurance. Sur le marché des ADBAI, ce sont l'ensemble des risques liés au contrat de cautionnement ou de garantie et qui proviennent de l'incertitude de l'assureur quant aux conséquences des souscriptions.

Conformément à solvabilité II, on peut identifier 4 types de risques de souscription sur le segment d'activité de la caution : le risque de prime, le risque de réserve, le risque catastrophe et le risque de rachat.

Le risque de prime peut être défini comme le risque pour l'assureur que les primes soient inférieures aux coûts des futurs sinistres. Le risque de réserve est le risque que les provisions calculées soient insuffisantes en raison d'une mauvaise estimation des cadences de paiement des sinistres, du niveau des recouvrements ou de leurs cadences. Le risque catastrophe est composé du risque catastrophe individuel et du risque catastrophe récession. Le risque catastrophe individuel est un risque de sévérité lié au fait que les clients aux encours les plus importants fassent défaut. Le risque catastrophe de récession est le risque de défaillance d'un grand nombre de clients à cause d'une situation de récession économique caractérisée par un ralentissement de l'activité immobilière dû à une situation de crise immobilière par exemple. Le risque de rachat, peut être défini

comme le risque lié à une dérive des taux de rachat sur les contrats offrant une possibilité de rachat anticipé avant la fin de la période de couverture ou de reconduction annuelle et unilatérale.

Identifier les facteurs engendrant ces risques permet de les modéliser et de déterminer le montant du SCR souscription par une approche modèle interne. Après cette première partie d'introduction et de présentation des données, la deuxième partie de ce mémoire traitera de la modélisation de la survenance du sinistre en comparant plusieurs méthodes d'estimation. Enfin la dernière partie traitera du calcul du SCR avec une proposition de modèle interne pouvant challenger la formule standard.

Chapitre 3: Analyse des données

Les données constituent le socle de toute étude statistique et actuarielle. L'objectif de ce chapitre est de présenter un ensemble de statistiques descriptives permettant l'analyse de notre historique de données et en particulier de quelques variables jugées pertinentes pour la suite de notre étude. Nous commencerons par présenter la répartition de notre portefeuille d'assurés vis-à-vis des variables étudiées telles que la sinistralité et le type de contrat. Nous analyserons ensuite, l'évolution de ce portefeuille au fil des années d'études. Dans une troisième partie, à partir de statistiques descriptives et en ayant à l'idée le but d'identifier les variables pouvant expliquer la survenance d'un sinistre, nous étudierons plus en détails certaines variables de notre base de données.

1. Présentation de la base de données

La base de données disponible regroupe les informations relatives aux contrats souscrits entre 1999 et 2015. Les variables de cette table peuvent être scindées en deux groupes : les variables concernant le contrat de garantie et les variables relatives aux sinistres en historiques. Dans un premier temps, nous présenterons quelques données relatives au contrat, l'une des plus importantes étant le score. Pour une bonne compréhension, nous reviendrons sur la méthode de construction de cette variable. Dans un second temps, nous nous intéresserons aux données liées aux sinistres. La variable de sinistralité a fait l'objet d'un retraitement, nous en exposerons les raisons.

1.1 Les données relatives au contrat

Parmi les données relatives au contrat on peut citer entre autres :

- L'identifiant de l'ADBAI: il s'agit d'un numéro attribué au client (ADBAI)
- La date d'entrée : il s'agit de la date à laquelle ce dernier a souscrit au contrat
- La date d'enregistrement au registre du commerce : il s'agit de la date de création de l'entreprise
- Le montant d'encours : il s'agit du montant sur lequel la compagnie s'engage dans le cadre de ce contrat
- Le segment : il s'agit du type de client avec deux possibilités, agent immobilier(AI) ou administrateurs de biens (ADB).
- L'âge : cette variable est construite en faisant la différence entre l'année d'étude et l'année de création de l'entreprise.
- L'ancienneté : cette variable est obtenue en faisant la différence entre l'année d'étude et l'année de première souscription de l'ADBAI
- Les scores : il s'agit de notations attribuées à chaque client. On en distingue 4 classes.

Dans les paragraphes suivants, nous présenterons et analyserons plus en détail les quatre dernières variables. Nous avons choisi de nous attarder sur ces variables car elles pourraient expliquer la survenance du sinistre.

1.2 La construction des variables de score

Dans leur quasi-totalité, les organismes financiers utilisent l'analyse statistique pour prédire si dans le cadre d'une caution ou garantie, une contrepartie va ou non honorer son engagement et prendre ensuite la décision appropriée : acceptation ou refus de garantie. Selon les méthodes de scoring, la construction de scores et la décision d'octroi de garantie se fondent sur l'observation d'informations du passé relatives à la contrepartie.

L'idée du scoring est d'organiser les éléments dont on dispose en classes. Le scoring peut être défini comme un ensemble de méthodes qui aboutissent à un classement d'individus au sein de groupes préalablement définis. On cherche à prévoir le groupe auquel appartient un individu donné sur la base des observations qu'il présente vis-à-vis des variables considérées. Ceci implique la définition a priori d'un certain nombre de groupes ou classes de scores et de certaines règles basées sur les caractéristiques des individus. À partir de ces règles, un individu sera affecté dans l'une ou l'autre des classes de scores. Les variables de scores ont été construites par les analystes. Chaque classe de score correspond à une notation codifiée par une couleur et reflète un certain niveau de risque. Les notations sont attribuées annuellement en s'appuyant sur les données chiffrées telles que les rapports d'audit, le dernier bilan et le compte de résultat, les notations externes telles que les notations Coface, les notations de la Banque de France et d'autres sources d'informations diverses et variées telles que l'environnement c'est-à-dire la qualité financière de société mères (holding) ou société sœurs.

C'est la combinaison de ces éléments qui permet de définir le niveau de risque associé à un dossier. Nous disposons en tout de quatre classes de scores représentées par des couleurs :

- La classe verte
- La classe orange
- La classe rouge
- La classe Watch List
- La classe NC (non classé)

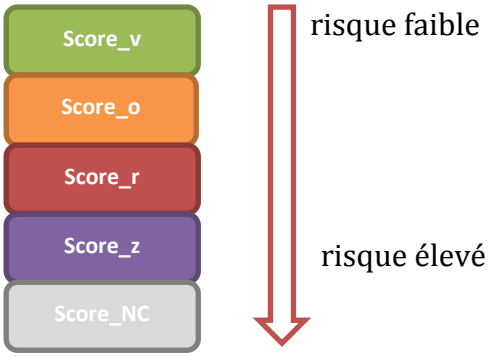


Figure 4: Les variables de score

Un ADBAI **vert** est un dossier ayant une bonne situation financière et dont les différentes notations sont plutôt correctes et convergent vers une capacité de la contrepartie à honorer ses engagements financiers. En définitive c'est un dossier présentant peu de risque financier.

Un ADBAI **orange** est un ADBAI pour lequel l'un ou l'autre des indicateurs financiers est dégradé et/ou dont le rapport d'audit laisse voir que l'ADBAI fait face à des difficultés

financières. Ici, on peut qualifier le risque de moyen. L'orange englobe toute contrepartie ne présentant ni un risque faible, ni un risque élevé. Cette couleur caractérise une dégradation du dossier qui est incertaine dans la durée.

Un ADBAI **rouge** est un ADBAI tel que le montant de fonds propres de l'ADBAI est négatif et la cotation de la Banque de France ou l'une des autres notations est très dégradée. C'est un dossier pour lequel, la très mauvaise situation financière laisse présager un risque très important. Cette classe de score englobe tous les dossiers présentant un risque de défaut très élevé.

Un ADBAI en **Watch List (Z)** est un ADBAI présentant un risque important et qui est placé en procédures collectives du type redressement Judiciaire, liquidation judiciaire.

Un ADBAI est **non classé (NC)** lorsque les informations dont dispose l'analyste concernant son dossier ne permettent pas de le classer dans l'une des catégories précédentes. L'absence de classement d'un ADBAI intervient en général en début de vie du contrat le liant à CEGC.

On remarque que les critères utilisés pour déterminer les classes de scores sont certes basés sur des indicateurs financiers mais en général la décision prise concernant la notation d'un ADBAI est essentiellement basée sur l'expertise au cas par cas de l'analyste et sa connaissance du marché. Cette notation n'est pas forcément fiable puisqu'il y a des dossiers bien notés par les métiers mais dont le risque associé s'avère important.

Bien que l'analyse financière au cas par cas soit plus fine qu'une analyse statistique, cette approche purement financière présente des limites. La construction d'un modèle statistique à partir des indicateurs financiers et avec validation des métiers, permettrait une meilleure connaissance du risque sur ce marché par les équipes actuarielles. Par ailleurs, l'étude des scores permet de remarquer la présence d'un nombre important de dossiers sans notation.

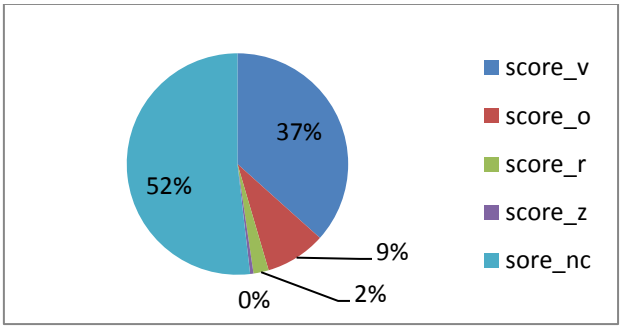


Figure 5: Répartition des dossiers par classe de scores

En fait, 52% soit plus de la moitié des dossiers n'est pas noté. Ce nombre important de dossiers non classés rend l'utilisation de cette variable peu fiable et pourrait créer un biais dans les résultats obtenus.

La répartition des dossiers par classes de scores au fil des années est représentée ci-dessous.

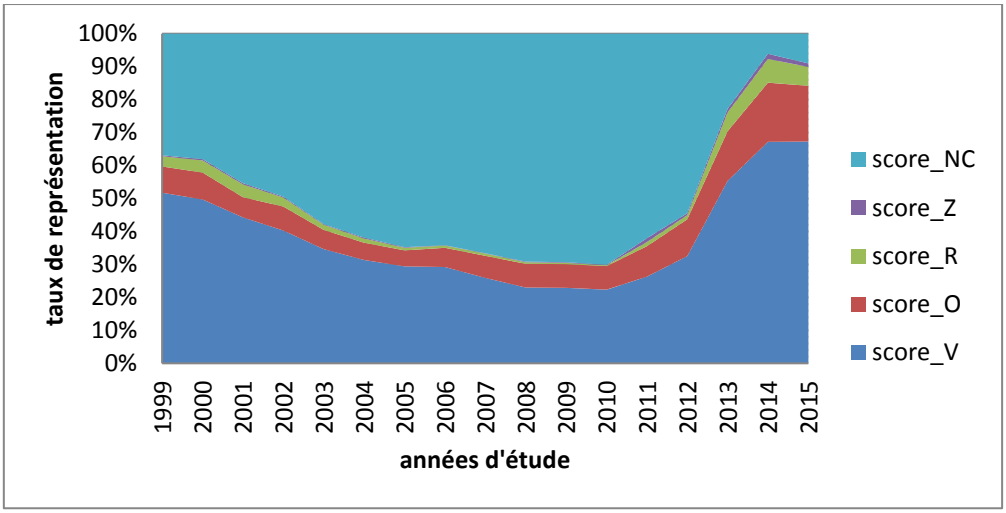


Figure 6: Taux de représentation des scores par année

À partir de 2012, le nombre de dossiers non classés connait une baisse importante à la faveur des autres classes. Des résultats plus pertinents pourraient être obtenus en ne considérant que les dernières années de notre historique. Cependant, il faudrait alors se contenter d'une profondeur d'historique très faible et se priver de données sans doute pertinentes. Il s'agit alors de trouver le meilleur compromis entre profondeur d'historique suffisante et fiabilité des données. Compte tenu de la faible proportion de sinistres dans cette base de données, nous avons fait le choix de considérer l'historique remontant à 1999. Pour les études futures, cet historique pourra être progressivement réduit.

1.3 Les données relatives au sinistre

Parmi les données relatives aux sinistres on peut citer entre autres :

- La variable indicatrice de la présence d'un sinistre : elle précise la présence ou non d'un sinistre en historique sur un dossier considéré.
- Une variable indicatrice du type de sinistre : elle précise si le sinistre survenu chez l'ADBAI est un vrai ou faux sinistre. Rappelons que la notion de vrai ou faux sinistre est liée au fait que la garantie financière soit ou pas en engagée dans le cadre de la survenance d'un sinistre. Si la garantie financière est mise en jeu, on parle de vrais sinistres sinon il s'agit de faux sinistres.
- Un numéro d'identification du sinistre : il s'agit d'un numéro attribué au sinistre dans la base de données des sinistres.
- La date de survenance du sinistre : c'est la date à laquelle la compagnie a enregistré le sinistre.
- Le montant de paiement des sinistres : il s'agit du montant payé par la compagnie dans le cadre du sinistre survenu.

1.4 Retraitement des données relatives au sinistre

L'un des paramètres les plus importants dans le cadre de la modélisation du risque de prime est la probabilité de survenance d'un sinistre. Afin de modéliser la probabilité de survenance, nous nous appuierons sur l'historique de sinistres et plus particulièrement sur la variable indicatrice du sinistre et la variable indicatrice du type de sinistre.

Après étude de ces variables, on remarque certaines anomalies dues à la méthode de construction des variables. Dans ce paragraphe, nous présenterons les anomalies constatées, nous discuterons des possibles causes de ces anomalies et de la démarche adoptée pour les résoudre afin obtenir des données de sinistralité plus fiables.

La variable « I_sinistre » prend deux modalités 0 ou 1 et répond à la question : « A-t'on observé un sinistre ? **»** Elle prend la valeur 1 en cas de sinistre et 0 le cas contraire. La variable « vrai_sinistre » prend aussi deux modalités 0 ou 1. Elle prend 1 en cas de vrai sinistre et 0 si on est en présence d'un faux sinistre. Ces deux variables sont liées. La variable vrai_sinistre vient préciser la nature du sinistre observé. Ainsi, elle ne doit exister qu'en cas de sinistre c'est–à–dire lorsque la variable I_sinistre a pour valeur 1. Cependant, on remarque les anomalies suivantes :

- Le nombre de sinistres comptabilisé pour la variable I_sinistre est 4 fois plus petit que le nombre de sinistre comptabilisé par la variable vrai-sinistre. Notons que ce dernier résultat est obtenu en sommant le nombre de vrai sinistre et de faux sinistre.

- On observe des ADBAI déclarés comme sinistrés (I_sinistre =1) mais qui n'apparaissent ni comme un vrai sinistre ni comme un faux sinistre (vrai_sinistre non renseignée).

- On observe des ADBAI déclarés comme non sinistrés (I_sinistre =0) mais pour qui on considère qu'il y a eu un vrai sinistre (vrai_sinistre=1)

- On observe des ADBAI déclarés comme non sinistrés (I_sinistre =0) mais pour qui on considère qu'il y a eu un faux sinistre (vrai_sinistre=0)

L'absence de concordance entre ces deux variables pour certaines observations de notre base d'historique peut s'expliquer par le fait qu'elles ont été construites différemment. Elles ont été construites à partir des historiques de CEGC, par des équipes différentes. En fait, la CEGC est née en 2008 de la fusion de plusieurs entités qui jusque-là fonctionnaient de façon indépendante.

Ainsi, nous disposons pour les sinistres survenus avant 2008 des sources d'informations différentes. Nous remarquons que les sinistres qui ne sont pas pris en compte par la variable I_sinistre sont des sinistres survenus avant 2008. Il existe probablement d'autres raisons pouvant expliquer ces incohérences notamment le processus de mise à jour des bases de données.

Pour résoudre ces anomalies, il a fallu dans un premier temps redéfinir les sinistres et dans un second temps définir le type de sinistres pour chacun des dossiers présentant une anomalie.

Dans un premier temps, il a été observé, plus de sinistres avec la variable vrai_sinistre qu'avec la variable I_sinistre. Une étude plus attentive de la base permet de se rendre compte de l'existence de plusieurs doublons parmi les sinistres répertoriés par la variable vrai_sinistre. Une fois les doublons de cette variable supprimés, le nombre de sinistres observé pour la variable vrai_sinistre construite par le pôle Actuariat Technique est moins important que celui obtenu avec la variable I_sinistre construite par l'équipe Risques Consolidés.

Mais un écart subsiste toujours. Par la suite, le choix effectué consiste à retenir tous les sinistres observés en considérant les deux variables et on trouve au total 111 sinistres. Parmi ces 111 sinistres, on compte 28 sinistres dont on ne saurait définir le type vrai ou faux sinistre. A défaut d'information précise, et par souci de concordance avec les informations obtenues des équipes métiers, le montant de paiement peut être considéré comme un paramètre permettant de déterminer le type d'un sinistre. Ainsi, les sinistres ayant un montant de paiement inférieur à 10000 euros sont classés en faux sinistres alors que les montants supérieurs sont considérés comme de vrais sinistres. Ce montant seuil a été défini à partir des montants de paiements observés pour les faux sinistres de notre historique (le montant maximal observé sur notre historique est de 8442 euros).

Bien qu'il puisse subsister des effets de biais liés à l'introduction de ce seuil, nous considérons qu'il s'agit d'une limite raisonnable et que le classement obtenu est suffisamment fiable.

En accord avec la règlementation Solvabilité 2 qui préconise également d'avoir une qualité des données suffisante pour pouvoir y appuyer un modèle interne, un travail important de fiabilisation devra être mis en œuvre au préalable de la mise en place du modèle interne proposé par la suite. Le temps imparti pour ce mémoire n'a pas permis d'effectuer ce travail approfondi qui nécessiterait l'intervention des équipes informatiques et des équipes métiers sur une période de temps prolongée. Toutefois, nous considérons que la base mise à notre disposition et retraitée des éléments présentés ci-dessus est satisfaisante pour la poursuite de notre étude.

2. Description du portefeuille d'assurés

L'historique de données dont nous disposons s'étend de 1999 à 2015. Pour toute année d'étude considérée sur cette période, notre portefeuille d'assurés est constitué d'agents immobiliers et d'administrateurs de biens ayant souscrit pour une garantie de transaction ou de gestion auprès de CEGC.

2.1 Répartition de la population d'assurés

La population d'assurés telle que définie plus haut peut être répartie en fonction de l'activité principale de l'ADBAI ou en fonction du type de garanties auxquelles ces derniers ont souscrit.

En se basant sur leur activité, nous distinguons 3 catégories :

- Les ADBAI ayant pour activité principale les transactions immobilières (AI). Il s'agit principalement des agents immobiliers. Pour cette catégorie, le montant de garantie consolidé lié à l'activité d'agent immobilier est supérieur au montant de garantie consolidé pour l'activité d'administration de biens.
- Les ADBAI ayant pour activité principale la gestion immobilière (ADB). Il s'agit principalement des administrateurs de biens. Dans cette catégorie, le montant de garantie lié à l'activité d'administration de biens est supérieur au montant de garantie consolidé pour l'activité d'agent immobilier.
- Les ADBAI ayant une activité de gestion et de transaction immobilière (mixte). Il s'agit des clients ayant une activité mixte et pour lesquels on a une équivalence entre le montant de garantie consolidé pour l'activité ADB et celui de l'activité AI.

Lorsqu'on considère notre historique de données, on remarque que la catégorie des agents immobiliers est celle qui prédomine avec 47% de présence pour un total de 54681 dossiers.

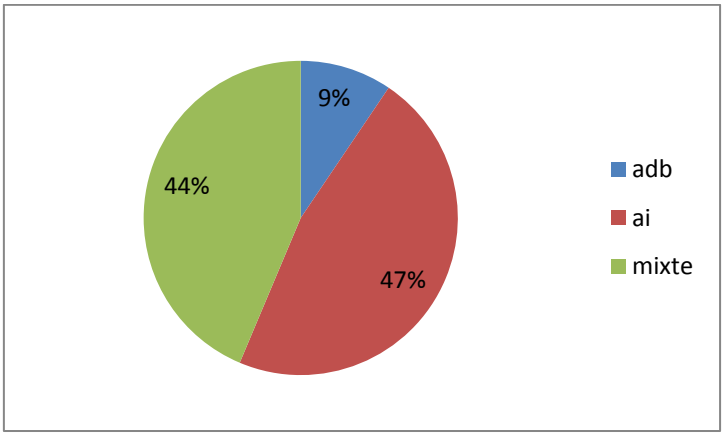


Figure 7: Répartition selon l'activité

En utilisant le type de garantie comme clé de répartition, on distingue deux catégories qui correspondent aux deux garanties proposées par CEGC sur ce marché:

- Les garanties de gestions immobilières qui représentent 53,11% des garanties.
- Les garanties de transactions qui représentent 46,89 % des garanties

2.2 Evolution du portefeuille d'assurés

La vie d'un portefeuille est caractérisée par divers évènements. On peut citer entre autres : l'entrée de nouveaux assurés dans le portefeuille, la sortie de certains ADBAI pour cause de départ vers la concurrence ou de dégradation de la note financière. Aussi, il peut y avoir la survenance d'un sinistre au cours de la période de validité de la garantie octroyée par CEGC à l'ADBAI. Ce sont ces différents paramètres que nous cherchons à étudier dans ce paragraphe.

Afin d'avoir une idée plus concrète des flux d'entrées/sorties qui affectent notre portefeuille, nous cherchons à quantifier le taux d'entrée, le taux de sortie et le taux de sinistralité global calculé sur l'historique de données.

Sur les 54681 dossiers de l'historique, on a observé 111 sinistres soit un taux de sinistralité global de 0,203%. Sur les 17 années d'historique, on observe 6298 nouveaux dossiers, ce qui correspond à un taux d'entrée de 11,518%. Le taux de sortie est de 8,495% soit 4645 sorties sur 54681 dossiers.

Évolution de la sinistralité : En ce qui concerne la sinistralité, les sinistres survenus sont catégorisés en faux et vrai sinistres selon qu'ils mettent en jeu ou non la garantie financière. Sur notre historique d'études, on observe 111 sinistres dont 24% sont des contentieux et les 76% restants sont de vrais sinistres. La prédominance des vrais sinistres s'observe aussi sur les 17 années d'historique. En effet, pour pratiquement toutes les années, la proportion de vrais sinistres est plus importante.

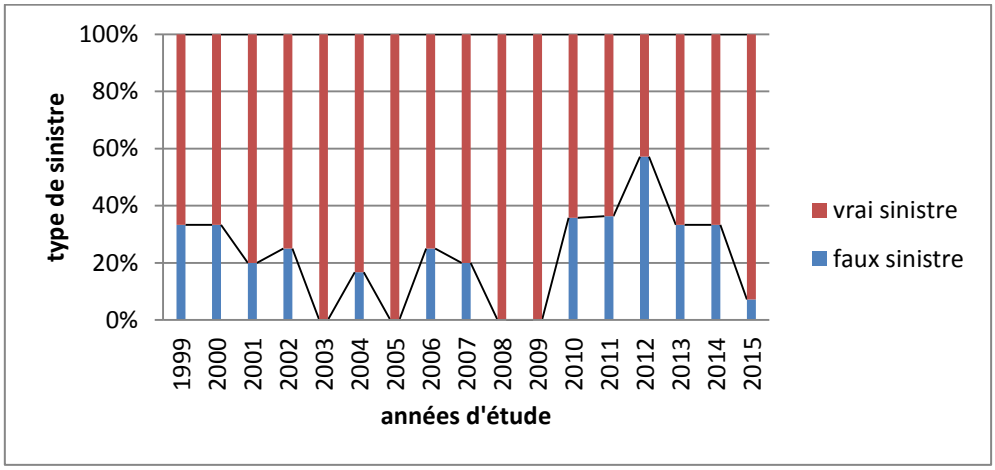


Figure 8: Répartition des sinistres par type de sinistre

Nous souhaitons évaluer la sinistralité par année d'étude. Pour cela, le taux de sinistralité est calculé. Sur notre historique le taux de sinistralité varie entre 0,09% (taux observé en 2008) et 0,4% (taux observé en 2015). On remarque que le taux de sinistralité sur ce marché est relativement bas.

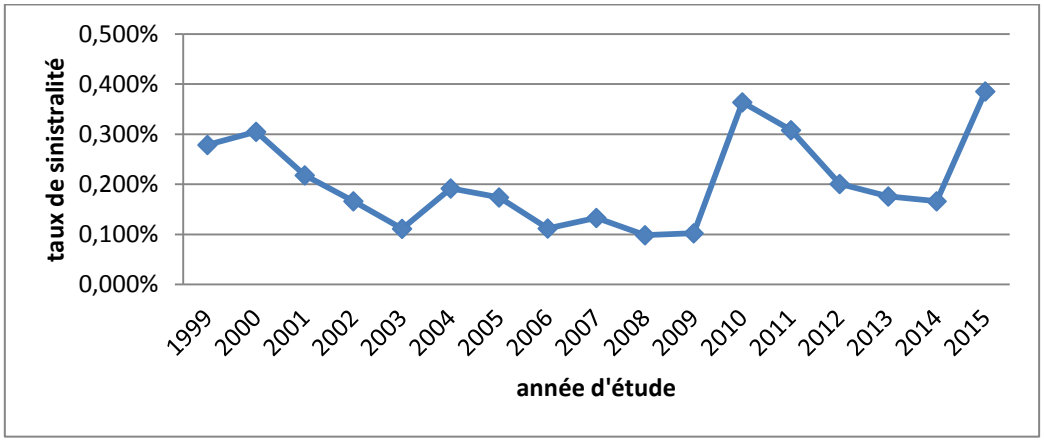


Figure 9: Evolution de la sinistralité

En observant ce graphique on remarque qu'aucune tendance ne se dégage à priori. Les variations brusques de la courbe s'observent en 2009 et en 2014. En effet, on note une légère tendance à la baisse de 1999 à 2008 et une nette augmentation de la sinistralité entre 2008 et 2009. De même entre 2010 et 2014 on a une tendance à la baisse puis une hausse en 2015.

Ces hausses de sinistralité correspondent aux périodes de reflux du marché immobilier français (en 2009 et en 2013) comme on peut le voir sur le graphique suivant :

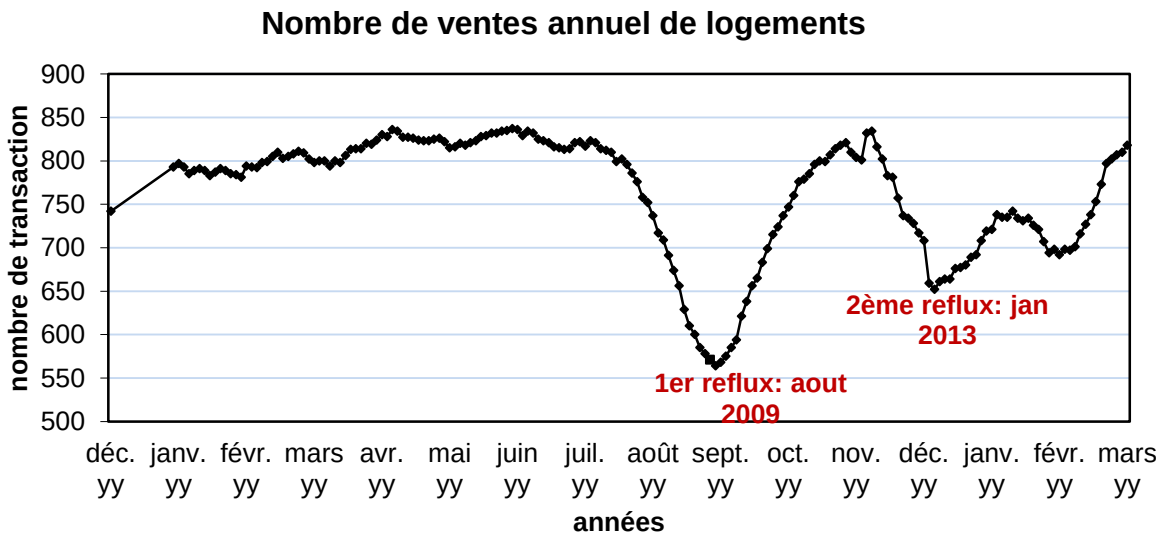


Figure 10: Evolution nombre annuel de transaction immobilière (Source : CGEDD[5])

[5] Conseil général de l'environnement et du développement durable

Ainsi la hausse brusque du taux de sinistralité observé sur le marché des ADBAI en 2010 et 2014 serait due aux baisses importantes du nombre de transactions immobilières observées sur le marché en 2009 et en 2013 soit un an avant et qui auraient impacter le chiffre d'affaire et la santé financière des professionnels de l'immobilier. Ces observations confirment les déclarations des métiers concernant une période de retour d'un an en cas de récession économique.

Evolution des taux d'entrée et taux de sortie : L'évolution du nombre de contrats souscrits et résiliés par année d'étude est représenté sur le graphique suivant :

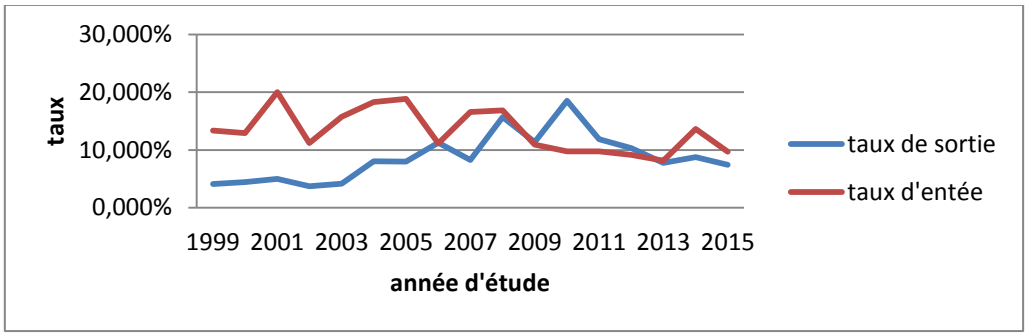


Figure 11: Evolution des taux d'entrée et de sortie

Le taux d'entrée est toujours supérieur au taux de sortie, sauf pour la période de 2009 à 2013. Cela parait logique car la capacité à engranger de nouvelles souscriptions tout en ayant peu de sorties du portefeuille dépend des tendances à la hausse du marché immobilier.

Évolution nombre de souscriptions : Chaque année de nouveaux ADBAI souscrivent à l'une des garanties proposées par la CEGC. Nous voulons voir comment a évolué le nombre de souscription annuelle sur le marché des ADBAI entre 1999 et 2015. Pour cela, nous avons réalisé le graphique suivant :

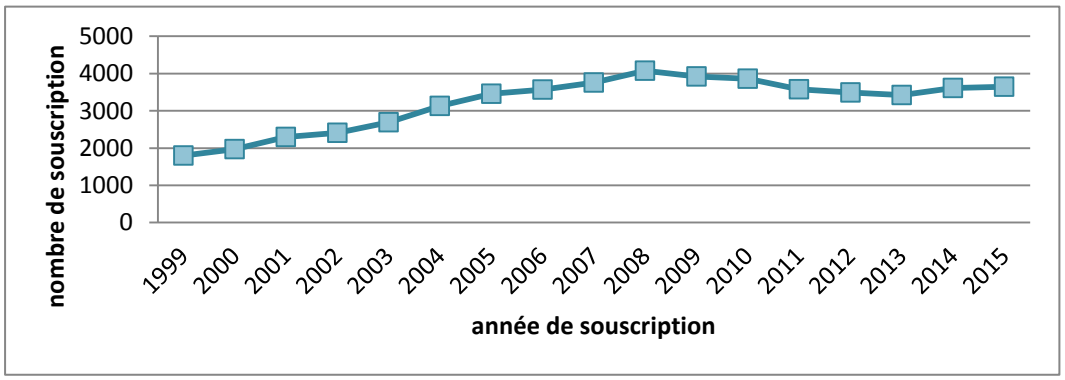


Figure 12: Evolution de la souscription

Sur toutes les années considérées le nombre moyen de souscription est de 3217. On remarque que la courbe est globalement croissante avec un nombre maximal de souscription égal à 4072 atteint en 2008. Depuis 2009 on observe une baisse du nombre de souscriptions qui pourrait s'expliquer par la crise immobilière de 2008 qui a affecté les

prix de l'immobilier, entrainé une baisse des transactions immobilières et la fermeture de plusieurs agences immobilières.

Évolution du montant d'encours global : Le montant d'encours pour un contrat se définit comme le montant sur lequel s'engage la compagnie dans le cadre de la garantie octroyée à l'ADBAI. Au fil des années de l'historique, le montant d'encours global a une tendance globalement croissante comme on peut le voir sur le graphique suivant :

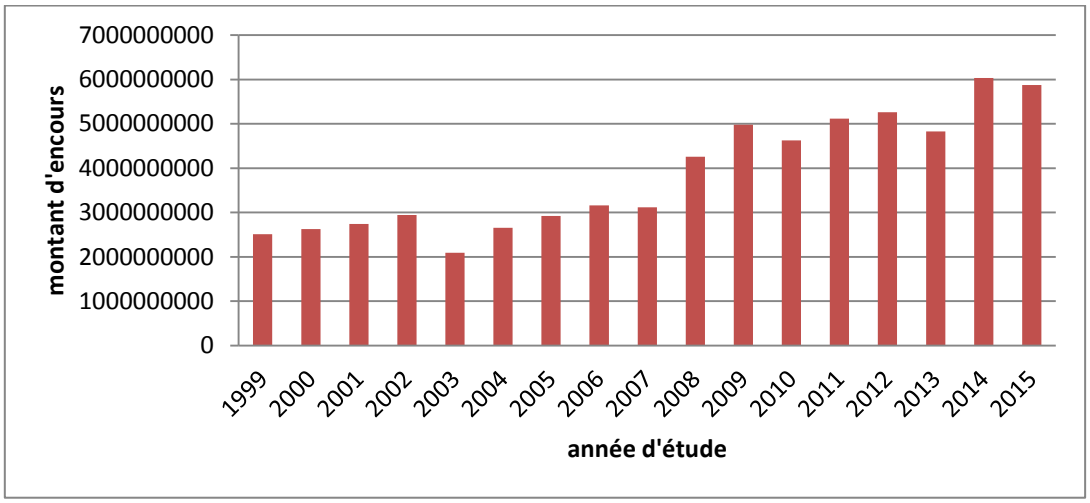


Figure 13: Evolution du montant d'encours global

La baisse la plus importante, s'observe en 2003. En 2014, on observe un pic du montant d'encours global. En effet, cette année-là le montant d'encours connait une augmentation de plus d'1 milliard par rapport à l'année précédente. La croissance du montant d'encours est en fait une conséquence de la croissance du nombre de souscriptions que l'on observe sur le graphique 12.

3. Études statistiques de quelques variables

L'objectif des études réalisées dans ce paragraphe est d'identifier les variables pouvant expliquer la survenance du sinistre. Nous recherchons donc parmi les variables de notre base de données celles qui sont les plus discriminantes vis-à-vis de la sinistralité. Nous reviendrons plus en détail sur les variables suivantes : le score, l'ancienneté, l'âge et le segment. Nous réaliserons plusieurs études et tests statistiques afin de retenir les variables les plus pertinentes.

3.1 Sinistralité par score

En considérant, l'ensemble des sinistres répertoriés dans notre base nous souhaitons voir la répartition de ces sinistres par classe de score. Le graphique suivant représente, le nombre de dossiers et le taux de sinistralité calculé pour chacune des classes de score. Pour les ADBAI dont la classe de score est renseignée, on observe les taux de sinistralité les plus élevés pour les dossiers classés rouge 1,06% et en Watch List 3,38%. Ce résultat est en adéquation avec la façon dont les classes ont été précédemment définies et implique qu'un dossier placé en Watch List ou rouge présente un risque de sinistralité plus important qu'un dossier vert ou orange.

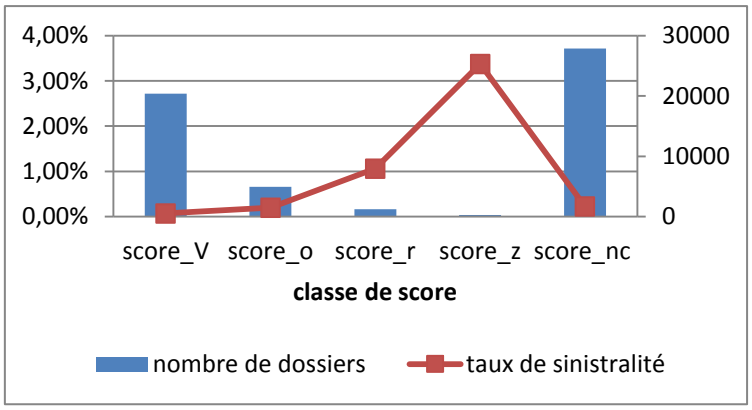


Figure 14: Sinistralité par classe de score

Le tableau suivant récapitule les taux de sinistralité obtenus par classe de score. L'objectif du découpage en classes est de permettre de discriminer les différents individus selon le risque de sinistralité. Nous souhaitons obtenir des classes robustes en termes de sinistralité c'est-à-dire qui présentent un nombre non négligeable de sinistres. En observant les classes et leurs taux de sinistralité respectifs, il n'apparait pas de classe qu'il serait pertinent de regrouper (les taux de sinistralité par classe étant très différents). Les classes de scores obtenues sont assez robustes.

Variable de score	sains	sinistres	total	taux de sinistralité
score_v	20379	14	20393	0,07%
score_o	4907	10	4917	0,20%
score_r	1226	13	1239	1,06%
score_z	266	9	275	3,38%
score_nc	28792	65	28857	0,23%

Table 1: Taux de sinistralité par classe de score

En faisant les rapports de sinistralité entre les classes score_v et score_o, on peut dire que les individus de la classe score_o ont quasiment trois (2,86) fois plus de chance de connaitre un sinistre que les ADBAI notés vert. Quand on passe à la classe score_r, on a 15 fois plus chance d'avoir un sinistre par rapport aux ADBAI de la classe score_v. De même

un individu placé en Watch liste, a trois (3,19) fois plus de chance d'être en sinistre qu'un individu noté rouge.

Afin de tester la pertinence de la variable de score, nous avons réalisé le test du khi-deux. Pour le test effectué sur les variables de sinistralité et de score, la p-value obtenue est 3,0014E-40 pour un seuil alpha fixé à 5%. Cette valeur étant inférieure à alpha, on peut donc affirmer avec moins de 5% de chances de se tromper que la variable de sinistralité et la variable de score sont liées.

Nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	191,149	4	3,00E-40

3.2 Sinistralité par segment

Sur le nombre total de sinistres observés 75% sont des ADB tandis qu'on compte seulement 25% des sinistres chez les agents immobiliers. Comme on peut le voir sur le graphique suivant, on observe une proportion de sinistres relativement plus élevée chez les ADB que chez les AI.

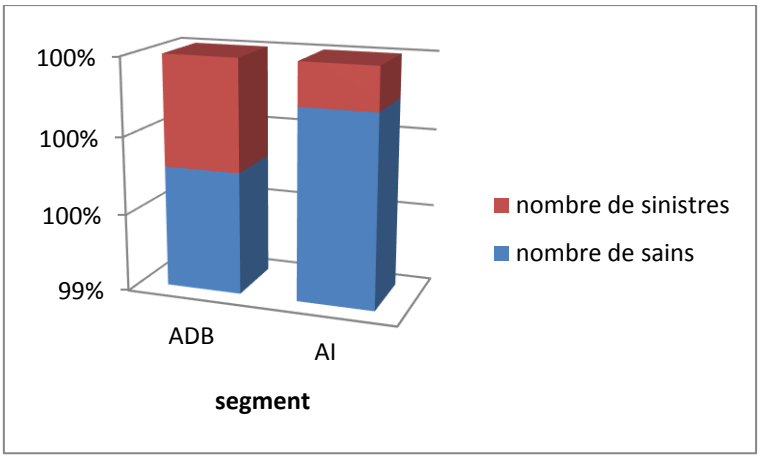


Figure 15: Proportion de sinistres par segment

Etant donné ces résultats, nous souhaitons calculer le taux de sinistralité par segment d'activité. Ainsi, on obtient 0,286% pour les ADB et 0,109% pour les AI. Cela signifie qu'un ADB a deux fois plus de chance d'avoir un sinistre qu'un AI. De ces résultats, le segment apparaît clairement comme un facteur pouvant déterminer la survenance d'un sinistre chez les ADBAI. Ces résultats ont d'ailleurs été confirmés par les experts métiers qui estiment que les ADB sont des clients plus à risque que les AI.

Par des tests supplémentaires nous évaluons le pouvoir discriminant de cette variable. Pour le test d'indépendance du khi-deux, la p-value obtenue est de 4,69 E-6.

nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	20,9566	1	4,7E-06

L'hypothèse H_0 d'indépendance de la variable segment d'activité et de la variable sinistralité est donc rejetée au seuil de 5%. La variable de segment est une variable discriminante de la sinistralité.

3.3 Sinistralité par ancienneté

La variable ancienneté est une variable quantitative. Dans notre base d'étude les modalités vont de 0 à 51 ans. Vu le nombre important de modalités, nous procédons à un découpage en classes. Pour réaliser une bonne discrétisation, les paramètres à déterminer sont le nombre de classes et les bornes des classes. Nous cherchons à construire des classes homogènes et séparées. Nous allons découper notre variable classe en utilisant deux méthodes.

Dans un premier temps, nous utiliserons une méthode d'équipartition pour définir les classes et leurs seuils puis nous utiliserons une méthode tenant compte de la dispersion des observations. Nous construisons 10 classes équiréparties avec la Proc Rank de SAS. Le tableau suivant regroupe toutes les classes de sinistres construites, les valeurs de la variable ancienneté regroupées et les taux de sinistralité par classe.

classes	modalités	Sains	sinistres	taux de sinistralité
1	0	6289	9	0,143%
2	1	6095	9	0,148%
3	2	5453	7	0,128%
4	3	5079	14	0,276%
5	4	4466	9	0,202%
6	5	3876	6	0,155%
7	[6 ; 7]	6260	20	0,319%
8	[8 ; 10]	6204	15	0,242%
9	[11;16]	5666	14	0,247%
10	[17;51]	5182	8	0,154%

Table 2: Taux de sinistralité par classes équiréparties d'ancienneté

Les classes obtenues sont robustes en termes de nombre de sinistres. Le graphique suivant nous permet d'appréhender une éventuelle relation entre les classes construites et la sinistralité.

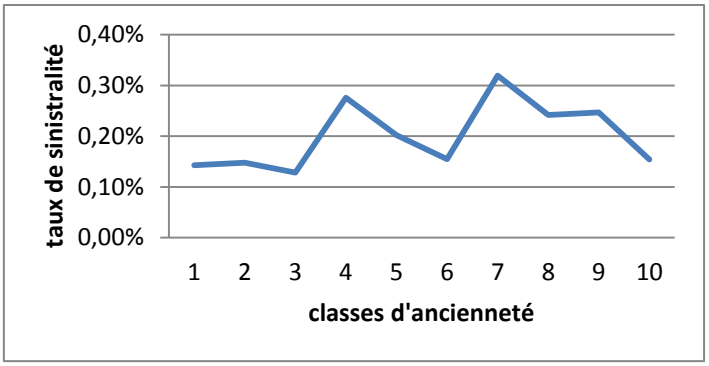


Figure 16: Courbe du taux de sinistralité par classe d'ancienneté équiréparties

Sur ce graphique, on ne remarque pas de tendance particulière. Il n'y a priori pas de relation entre les classes d'ancienneté construites et la sinistralité. Le test du Khi-deux permet de confirmer ce résultat puisqu'on obtient une p-value de 0,27346 et qui est supérieure au seuil 0,5 %. L'hypothèse d'indépendance est acceptée au seuil de 0,5%.

nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	11,033	9	0,2735

A présent, nous faisons un découpage en clusters basé sur une méthode de classification mixte. Pour cela, nous utilisons pour la construction des clusters une méthode de classification mixte. Le principe de la classification est de regrouper entre elles certaines observations que l'on juge similaires selon un critère défini. L'idée est de répartir les n observations dont on dispose et qui sont caractérisées par des variables, en p classes les plus homogènes et différenciées les unes des autres. La classification mixte consiste à réaliser une classification ascendante hiérarchique puis à appliquer la méthode des centres mobiles. Ces méthodes sont décrites en annexe du mémoire. Les caractéristiques des classes obtenues sont reprises dans le tableau suivant :

Cluster	sains	sinistres	Taux de sinistralité	Ancienneté (barycentre)
1	17837	25	0,14%	1
2	2201	5	0,23%	41
3	1752	3	0,17%	27
4	5452	12	0,22%	15
5	10556	26	0,25%	9
6	16772	40	0,24%	4

Table 3: Taux de sinistralité par classes homogènes d'ancienneté

Pour cette classification aussi, il ne semble pas avoir de lien entre les classes d'ancienneté et la sinistralité. Ainsi, l'ancienneté n'est pas un facteur discriminant de la sinistralité. Nous réalisons à nouveau le test du khi-deux.

Nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	5,692	5	0,11751

La p-value obtenue étant supérieure à 0,5%, on ne rejette pas l'hypothèse d'indépendance entre la sinistralité et l'ancienneté. L'ancienneté n'apparait donc pas être une variable discriminante de la sinistralité. Les classes obtenues par cette méthode sont plus compliquées à interpréter que les classes équiréparties car l'algorithme peut regrouper au sein d'une même classe des anciennetés qui sont éloignées par exemple 50, 51 et 3. Pour régler ce problème, il aurait suffi de rajouter un critère dans la construction des classes de façon à regrouper uniquement les classes proches mais étant donné les résultats des différents tests du khi-deux effectués, il ne semble pas nécessaire d'approfondir l'étude.

3.4 Sinistralité par âge

Compte tenu du nombre important de modalités de la variable âge, nous réalisons la même étude que pour la variable ancienneté. Ici l'objectif est de voir l'éventuel impact de l'âge de l'entreprise sur la sinistralité. Dans notre historique la variable âge prend les valeurs allant de 0 à 88 mais pour certains clients cette variable n'a pas été renseignée. On note en tout 4488 valeurs manquantes sur notre historique. Les valeurs manquantes sont regroupées dans une classe distincte, et nous construisons 12 classes équiréparties caractérisées comme suit :

Classe	Modalités	Sains	Sinistres	Taux de sinistralité
1	NA	4465	23	0,51%
2	0	3893	0	0,00%
3	1	4105	2	0,05%
4	2	3889	9	0,23%
5	[3;4]	6885	17	0,25%
6	5	2927	4	0,14%
7	[6;7]	4995	9	0,18%
8	[8;10]	5706	10	0,18%
9	[11;13]	4008	9	0,23%
10	[14;18]	4538	9	0,20%
11	[19;28]	4577	9	0,20%
12	[29;88]	4582	10	0,22%

Table 3: Taux de sinistralité par classes équiréparties d'âge

On remarque que le taux de sinistralité par classe le plus élevé est de 0,51% et est obtenu pour la classe NA et que le taux le plus faible est 0% et est obtenu pour les ADBAI entré dans le portefeuille la même année que l'année d'étude.

Cela pourrait s'expliquer par le fait que l'année de création les ADBAI dispose de fonds notamment de capitaux sociaux substantiels pour faire tourner l'activité. Les difficultés de l'ADBAI ne se manifesteront qu'après l'établissement du bilan annuel. Le graphique suivant montre les taux de sinistralité par classe d'âge de l'entreprise.

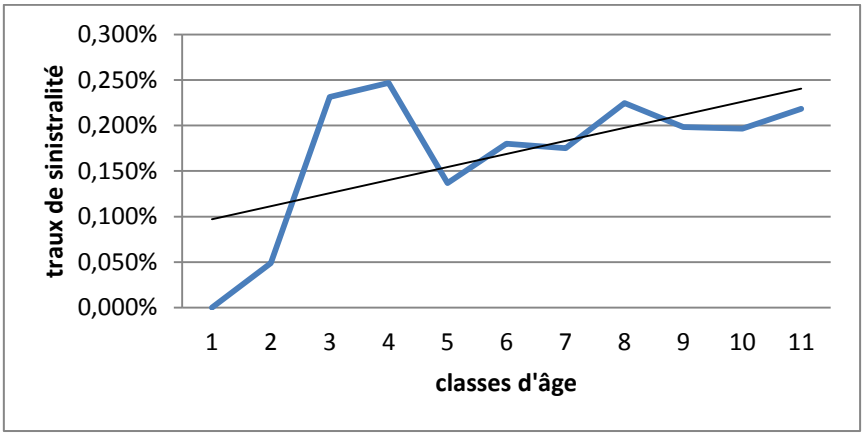


Figure 17: Courbe du taux de sinistralité par classe d'âge équiréparties

On remarque une légère tendance linéaire lorsqu'on considère uniquement les classes renseignées ce qui ferait penser à une dépendance entre classe d'âge et sinistralité. Ce

résultat est confirmé par le test du khi-deux effectué sur ces classes dont la p-value est 0,000176.

nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	35,901	11	0,00017579

Mais ce résultat n'est pas logique puisque cela signifierait l'âge de l'ADBAI pourrait expliquer la survenance du sinistre. En réalité, il apparaît que l'étude de cette variable introduit un biais car les ADBAI ayant un âge plus élevé ont été exposés pendant plus longtemps au défaut. Par ailleurs, la pertinence de ce résultat peut être remise en question en raison du nombre important de données manquantes pour cette variable.

Afin de réduire le nombre de classe et d'infirmer ou de confirmer ce résultat statistique, nous faisons une analyse en cluster en appliquant une méthode mixte identique à celle utilisée pour la variable ancienneté. Pour cette analyse nous éliminons les variables manquantes. En définitive, l'analyse en cluster nous permet de définir 3 classes d'âge pour définit dans le tableau suivant :

Cluster	Sains	Sinistres	Taux de sinistralité	Âge (barycentre)
1	6482	15	0,23%	30
2	1410	3	0,21%	52
3	46678	93	0,20%	6

Table 4 : Taux de sinistralité par classes homogènes d'âge

Pour cette classification aussi, on effectue le test du khi-deux.

nombre d'observations	Valeur khi-deux	DDL	p-value khi-deux
54681	0,295	2	0,13102

On obtient une p-value de 0,13102. Ce qui veut dire qu'on accepte l'hypothèse d'indépendance au seuil de 5%. L'âge de l'entreprise n'est donc pas une variable discriminante de la sinistralité.

De ces études, il ressort que les variables de score et de segment sont des variables pouvant expliquer la survenance du sinistre alors que les variables relatives au temps plus précisément l'âge et l'ancienneté ne sont pas des variables discriminantes de la sinistralité. Nous nous appuierons sur ces résultats pour définir les variables explicatives du sinistre. Il serait aussi intéressant d'utiliser un indicateur de l'évolution (baisse ou hausse) des transactions immobilières par région géographique et de voir son impact sur la sinistralité. Pour mener ces études, il faudrait collecter de nouvelles données,

construire un historique fiable, ce qui serait coûteux en temps. C'est pourquoi elles n'ont pas été effectuées mais cela donne déjà des pistes de réflexion et idées pour des études ultérieures.

PARTIE 2 : LA SURVENANCE DU SINISTRE EN CAS DE DESEQUILIBRE DES CLASSES

Chapitre 1:Régression logistique et arbres de décisions CART en cas de déséquilibre des classes

1. Introduction aux déséquilibres des classes

1.1 Définition

Le déséquilibre des classes est un problème assez courant en analyse de données. Il est relatif aux variables possédant deux classes et s'observe dans des situations assez variées notamment en marketing lorsqu'on s'intéresse à la perte ou non d'un client par une entreprise, en banque lorsqu'on veut détecter la présence ou non de fraude. Dans ces cas, bien souvent le nombre de négatifs est très minoritaire par rapport aux positifs : on parle alors de déséquilibre des classes. Le problème vient du fait que l'on cherche à modéliser une variable dont l'une des modalités concerne un nombre très faible d'observations. En ce qui concerne la proportion à partir de laquelle, le problème se pose, les avis sont plutôt partagés. Plusieurs auteurs s'accordent à dire que la situation de déséquilibre des classes est avérée à partir du moment où l'une des classes de la variable d'intérêt représente en proportion moins de 10% des observations, pour d'autres le seuil retenu est plutôt de 5%.

Dans ce genre de situation, il suffit par exemple de classer tous les dossiers en négatifs pour avoir un taux d'erreur de 5 % et il s'avère en général compliqué de trouver un classificateur qui fait mieux c'est ce qu'on appelle l'« accuracy paradox ». Cependant, l'enjeu étant de modéliser les positifs, le problème reste entier. Sur le marché des ADBAI, les positifs sont en forte minorité ce qui rend délicat la recherche d'un modèle pertinent. En optant pour un modèle qui réduit l'erreur en classant tous les individus en négatif, les coûts engendrés sont plus importants en cas de sinistre comparé à un modèle qui prédit des sinistres qui s'avère en réalité être sains. C'est ce qui explique la recherche d'un modèle qui prédit le sinistre même s'il présente un taux d'erreur plus élevé.

La stratégie la plus couramment utilisée pour faire face au déséquilibre des classes est le rééchantillonnage. Cette technique consiste à rééquilibrer la base par des techniques de sélection aléatoire des données. Dans le cadre de ce mémoire, nous considérerons cette approche selon deux points de vue différents : la régression logistique et les techniques d'apprentissage. Dans la section suivante nous présentons une liste non exhaustive des travaux réalisés en correction d'une situation de déséquilibre de la base de données.

1.2 Quelques travaux importants

- En régression logistique :

En régression logistique, plusieurs chercheurs se sont interrogés sur la légitimité d'utiliser une régression logistique classique dans le cas d'une base présentant un déséquilibre des classes. En effet l'avantage principal lorsqu'on utilise le maximum de vraisemblance pour estimer les coefficients d'une régression est que l'on obtient des estimateurs possédant asymptotiquement les propriétés de convergence, mais ces estimateurs pourraient présenter un biais lorsque l'échantillon étudié est de petite taille. En cas de déséquilibre des classes de l'échantillon, le degré de biais dépend fortement du nombre d'observations de la classe minoritaire. Ainsi les résultats obtenus sont biaisés et la probabilité de réalisation de la modalité minoritaire est mal estimée.

En régression logistique, en raison du déséquilibre on est face au problème de biais des estimateurs et par conséquent à la sensibilité de la variance. Pour une meilleure compréhension de la suite de ce paragraphe, le lecteur pourra lire au préalable le premier point du chapitre 2 qui introduit la régression logistique.

Le biais des estimateurs

Pour rendre plus facile la compréhension du biais des estimateurs, KING et ZENG en proposent une expression simplifiée. Ils considèrent le cas où le modèle possède deux paramètres : β_0 le terme constant qui sera estimé par le modèle et β_1 le second paramètre qui est fixé à 1. Le modèle s'écrit alors :

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + x_i \text{ Avec } \pi(x_i) = \mathbb{P}(y_i = 1|x = x_i) \text{ et donc } \pi(x_i) = \frac{e^{\beta_0+x_i}}{1+e^{\beta_0+x_i}}$$

Dans ce cas précis l'expression mathématique proposée pour le biais est :

$$E(\widehat{\beta_0}-\beta_0) \approx \frac{\bar{\pi}-0{,}5}{n\bar{\pi}(1-\bar{\pi})} \text{ où } \bar{\pi} = \frac{1}{n}\sum_{i=1}^{n}\pi_i$$

En fait $\bar{\pi}$ représente la probabilité moyenne de réalisation de la modalité 1.

Dans le cas de déséquilibre des classes, on a $\bar{\pi} < 0{,}5$ car la probabilité pour chaque observation de réalisation de la modalité minoritaire est petite. Par conséquent le numérateur et le biais seront négatifs. Le résultat négatif obtenu veut dire que le coefficient estimé $\widehat{\beta_0}$ est trop petit et donc que la probabilité $\pi(x_i)$ est sous-estimée.

La sensibilité de la variance

En effet dans une régression logistique classique telle que décrite ci-dessus, les coefficients $\hat{\beta}$ obtenus par maximisation de la vraisemblance du modèle, admettent une matrice de variance-covariance. Cette matrice notée $V(\hat{\beta})$ se calcule par la formule suivante :

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \hat{\pi}(X_i) * \left(1 - \hat{\pi}(X_i)\right) * X_i{}' X_i\right]^{-1}$$

Où

- $X_i = (x_{i0}, x_{i1}, \dots, x_{ik})$ est le vecteur de réalisation des variables explicatives pour l'observation i
- $\hat{\pi}(X_i)$ est la probabilité de survenance du sinistre estimée par le modèle à partir des variables explicatives X_i.

Dans cette expression, le facteur $\hat{\pi}(X_i) * \left(1 - \hat{\pi}(X_i)\right)$ est affecté par le déséquilibre des classes. En effet, on remarque que dans une situation où la classe 1 serait très minoritaire par rapport à la classe 0, la probabilité estimée $\hat{\pi}(X_i) = \mathbb{P}(\mathrm{Y} = 1 | X = X_i)$ sera très faible pour toutes observations. Cependant d'après ZENG et KING [7], lorsque le modèle utilisé a un fort pouvoir explicatif, cette probabilité sera plus grande et proche de 0,5 en ce qui concerne les éléments pour lesquels la classe 1 est effectivement observée. Car, de façon générale, on remarquera qu'en cas de déséquilibre des classes, quelque soit le modèle considéré la probabilité estimée pour $\mathrm{Y} = 1$ sera très faible comparée à celle estimée pour $\mathrm{Y} = 0$. Ainsi le terme $\hat{\pi}(X_i) * \left(1 - \hat{\pi}(X_i)\right)$ sera toujours plus grand pour les observations de la classe 1 que pour celle de la classe 0. Cela s'explique par les variations de la fonction $x(1 - x)$. Cette dernière croit entre 0 et 0,5 puis décroit entre 0,5 et 1. Par conséquent la variance inverse du terme $\hat{\pi}(X_i) * \left(1 - \hat{\pi}(X_i)\right)$ sera petite pour les observations de classe 1 et grande pour celles de classe 0. En conséquence plus il y a d'éléments de classe 0 et plus importante sera la variance et plus il y a d'éléments de classe 1 et plus faible sera la variance.

En solution à ces problèmes on retient 3 importants travaux :

- Les méthodes de correction de biais après rééchantillonnage de KING et ZENG
- La régression logistique exacte de COX
- La pénalisation du maximum de vraisemblance par FIRTH

KING et ZENG proposent une stratégie basée sur le rééchantillonnage. L'objectif est de réduire le biais et la variance des estimateurs. Pour cela on cherche à ajouter le nombre d'observations de classe 1 dans l'échantillon et à sélectionner aléatoirement les observations de classe à 0 tout en ayant un certain équilibre entre les deux classes. C'est ce qu'on appelle « Response-Based Sampling ». Une fois la base rééchantillonnée, on applique la régression logistique classique puis des corrections aux paramètres obtenus.

La seconde approche a été proposée par COX et l'algorithme permettant de l'implémenter a été proposé par gail et al. Cette méthode consiste à corriger le biais engendré en cas de base de données de petites tailles, ayant un déséquilibre des classes. Cette méthode fonctionne uniquement lorsque la taille de l'échantillon est inférieure à 200.

La troisième méthode appelée aussi correction de Firth, a été proposée par David FIRTH pour réduire le biais des estimateurs par maximum de vraisemblance. Cette méthode fonctionne très bien en cas de petites bases et a l'avantage de fournir en tout cas des estimateurs convergents. Elle consiste à ajouter une fonction de pénalité à la log vraisemblance du modèle de régression logistique classique selon la formule suivante

$$logL(\beta)^f = \log L(\beta) + \frac{1}{2} log|I(\beta)|$$

Où

- $|I(\beta)|$ est le déterminant de la matrice d'information de Fisher
- $L(\beta)$ est la vraisemblance du modèle de régression logistique telle que présentée ci-dessus.

Ces deux dernières approches sont souvent utilisées lorsqu'on observe un problème de séparation complète ou de quasi séparation des classes. La séparation des classes est un problème connexe en cas de déséquilibre des classes qui survient dans une base de données lorsqu'une combinaison de variables explicatives permet de séparer la base en deux groupes distincts correspondant aux modalités de la variable à expliquer.

Dans le cadre de notre base, il y a un déséquilibre des classes mais pas de problème de séparation. De plus la taille de la base est importante (54681 observations). La méthode de la régression logistique exacte de Cox et la méthode de pénalisation du maximum de vraisemblance par firth, ne sont pas pertinentes dans le cadre de notre étude. Notre intérêt portera donc sur la première méthode.

- <u>En apprentissage :</u>

En apprentissage, l'une des plus grandes contributions pour remédier au déséquilibre des classes est celle de WEISS. Dans, l'article référencé [10], il décrit le problème de déséquilibre des classes en mettant une nuance sur le degré de déséquilibre des classes. Il fait une distinction entre une situation d'absolue rareté dans laquelle on ne dispose pas d'assez d'observations de la classe minoritaire pour apprendre une décision et une situation de rareté relative où malgré la faible quantité d'observations de la classe minoritaire, l'apprentissage est possible. Il décrit également les problèmes rencontrés en apprentissage à cause du déséquilibre des classes notamment le problème de manque de données, la faible performance des algorithmes, l'inadéquation des métriques utilisées pour tester la performance des classificateurs. Ensuite, il décrit un certain nombre de méthodes pour corriger ces problèmes notamment le rééchantillonnage, le *boosting* des algorithmes d'apprentissage et l'utilisation de métriques plus pertinentes. Ces notions seront présentées plus en détails dans la suite. Ses travaux servent de socle lorsqu'on s'intéresse à l'apprentissage en cas de déséquilibre des classes. Depuis, des améliorations ont été apportées à ces différentes techniques et d'autres techniques ont été proposées.

En apprentissage, il y a trois approches principales permettant de régler le problème de déséquilibre des classes : l'approche par la base, l'approche par algorithme et l'approche par apprentissage des matrices de coûts.

1. La première méthode encore appelée approche externe est la plus couramment utilisée. Elle consiste à rééquilibrer la base en jouant sur les proportions de positifs et de négatifs dans la base. Cette technique est indépendante du classificateur utilisé et de ce fait est plus souple et flexible. Il est donc possible d'équilibrer la base et appliquer le classificateur de son choix.

2. La deuxième méthode appelée approche interne consiste à jouer sur les algorithmes de classification. L'idée est d'essayer de modifier ou d'adapter certaines caractéristiques des classificateurs de façon à améliorer leurs performances.

3. La troisième méthode consiste à introduire une matrice de coûts de mauvais classements de façon à réduire les erreurs de classification. les matrices de coûts sont utilisées pour décrire les coûts engendrés par la mauvaise classification de l'un ou l'autre des exemples (observation). On se retrouve dans une situation de pondération des exemples, les poids étant ici les coûts. La technique consiste à modifier le processus d'apprentissage en tenant compte des coûts. On intègre une fonction de coûts à minimiser lors de la séparation en classes des exemples.

L'inconvénient majeur des méthodes basées sur les coûts de mauvais classement est la nécessité de définir les coûts ou la fonction de coûts de mauvaises classifications.

En apprentissage, les méthodes proposées sont diverses et variées et il est possible d'associer plusieurs méthodes. Certaines méthodes fonctionnent plus ou moins bien que d'autres selon les données dont on dispose.

L'objet de ce mémoire n'est pas de faire une revue de toutes les techniques existantes et qui permettent de régler le problème de déséquilibre des classes mais de présenter quelques techniques couramment utilisées.

Pour l'approche par régression logistique, nous travaillerons sur les méthodes de correction de KING et ZENG. En ce qui concerne les techniques d'apprentissage, l'étude portera sur les deux premières méthodes évoquées dans le paragraphe portant sur l'apprentissage.

2. Les méthodes de rééquilibrage de la base

Modifier l'échantillon d'étude de façon à ce que le volume de réponses positives pour la variable d'intérêt soit plus important, permet d'obtenir un équilibre des classes de la variable d'intérêt. L'avantage de cette méthode est qu'elle est transverse. Elle peut

s'appliquer pour n'importe qu'elle modèle. Ils existent plusieurs méthodes d'échantillonnage notamment:
"Simple Random Sampling", "Exogenous Stratified Sampling", "Response-Based Sampling".

- Simple Random Sampling : cette technique consiste a sélectionné aléatoirement les observations dans la base d'origine pour constituer le nouvel échantillon, chaque élément ayant la même probabilité d'être sélectionné.

- Exogenous Stratified Sampling : il s'agit d'une méthode pour laquelle les observations sont dans un premier temps classés en sous-groupes formés en fonction des variables explicatives. On parle de strate. Dans un second temps les observations sont sélectionnées aléatoirement au sein de chaque groupe selon des taux fixés. Pour cette méthode, chaque élément d'un sous-groupe a la même probabilité d'être sélectionné.

- Response-Based Sampling : cette méthode est similaire à la précédente à la différence que l'échantillon est stratifié par rapport à la variable réponse. En fait dans un premier temps les sous-groupes sont créés en fonction des modalités de la variable réponse. Dans un second temps les observations sont sélectionnées de façon aléatoire dans chaque sous-groupe en fonction de taux différents pour chaque groupe, chaque élément d'un sous-groupe ayant la même chance d'être sélectionné.

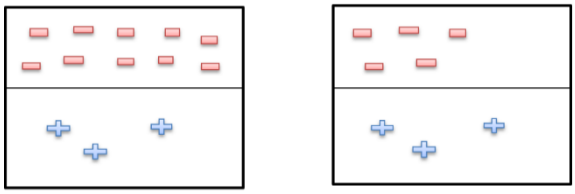
La méthode la plus pertinente pour modifier notre base est la dernière puisque l'objectif est de construire une base stratifiée en fonction des modalités de la variable de sinistre à expliquer. En jouant sur la proportion de représentation de chaque classe, il est possible de définir plusieurs stratégies.

2.1 Les méthodes classiques

Elles consistent à réduire ou à augmenter par sélection aléatoire le nombre d'observations d'une classe donnée. Il s'agit de l'undersampling (réduction des observations) et de l'oversampling (augmentation des observations).

- ✓ **l'undersampling**

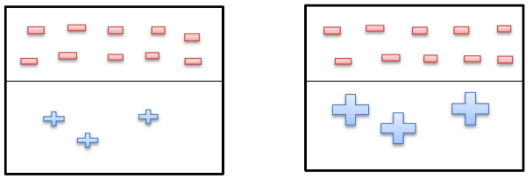
Cette technique consiste à éliminer certains les éléments de la classe majoritaire. L'échantillon obtenu contient un nombre réduit d'observation par rapport à la base d'apprentissage initiale. Elle s'appuie sur une suppression de certaines observations de la classe majoritaire. L'inconvénient de cette méthode est la perte d'information sur la classe majoritaire par rapport à la base initiale. On obtient une base certes équilibrée mais qui ne reflète pas toujours l'information contenue dans la base initiale



Actuel undersampling

- ✓ **l'oversampling**

La méthode consiste à dupliquer les éléments de la classe minoritaire avec pour conséquence une augmentation de la taille de l'échantillon. Elle s'appuie sur la classe minoritaire et consiste à répéter plusieurs fois les observations de la base d'origine possédant cette modalité. L'avantage de cette méthode est qu'elle reprend toutes les observations existant dans la base d'origine. Mais l'inconvénient est qu'elle pourrait donner lieu à un sur-apprentissage auquel cas le modèle ne serait pas robuste pour d'autres données. La notion de sur-apprentissage est décrite dans la suite du mémoire.



Actuel oversampling

Undersampling vs oversampling

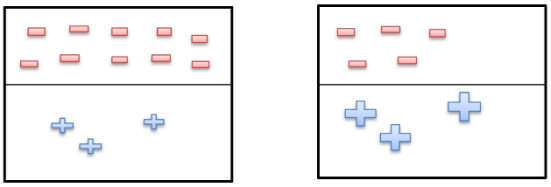
Ces deux méthodes conduisent à diminuer de façon considérable le déséquilibre entre les classes avec pour finalité, l'obtention des bases où la classe minoritaire est mieux représentée. Cependant comme évoqué lors de la description des deux méthodes, il apparait qu'elles présentent des conséquences sur la qualité des modèles construits.

En ce qui concerne le temps de calcul, l'undersampling est meilleure. Les études antérieures n'ont pas permis d'aboutir à une conclusion définitive quant à la meilleure des deux méthodes. Selon certains auteurs, l'oversampling est toujours meilleure en ce sens que toute l'information de la base initiale est conservée.D'autres auteurs pensent que l'undersampling est meilleur en termes de précision.
Il existe des approches dites hybrides qui permettent de corriger ces méthodes et de tirer parti de leur avantage. Parmi ces techniques, la méthode combinaison de l'oversampling et de l'undersampling appelé « Both sampling » dans ce mémoire, le SMOTE (Synthetic Minority Oversampling Technique) et le ROSE (Random Oversampling Examples) seront présentés.

2.2 Les méthodes hybrides

✓ Both sampling

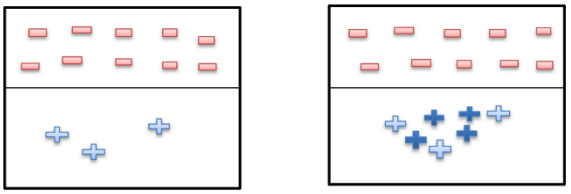
Cette méthode est une association de l'undersampling et de l'oversampling, Elle a été introduite par Ling and Li. L'application de cette méthode conduit à dupliquer les observations de la classe minoritaire (oversampling) tandis que les observations de la classe majoritaire sont réduites (undersampling). L'avantage de cette méthode est qu'elle permet d'obtenir une base de même taille que la base d'origine sans perte d'informations sur les observations de la classe majoritaire.



Actuel both

✓ Synthetic Minority Oversampling Technique (SMOTE)

L'algorithme SMOTE construit de artificiellement de nouvelles observations de façon à augmenter les éléments de la classe minoritaire.



Actuel smote

Cette technique a été introduite par CHAWLA. Elle consiste à construire de nouvelles observations ayant une similarité avec les observations de la classe minoritaire. Cette construction est basée sur la méthode des K-plus proches voisins (K-ppv). Le principe de la méthode k-ppv est la suivante :

Une observation est sélectionnée puis elle est comparée à toutes les observations stockées dans la base par un calcul de la distance entre l'observation sélectionnée et le reste des observations. Les distances obtenues sont stockées puis classées et on retient les k observations possédant les plus petites distances qui sont appelés les k- plus proches voisins.

Notons S_+ l'ensemble formé par les éléments de la classe minoritaire. Pour chaque observation x de la classe minoritaire S_+, les k observations ayant les plus petites distances euclidiennes par rapport à x sont déterminées. Les nouvelles observations sont générées le long des segments joignant les éléments les plus proches de x.

Concrètement pour créer le nouvel échantillon contenant de nouvelles observations de la classe minoritaire, la procédure est la suivante :

1. pour chaque x, un de ses plus proches voisins est sélectionné aléatoirement.
2. On calcule la différence (en termes de distance) entre l'élément x et son voisin sélectionné noté $\hat{x}$.
3. Cette différence est multipliée par un paramètre λ aléatoire prenant ses valeurs dans l'intervalle [0, 1].

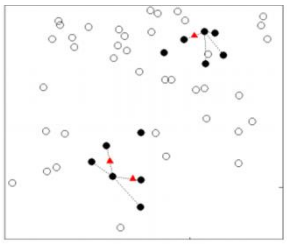
La formule finale pour le calcul des nouvelles observations synthétiques est :

$$x_{syn} = x + (\hat{x} - x) * \lambda$$

Avec:

- x une observation appartenant à la classe minoritaire
- $\hat{x}$, un des plus proches voisins de x
- λ, un paramètre dont la valeur est générée aléatoirement dans l'intervalle [0,1]

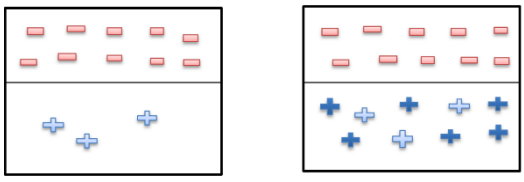
Pour k=4, une représentation schématique des observations synthétiques construites est la suivante :



Classe majoritaire • classe minoritaire ▲ observation synthétique

✓ **Random OverSampling Examples(ROSE)**

Cet algorithme a été proposé par MENARDI et TORELLI [9] pour répondre au besoin de classificateur possédant de bonne capacité de précision. Cette technique d'échantillonnage vise à construire un échantillon permettant de remédier au manque de précision des modèles dû au déséquilibre des classes. L'algorithme ROSE consiste à générer des données artificielles.



Actuel rose

Dans l'article d'introduction à cette technique, cet algorithme est décrit comme suit :

Soit S_n l'ensemble des éléments de la base d'origine. S_n contient n observations chacune caractérisée par le couple de variable(x_i, y_i). y_i correspond à la réalisation de la variable d'intérêt pour l'observation i. y_i a donc deux possibilités 0 et 1 notée respectivement y_0, y_1. x_i est un vecteur contenant les réalisations des variables

prédictives de la i[ème] observation et a une densité de probabilité inconnue $(x) = p(x)$. Notons n_i le nombre d'éléments de chaque classe. Ainsi n_0, n_1 correspondent respectivement au nombre d'observation appartenant aux classes 0 et 1. En notant(x^*, y^*), les nouvelles observations générées artificiellement, leur procédure de construction est la suivante :

1. La première étape consiste à sélectionner y^* parmi les deux possibilités de y_i avec la probabilité π_i qui dépend de la proportion d'observations voulue dans chacune des classes. Si $\pi_i = 1/2$ alors la moitié des observations générés appartient à la classe 0 et l'autre moitié appartient à la classe 1.
2. La deuxième étape consiste à sélectionner aléatoirement une observation (x_j, y_j) dans la base d'origine S_n tel que $y_j = y^*$. Chaque observation vérifiant cette hypothèse a une probabilité $p_j = \frac{1}{n_i}$ d'être sélectionnée.
3. La dernière étape consiste à construire le vecteur x^*de réalisation des variables prédictives. x^* est sélectionné dans le voisinage de l'observation (x_j, y_j) à partir de $K_{H_i}(x_j)$ où K_{H_i}est une distribution de probabilité centrée en x_j et H_i est la matrice des paramètres de lissage. En général K_{H_i}est choisi parmi les distributions unimodales et symétriques.

Ainsi une fois la classe de y sélectionnée, x est modélisé par la densité $f(x|y = y_i)$ une estimation de cette densité est alors calculée par la formule suivante :

$$\hat{f}(x|y = y_i) = \sum_{j=1}^{n_i} p_j * proba(x|x_j)$$

$$= \sum_{j=1}^{n_i} \frac{1}{n_i} * proba(x|x_j)$$

$$= \sum_{j=1}^{n_i} \frac{1}{n_i} K_{H_i}(x - x_j)$$

Cette expression de la densité correspond à la densité de Kernel encore appelé méthode du noyau. Elle propose une estimation de la densité d'une variable aléatoire lorsqu'on dispose d'un échantillon d'observations de la variable. Soit x une variable aléatoire ayant une densité f(x) qui est inconnue. La méthode de kernel propose une estimation de f(x) à partir d'un n-échantillon. Soit $x_1, x_2, \dots, x_n$, l'échantillon de la variable aléatoire x, alors l'estimateur de la densité est :

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{n} K_h(x - x_i) = \sum_{i=1}^{n} \frac{1}{n\,h} K\left(\frac{x - x_i}{h}\right)$$

Où :

- $K(x)$ *est appelé noyau, souvent choisi comme la densité gaussienne centrée réduite*
- h *est un paramètre de lissage nommé* $fenêtre$

Une fois la base équilibrée, il est possible d'appliquer différentes méthodes pour modéliser la survenance du sinistre. Dans la suite nous présentons les méthodes de corrections en régression logistique et les méthodes utilisées pour améliorer les performances de l'arbre CART.

Chapitre 2: Théorie sur les méthodes de correction en régression logistique

1. Introduction à la régression logistique

La régression logistique fut historiquement la première méthode utilisée, notamment en épidémiologie, pour modéliser l'absence ou la présence d'une pathologie en fonction de symptômes. Les exemples d'utilisation de cette méthode dans d'autres domaines sont nombreux. En banque, ce modèle peut être utilisé pour déterminer l'acceptation ou le refus de crédit à un client en fonction des caractéristiques de celui-ci. L'objet de ce paragraphe est d'exposer les résultats théoriques sur la régression logistique, en particulier la régression logistique binaire.

1.1 Formulation du modèle

Soit Y la variable d'intérêt et X le vecteur de variables explicatives. La variable Y prend deux modalités 0 et 1. Le vecteur X est constitué de $x_{i;i=1,\dots,k}$ variables explicatives quantitatives ou qualitatives. Pour chaque individu considéré, la régression logistique vise à modéliser la variable y à partir des réalisations des variables x_i. En particulier c'est la probabilité que Y prenne la valeur 1 (ou 0 la démarche serait la même) sachant les valeurs des variables explicatives.

Soit π cette probabilité et $\beta_0, \beta_1, \dots, \beta_k$ les coefficients de l'estimation alors pour la j^{ème} observation, on a :

$$\pi(X_j) = \mathbb{P}(Y = 1 | X = X_j)$$

La particularité de ce modèle est que l'on souhaite restreindre les possibilités de valeurs prises par Y à {0,1}. π(X) étant une probabilité, ses valeurs doivent être comprises entre 0 et 1 pour tout X. Pour cela, deux fonctions sont utilisées : le logit et le probit.

Le modèle logit

La fonction lien logit à pour expression : $\boldsymbol{logit(p) = \ln(\frac{p}{1-p})}$

Ainsi l'expression du modèle devient alors :

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \Leftrightarrow \quad \pi(X) = logit^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$\Leftrightarrow \quad \pi(X) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

$$\Leftrightarrow \quad \pi(X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}$$

Le modèle probit

La fonction lien probit a pour expression : $\boldsymbol{probit(p) = \phi^{-1}(p)}$

ϕ est la fonction de répartition de la loi normale centrée réduite N(0,1) et a pour expression :

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} * e^{-\frac{t^2}{2}}\, dt, \forall\, x\epsilon\, \mathbb{R}$$

Le modèle s'écrit alors :

$$\phi^{-1}\big(\pi(X)\big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \Leftrightarrow\ \pi(X) = \phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$\Leftrightarrow \pi(X) = \int_{-\infty}^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} \frac{1}{\sqrt{2\pi}} * e^{-\frac{t^2}{2}}\, dt, \forall\, x\epsilon\, \mathbb{R}$$

Lorsqu'on parle de modèle logistique on fait plutôt référence au modèle logit.

1.2 Estimation des coefficients par maximisation de la vraisemblance

La méthode classique utilisée pour estimer les coefficients dans un modèle de régression logistique est le maximum de vraisemblance.

Considérons l'échantillon $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ contenant n observations. Sur cet échantillon, la vraisemblance L du modèle qui se définit comme la probabilité d'observer cet échantillon peut être calculé par la formule classique de vraisemblance.

$L = \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$ où $y_{i,i=1,\dots,n} \in \{0,1\}$

$L = \mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = 2\} \cap, \dots, \cap \{Y_n = y_n\})$

$L = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i)$ car les observations sont supposées indépendantes.

Or $\mathbb{P}(Y_i = y_i)$ suit une loi de Bernoulli de paramètre $\pi(X)$ donc

$$\mathbb{P}(Y_i = y_i) = \pi(X_i)^{y_i} \left(1 - \pi(X_i)\right)^{1-y_i}$$

$$L = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi(X_i)^{y_i} \left(1 - \pi(X_i)\right)^{1-y_i}$$

$$\log(L) = \sum_{i=1}^{n} y_i * \log\big(\pi(X_i)\big) + (1 - y_i) log\big(1 - \pi(X_i)\big)$$

Le modèle logistique étant tel que $\pi(X_i) = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}}$

Où :

- $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ est le vecteur des coefficients
- $X_i = (x_{i0}, x_{i1}, \dots, x_{ik})$ est le vecteur de réalisation des variables explicatives pour l'observation i

Maximiser la vraisemblance revient alors à écrire les expressions suivantes :

$$\max_{\beta}(L) = \max_{\beta} log(L) = \max_{\beta}(\prod_{i=1}^{n} \pi(X_i)^{y_i} \left(1 - \pi(X_i)\right)^{1-y_i})$$

$$= \max_{\beta} \sum_{i=1}^{n} y_i * \log\left(\pi(X_i)\right) + (1 - y_i) log\left(1 - \pi(X_i)\right)$$

La solution à ce problème d'optimisation est le vecteur $\hat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_k})$ tel que :

$$\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial \beta_1} = \cdots = \frac{\partial L}{\partial \beta_k} = 0$$

Pour le résoudre, il n'y a pas de méthode directe d'où le recours aux techniques numériques. La méthode la plus utilisée est l'algorithme de Newton-Raphson.

Une fois le vecteur des coefficients estimés $\hat{\beta}$ calculé, il est alors possible de déterminer la probabilité $\hat{\pi}(X_i)$ par la formule suivante :

$$\hat{\pi}(X_i) = \mathbb{P}(\mathrm{Y} = 1 | X = X_i) = \frac{e^{\hat{\beta}\prime X_i}}{1 + e^{\hat{\beta}\prime X_i}}$$

2. Les méthodes de correction de la régression logistique

Pour remédier aux inconvénients générés par le problème du déséquilibre des classes plusieurs auteurs ont travaillé sur quelques méthodes. C'est le cas de Gary KING et Langche ZENG [6] qui présentent deux approches.

L'une des approches consiste à utiliser les méthodes de rééchantillonnage puis à appliquer une correction aux estimateurs du maximum de vraisemblance. Comme présenté plus haut, les estimateurs obtenus par une régression logistique en cas déséquilibre des classes sont biaisées et ne vérifient pas la propriété de convergence. Les méthodes suivantes ont été proposées pour réajuster les coefficients estimés par le maximum de vraisemblance en cas d'asymétrie des classes.

2.1 Ajustement préalable (Prior correction)

Les auteurs de cette méthode ont prouvé que lorsqu'on a recours à une méthode de rééchantillonnage de type « response-based sampling » puis qu'on applique la régression logistique, les coefficients liés aux variables explicatives restent convergeant à l'exception du coefficient β_0. Ils proposent alors un ajustement du coefficient β_0 selon la procédure suivante :

Dans un premier temps, la base de données initiale est rééchantillonée par l'une des techniques permettant de rééquilibrer la base. Ensuite, une procédure de régression logistique classique est appliquée à la nouvelle base. Les coefficients des variables explicatives, obtenus sont conservés et le coefficient β_0 est réestimé par la formule suivante :

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

$$\tilde{\beta}_1 = \hat{\beta}_1$$

$$\vdots$$

$$\tilde{\beta}_k = \hat{\beta}_k$$

Où :

- $\hat{\beta}_{i,i=0,\dots,k}$ représente les coefficients estimés par le maximum de vraisemblance
- $\tilde{\beta}_{i,i=0,\dots,k}$ représente les coefficients corrigés
- τ représente la proportion de 1 (sinistres) dans la population initiale
- $\bar{y}$ représente la proportion de 1 (sinistres) dans la base rééchantillonée.

Cet ajustement des coefficients de l'estimation, entraine une modification dans la probabilité de sinistre calculée qui devient :

$$\tilde{\pi}(X_i) = \mathbb{P}(\mathrm{Y} = 1|X = X_i) = \frac{e^{\tilde{\beta}\prime X_i}}{1 + e^{\tilde{\beta}\prime X_i}}$$

Avec $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_k)$ les coefficients ajustés

2.2 Méthode par pondération (weighting method)

La seconde approche consiste à pondérer les données de façon à réduire l'impact de la différence observée entre la proportion de sinistres dans la base initiale et la proportion de sinistres dans la nouvelle base (après rééchantillonnage). Cette méthode appelée « weighted maximum likehood estimation » a été développée par Manski et Lerman en 1977. Le principe de cette méthode consiste à maximiser la vraisemblance pondérée. La vraisemblance classique s'écrit :

$$L = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi(X_i)^{y_i} \left(1 - \pi(X_i)\right)^{1-y_i}$$

$$= \prod_{\{y_i=1\}} \pi(X_i) * \prod_{\{y_i=0\}} (1 - \pi(X_i))$$

$$\log(L) = \sum_{\{y_i=1\}} \log\left(\pi(X_i)\right) + \sum_{\{y_i=0\}} log\left(1 - \pi(X_i)\right)$$

Alors la log vraisemblance pondérée log (L_w) s'obtient par la formule :

$$Log(L_w) = W_1 \sum_{\{y_i=1\}} \log(\pi(X_i)) + W_0 \sum_{\{y_i=0\}} log\left(1 - \pi(X_i)\right)$$

Avec :

- $W_0 = \frac{1-\tau}{1-\bar{y}}$
- $W_1 = \frac{\tau}{\bar{y}}$

Pour généraliser on écrit : $W_i = W_0 * (1 - Y_i) + Y_i * W_1$

L'avantage de cette méthode est qu'elle peut s'appliquer à n'importe quel modèle de régression contrairement à la méthode précédente qui s'applique uniquement à la régression logistique.

3. Evaluation de la qualité des modèles

3.1 Evaluation de la qualité du modèle : les pseudos R^2

Pour évaluer la qualité des modèles de régressions logistiques, les pseudos R^2 sont de bons indicateurs. Ils sont calculés à partir de ratios de la vraisemblance du modèle trivial contenant uniquement la constante et de la vraisemblance du modèle étudié constitué de l'ensemble des variables explicatives. Ils permettent d'apprécier la qualité du modèle étudié en vérifiant que ce dernier est meilleur comparé au modèle trivial. Il existe 3 types de pseudo R^2 dont le détail des formules est présenté ci-dessous. En notant L_0 la vraisemblance du modèle trivial et L_M celle du modèle étudié, on a :

♣ R^2 de McFadden $R^2_{MF} = 1 - (\frac{log\ L_M}{log L_0})$

♣ R^2de Cox & Snell $R^2_{CS} = 1 - (\frac{L_0}{L_M})^{2/n}$

♣ R^2 ajusté de Nagelkerke (par rapport au R^2 maximum possible) $R^2_N = \frac{R^2_{CS}}{1-L_0^{2/n}}$

Les pseudos R^2 sont similaires au coefficient de régression R^2 pour une régression classique. Selon Menard (2000), le R^2 de McFadden serait le R^2 le plus adapté lorsqu'on réalise une régression logistique car il n'est pas sensible aux variations de proportion de positifs dans la base étudiée [15]. C'est donc le R^2 de McFadden que nous utiliserons pour évaluer l'ajustement des modèles.

3.2 Performance des modèles

➢ Tableau de contingence

Le tableau de contingence permet de confronter les résultats obtenus par le modèle aux observations. Il se présente sous la forme suivante :

	Sinistre (y_i=1)	**Non-sinistre (y_i=0)**
Prédit sinistre ($\tilde{y}_i$=1)	**VP**	**FP**
Prédit non sinistre ($\tilde{y}_i$=0)	**FN**	**VN**

Où :

- VP est le nombre de vrais positifs c'est-à-dire ceux qui ont été prédit comme sinistre et pour qui on observe un sinistre en réalité.
- VN est le nombre de vrais négatifs c'est-à-dire ceux qui ont été prédit comme sains et pour qui on n'observe pas de sinistre en réalité.
- FP est le nombre de faux positifs c'est-à-dire ceux qui ont été prédit comme sinistre et pour qui on n'observe pas de sinistre en réalité.
- FN est le nombre de faux négatifs c'est-à-dire ceux qui ont été prédit comme sains et pour qui on observe un sinistre en réalité.

A partir du tableau, les indicateurs suivant sont calculés pour évaluer les performances prédictives du modèle étudié:

- Le taux de bonnes prédictions qui se calcule par l'expression $\frac{VP+VN}{\text{VP+VN+FP+FN}}$
- Le taux de mauvaises prédictions qui se calcule par $\frac{FP+FN}{\text{VP+VN+FP+FN}}$
- La sensibilité se définit comme la proportion de positifs bien prédits parmi les observations de positifs et se calcule par la formule suivante : sensibilité $= \frac{VP}{VP+FN}$
- La spécificité est le taux de vrais négatifs parmi les observations de négatifs et se calcule par la formule suivante : spécificité $= \frac{VN}{VN+FP}$

En régression logistique, les résultats de la matrice de confusion sont conditionnés par le choix d'un seuil de probabilité. En effet, pour attribuer les observations à l'une ou l'autre des classes de la variable d'intérêt, on s'appuie sur les probabilités construites par le modèle de régression logistique. L'estimation $\tilde{y}_i$ de la variable réponse par le modèle permet d'obtenir les probabilités associées. Il est alors possible de connaitre le nombre de bonnes et de mauvaises prédictions par rapport à un seuil donné. Pour un seuil fixé à 50%, tous les éléments dont la probabilité est supérieure à 50% seront classés dans la modalité 1 (réalisation du sinistre par exemple) et les éléments ayant une probabilité inférieure, seront classés dans la modalité 0 (absence de sinistre par exemple). Il est donc indispensable de bien choisir le seuil de probabilité si l'on veut utiliser la matrice de confusion. On peut calculer le seuil optimal et en déduire la matrice de confusion optimale ou s'appuyer sur d'autres indicateurs tels que la courbe ROC.

- **Performance du modèle : courbe ROC**

La courbe ROC (Receiver Operating Characteristic curve) est un outil très intéressant lorsqu'on veut évaluer et comparer la performance de modèles. Il s'agit en fait d'une courbe représentant le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1-spécificité). L'idée derrière cette courbe est de faire varier la probabilité seuil entre 0 et 1, et de regarder l'évolution de ces deux paramètres (sensibilité et 1-spécificité). Elle se représente en général avec la bissectrice et l'interprétation est que plus la courbe est éloignée de la bissectrice en terme d'aire et meilleur est le modèle.

- **L'indice de Youden : choix du seuil optimal**

Pour le choix du seuil optimal, nous nous appuyons sur l'indice de Youden en conjonction avec la courbe ROC. L'indice de Youden peut être défini comme une mesure de la performance d'un modèle. Il se calcule par la formule suivante :

$$Indice\ de\ Youden\ =\ (sensibilité\ +\ spécificité\ -\ 1)$$

Il se définit aussi comme la différence entre le taux de vrais positifs et le taux de faux positifs. Il prend ses valeurs dans l'intervalle [-1 ; 1]. Lorsque l'indice est négatif, le modèle est considéré comme non performant. Lorsqu'il est proche de 1, alors le modèle est performant. L'indice de Youden permet de déterminer le cut-off optimal utilisé pour réaliser une bonne discrimination des individus. Le seuil optimal est alors obtenu par la formule suivante :

$$Seuil\ optimal\ = max\ (sensibilité\ +\ spécificité - 1)$$

Notons que dans certaines littératures, l'indice de Youden est définit comme étant le point maximal lui-même. Le graphique représente une courbe ROC et l'indice de Youden :

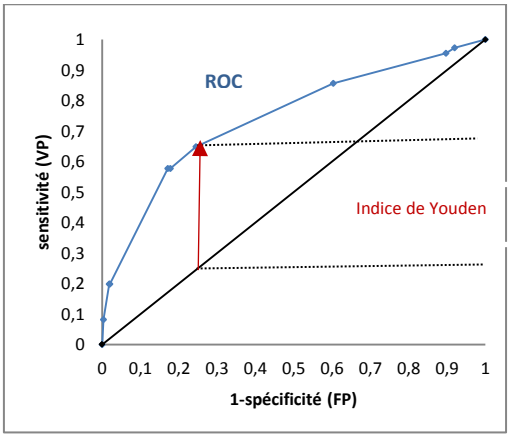


Figure 18: Représentation de la courbe ROC et l'index de Youden

L'AUC

L'un des indicateurs associés à la courbe ROC est l'AUC (Area Under the Curve). Il s'agit de l'aire comprise entre la courbe de discrimination et la première bissectrice. L'AUC correspond à la probabilité pour qu'un individu soit bien classé par le test en fonction du seuil choisi. Plus l'AUC est grand et meilleur est le modèle. De façon générale, l'interprétation de cet indicateur est la suivante :

Valeur AUC	Interprétation
0,5<AUC ≤0,6	Discrimination médiocre
0,6<AUC ≤0,7	Discrimination moyenne
0,7<AUC ≤0,8	Discrimination bonne
0,8<AUC ≤0,9	Discrimination très bonne
0,9<AUC ≤1	Discrimination excellente

Le GINI

C'est aussi un indicateur de performance d'un modèle. Il est le double de la surface entre la courbe ROC et la diagonale et se calcule par la formule :

$$Gini = 2AUC - 1$$

Ce coefficient est compris entre 0 et 1 et plus il est proche de 1 et meilleur est le modèle.

Chapitre 3: Théorie sur les techniques d'apprentissage

En apprentissage, le déséquilibre des classes peut conduire à une mauvaise performance des classificateurs. Dans le cadre de ce mémoire, nous avons choisi de travailler sur deux techniques permettant corriger le déséquilibre des classes à savoir le rééchantillonnage de la base et l'utilisation des méthodes ensemblistes telles que le *boosting*. Dans la suite, nous introduirons la notion d'apprentissage puis nous présenterons ces méthodes.

1. Généralités sur l'apprentissage et les arbres de classification

L'apprentissage désigne un ensemble de techniques permettant de classer des données. Il existe deux types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage non-supervisé consiste à classer les données à partir de leur valeur sans tenir compte d'une quelconque variable explicative ce qui aboutit à regrouper les données selon leur similarité. A contrario, en apprentissage supervisé, à partir de variables qualitatives ou quantitatives caractéristiques, on cherche à définir des règles permettant de répartir des objets en classes. Supposons que l'on dispose des données suivantes :

- y la variable à prédire ou expliquer
- x les attributs ou variables explicatives
- $(x_1, y_1), \dots, (x_m, y_m)$ les données appelés « exemples »

Nous souhaitons définir une règle permettant de prédire la variable y à partir des attributs x pour toute nouvelle donnée. Pour ce faire, il est nécessaire de choisir un algorithme d'apprentissage qui permet de classer les exemples. Nous avons choisi de nous intéresser aux arbres de décisions qui sont souvent utilisés comme classificateurs.

Les arbres de décisions regroupent les arbres de régression et les arbres de classification. Lorsque la variable Y est quantitative on parle d'arbre de régression et lorsqu'elle est catégorielle on parle d'arbre de classification. Les arbres de classification sont des outils non paramétriques de segmentation. Le but d'un arbre de classification est de répartir les individus en classes en se basant sur les variables explicatives. Ces méthodes de segmentation, ont la particularité de définir les résultats de la classification à partir de règles logiques ce qui rend les résultats plus faciles à interpréter du fait de la visualisation sous forme d'arbre.

1.1 Introduction aux arbres de classification

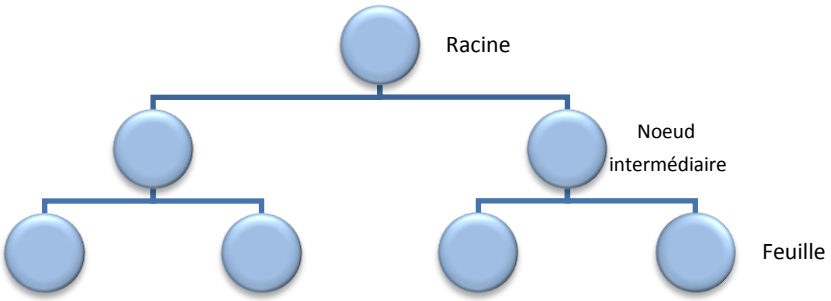
Grâce à l'essor des méthodes d'apprentissage automatique, les arbres de décisions sont de plus en plus utilisés en statistique et data mining. Un arbre de décision est constitué d'un ensemble de règles permettant de classer les données. En se basant sur une série de tests sur des variables appelées attributs, ils permettent au prédicteur de fournir une réponse parmi un ensemble discret de possibilités. L'idée est de construire des classes à

partir des données et des conditions sur les attributs dont on dispose et de pouvoir déterminer la classe correspondante à une nouvelle observation.

- **Structure des arbres de classification**

Un arbre de décision est constitué :

- de nœuds en arborescence
- de branches



Le nœud initial est appelé racine de l'arbre et représente toutes les observations. Chaque nœud représente un test sur les attributs. Le nœud est défini par le choix d'une variable parmi les variables explicatives et d'une division qui implique une partition en classes. Lorsque le nombre de classes est égal à deux on parle d'arbre binaire de décision.

Chaque branche représente l'une des valeurs possibles prises par les attributs à l'issue du test réalisé dans le nœud dont elle découle. Les nœuds terminaux sont appelés feuilles. Ils sont représentatifs des groupes crées à partir des règles de classification construites.

- **Apprentissage des arbres de décisions**

Un arbre de décision met en jeu deux types de variables : une variable réponse et des variables explicatives. Les arbres de décisions sont construits autour d'un algorithme. Les choix des attributs à placer à chaque nœud sont réalisés de telle sorte que l'arbre construit soit de taille raisonnable et ait une bonne capacité de prédiction. Pour implémenter une méthode d'apprentissage, la base de données doit être séparée en base d'apprentissage et en base test. La première est utilisée pour l'apprentissage tandis que l'autre sert à tester la qualité prédictive du modèle. L'algorithme de construction est constitué des étapes suivantes :

1. Définition de l'ordre dans lequel les variables explicatives seront examinées dans l'arbre.
2. Définition d'une règle permettant de choisir la meilleure division possible parmi les possibilités pour chaque variable explicative.
3. Définition d'une règle permettant de stopper l'évolution de l'arbre à un nœud donné c'est-à-dire transformer le nœud en feuille.
4. Attribution de chaque nœud terminal à l'une des modalités de la variable réponse.

Critère d'arrêt

Pour ce qui est du troisième point, de façon générale, un nœud devient terminal

- Soit lorsqu'il est homogène c'est-à-dire que toutes les observations du nœud appartiennent à la même classe et donc il n'existe plus de possibilité de partitionnement.
- Soit parce que le nombre d'observations dans le nœud est faible et inférieur à une valeur seuil fixée dans l'algorithme.

Règle d'affectation

En ce qui concerne le quatrième point, l'attribution de chaque nœud terminal à l'une des modalités de la variable qualitative à expliquer se fait en général en considérant la classe la mieux représentée dans le nœud.

Les critères de segmentation

Ce sont des critères statistiques qui permettront de répartir au mieux les observations selon les différents attributs possibles. Les critères les plus souvent utilisés sont :

- L'entropie

$$Gini(S) = -\sum_{i=1}^{K} \frac{N_i}{N} log\left(\frac{N_i}{N}\right)$$

- L'indice de GINI

$$Gini(S) = \sum_{i=1}^{K} \frac{N_i}{N}\left(1 - \frac{N_i}{N}\right) = 1 - \sum_{i=1}^{K} \left(\frac{N_i}{N}\right)^2$$

Où :

N_i Nombre d'éléments ayant la modalité i

N Nombre d'éléments dans le nœud considéré

K Nombre de modalités de la variable cible

Taille optimale de l'arbre

Lorsque l'arbre construit est de taille maximale, les erreurs entre données observées et prédites sont minimisées sur la base d'apprentissage. Mais puisque le modèle est fortement lié aux échantillons ayant permis sa construction, il y a un risque de sur-apprentissage et de mauvaise estimations des classes pour toute nouvelle observation de la base test. Pour éviter cela et produire des modèles performants aussi bien en apprentissage qu'en test, les arbres sont élagués.

L'élagage est une technique permettant de réduire la taille de l'arbre de façon à éviter que l'on ne se retrouve avec des nœuds contenant un seul élément. Elle se décline sous deux formes :

- L'élagage a priori consiste à introduire un critère d'arrêt d'expansion lors de la construction de l'arbre. Plus concrètement, lorsque ce critère s'applique, l'expansion de l'arbre est freinée quand bien même il resterait des attributs n'ayant pas encore été utilisés pour classer les exemples. Il peut s'agir par exemple de fixer un nombre minimum d'observations dans les nœuds.

- L'élagage a posteriori est la technique la plus utilisée dans les algorithmes d'apprentissage. Elle consiste à agir a posteriori sur la taille de l'arbre une fois son expansion maximale atteinte. A partir d'une base autre que celle d'apprentissage, un emboitement de sous arbres est construit. Les nœuds terminaux sont réexaminés et les moins significatifs sont supprimés de façon à obtenir un arbre de taille optimale. S'il existe un nœud pour lequel la division n'entraine pas un gain de significativité dans l'estimation, alors les feuilles de ce nœud peuvent être supprimées. En d'autres termes, tant qu'il existe un sous arbre que l'on peut remplacer par une feuille sans faire augmenter de façon significative l'erreur de mauvais classement, alors on élague ce sous arbre en réunissant les feuilles qui le composent. On remonte ainsi progressivement le long de l'arbre jusqu'à ce qu'on ne puisse plus remplacer un sous arbre par une feuille sans perdre en significativité.

On distingue plusieurs algorithmes de classification dont les plus connus sont : C5.0, CART (Classification and Regression Tree), CHAID (Chi-Square Automatic Interaction Detection), QUEST (Quick Unbiased Efficient Statistical Trees).

L'algorithme C5.0 est la version la plus récente et améliorée des algorithmes ID3 (Inductive Decision Tree) et C4.5. La séparation des observations est basée sur le critère d'entropie.

Pour la méthode CART, le critère de sélection utilisé est l'indice de GINI. Elle aboutit uniquement à la construction d'arbres binaires. La littérature existante met l'accent sur le fait que cette méthode est plus adaptée lorsque la finalité est de faire une prédiction.

La méthode CHAID permet de construire des arbres non binaires avec plusieurs ramifications. Cette méthode utilise le test du khi-deux comme critère de sélection et semble donc plus appropriée lorsque le but de l'étude est une analyse de l'impact des variables explicatives sur la variable à expliquer.

Le risque lorsqu'on construit un arbre est que celui-ci soit trop détaillé auquel cas, on est soumis à une instabilité et une mauvaise estimation de la prévision pour toute autre observation ou à l'inverse un arbre très petit ne contenant pas suffisamment d'informations pour réaliser une bonne prédiction. La différence entre la dernière méthode et les deux précédentes est la façon dont est traitée la réponse à ce problème. Pour les méthodes C5.0 et CART, on construit l'arbre maximal possible puis on sélectionne le sous arbre optimal par une procédure dite d'élagage alors que pour la méthode CHAID,

la taille de l'arbre est limitée puisque chaque nœud n'est construit que s'il remplit un critère statistique minimum de significativité construit à partir de la statistique du test du khi-deux.

La méthode QUEST est plus intéressante que les autres méthodes lorsque les variables explicatives catégorielles ont un grand nombre de modalités.

Pour ces différentes raisons, notre choix s'est porté sur l'arbre CART. Il est plus adapté pour prédire une variable ayant deux modalités telle que la survenance de sinistre.

1.2 L'arbre de classification CART

L'arbre CART est un arbre binaire de classification qui utilise comme critère de segmentation l'indice de GINI. Cet indice ne doit pas être confondu avec l'indice de GINI calculé à partir de l'AUC. Il peut être calculé en sommant le produit des probabilités d'être bien classé et d'être mal classé. Lorsque tous les éléments du nœud considéré appartiennent à la même modalité de la variable d'intérêt, il atteint sa valeur minimale. Le critère de Gini peut être évalué par la formule suivante :

$$Gini(S) = \sum_{i=1}^{2} \frac{N_i}{N}\left(1 - \frac{N_i}{N}\right) = 1 - \sum_{i=1}^{2} \left(\frac{N_i}{N}\right)^2$$

Où :

N_i Nombre d'éléments ayant la modalité i

N Nombre d'éléments dans le nœud considéré

Construction de l'arbre CART

La construction de l'arbre CART se fait en deux phases. Dans un premier temps, l'arbre maximal est construit puis dans un second temps les feuilles les moins significatives sont supprimées de façon à avoir un arbre de taille optimale et moins sujet au sur-apprentissage.

1- Construction de l'arbre maximal

Considérons les éléments suivants :

- Y la variable d'intérêt qualitative possédant l modalités $Y \in \{1,2, \dots, l\}$
- $X = (X_1, X_2, \dots, X_k)$, le vecteur de variables explicatives quantitatives ou qualitatives.
- $(X_i, Y_i)_{i=1,\dots,n}$ Les exemples disponibles dans la base d'apprentissage.

L'objectif en construisant un arbre est de pouvoir déterminer à partir de ces éléments, un classificateur $C: X \to m \; avec \; m \in \{1,2, \dots, l\}$ permettant de classer toute nouvelle observation pour laquelle on dispose des réalisations des variables explicatives. Au départ l'arbre est constitué d'un seul nœud contenant tous les exemples et appelé racine

de l'arbre. Ce nœud est splitté en plusieurs branches selon une variable de coupure. Cette variable de coupure est choisie de façon à créer des nœuds les plus homogènes possibles en s'appuyant sur le critère de Gini. La variable de coupure retenue pour le nœud considéré est celle qui minimise le critère de Gini. Pour un arbre binaire, le critère de Gini au nœud S, se calcule suivant la formule :

$$Gini(S) = \frac{N_1}{N}\left(1 - \frac{N_1}{N}\right) + \frac{N_2}{N}\left(1 - \frac{N_2}{N}\right)$$

Où: N_i nombre d'éléments ayant la modalité i et N nombre d'éléments dans le nœud considéré.

Lors du split du nœud S, la classification des exemples dépend du type qualitatif ou quantitatif de la variable de coupure X_i.

- S'il s'agit d'une variable quantitative X_i, un seuil de coupure α_s est défini :
  Si $x_j \leq \alpha_s$ alors l'observation j va dans le nœud fils gauche s_g sinon elle est classée dans le nœud fils droit s_d avec x_j la valeur de la variable X_i pour l'observation j
- S'il s'agit d'une variable qualitative, une partition en deux groupes de modalités $\{M_s;\ {}^cM_s\}$ est définie :
  Si $x_j \in M_s$ alors l'observation j va dans le nœud fils gauche s_g sinon elle est classée dans le nœud fils droit s_d avec x_j la valeur de la variable X_i pour l'observation j

Tous les attributs sont testés et celui retenu pour le split du nœud S en nœud fils S_g et S_d est celui qui maximise le gain informationnel qui se calcule selon la formule :

$$Gain(S, S_g, S_d) = gini(s) - p_d * gini(s_d) - p_g * gini(s_g)$$

Avec :

p_d la proportion de données du nœud S classée dans le nœud S_d

p_g la proportion de données du nœud S classée dans le nœud S_g

Ce processus est répété et se poursuit selon le principe suivant :

- Un nœud est identifié comme terminal dans l'une des situations suivantes :
  - Il comporte une seule observation
  - Tous les éléments du nœud appartiennent à la même classe
  - Le gain informationnel obtenu en découpant ce nœud est trop faible
- Si un nœud n'est pas terminal alors il est segmenté par un test sur un attribut selon le processus décrit précédemment en choisissant l'attribut qui maximise le gain informationnel
- Lorsque le nœud est terminal, on lui associe la classe majoritaire et on s'arrête lorsque tous les nœuds ont été classés.

2- Elagage

La stratégie d'élagage consiste à construire l'arbre maximal A_{max} puis à construire une suite de sous arbre en partant de l'arbre maximal. L'arbre optimal est ensuite choisi parmi la sélection d'arbres emboités. Pour cela, un critère de coût-complexité est introduit. Il s'obtient en sommant la proportion de mal classés sur les feuilles de l'arbre auquel on ajoute une pénalisation de la complexité de l'arbre. La complexité d'un arbre peut être exprimée par le nombre F de feuilles de celui-ci et dans ce cas une expression du critère coût-complexité est la suivante :

$$CC = \frac{1}{n}\sum_{j=1}^{n} 1_{C(X_j)\neq Y_j} + \gamma * \frac{F}{n}$$

Minimiser ce critère revient à minimiser l'expression suivante :

$$\sum_{j=1}^{n} 1_{C(X_j)\neq Y_j} + \gamma * F$$

Lorsque $\gamma = 0$, la complexité de l'arbre n'est pas pénalisé et l'arbre qui minimise le critère coût complexité est l'arbre maximal ayant F feuilles $A_{max} = A_F$. En augmentant le coefficient de pénalisation γ, l'une des divisions de l'arbre A_F devient superflue et les feuilles obtenues lors de cette division sont regroupées au sein du nœud père qui devient alors terminal. L'arbre qui minimise le critère est alors A_{F-1}. Ce processus est réitéré jusqu'à obtenir la suite de sous arbres suivante :

$$A_{max} = A_F \supset A_{F-1} \supset \cdots \supset A_1 = S$$

L'arbre optimal est ensuite choisi parmi cette suite de sous arbres en minimisant le nombre de mauvaises prédictions sur l'ensemble de validation.

1.3 La notion d'erreur en apprentissage

Dans ce paragraphe nous présenterons brièvement différents concepts d'erreurs importants lorsqu'on s'intéresse aux méthodes de classification.

- **L'erreur de prédiction**

Un modèle d'apprentissage est construit à partir d'un échantillon donné. Une fois le modèle construit, il est intéressant de regarder la performance en termes de prédiction. Lorsque l'erreur de prédiction est calculée sur la base ayant servie à l'apprentissage on parle d'erreur en resubstitution. Lorsque l'erreur est calculée sur une base test différente de la base d'apprentissage on parle d'erreur en test. L'erreur en resubtitution n'est pas vraiment pertinente puisqu'elle ne reflète pas les performances du modèle et sa capacité à bien classer de nouvelles observations. De façon générale, l'erreur du modèle peut être décomposée en deux composantes le biais et la variance.

Erreur = (biais)² + variance

Biais :

Encore appelée erreur d'approximation, le biais traduit l'incapacité du modèle à traduire le concept que l'on définit comme la « vraie » fonction reliant la variable de classification Y, aux attributs (variables explicatives) X.

Variance :

Encore appelée erreur d'estimation, la variance se définit comme la sensibilité du modèle. C'est la variabilité du modèle due aux fluctuations de l'échantillon d'apprentissage.

En notant $g(x) = E(Y|X = x)$, la fonction de classification idéale alors l'erreur commise par le classificateur $\widehat{g_c}(x)$ a pour formule :

$$E\left[\left(\widehat{g_c}(x) - g(x)\right)^2\right] = E[\widehat{g_c}(x)^2] - 2g(x)E[\widehat{g_c}(x)] + g(x)^2$$

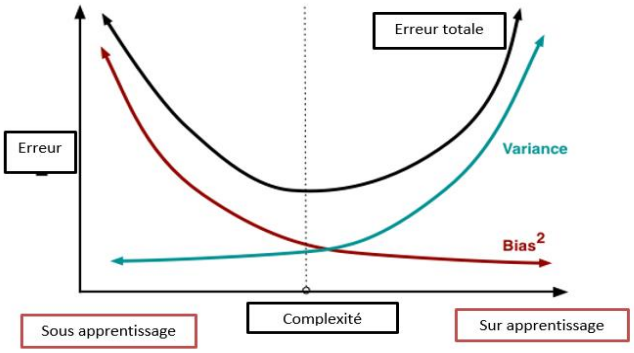
$$= E[\widehat{g_c}(x)^2] - E[\widehat{g_c}(x)]^2 + E[\widehat{g_c}(x)]^2 - 2g(x)E[\widehat{g_c}(x)] + g(x)^2$$

$$= E[\widehat{g_c}(x)^2] - E[\widehat{g_c}(x)]^2 + E\left[\left(\widehat{g_c}(x) - g(x)\right)^2\right]$$

$$= V\left(\widehat{g_c}(x)\right) + E\left[\left(\widehat{g_c}(x) - g(x)\right)^2\right]$$

$$Erreur = Variance + biais^2$$

Les modèles simples tels que les modèles linéaires, les modèles possédant peu de paramètres à estimer présentent un biais fort, mais une faible variance alors que les modèles complexes ayant beaucoup de paramètres à estimer présentent un faible biais, mais une forte variance. Cela pose le problème de la complexité du modèle et donc de sous-apprentissage et de sur-apprentissage.

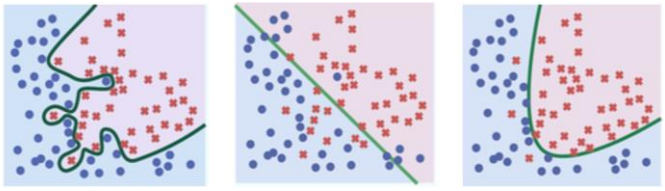


En apprentissage lorsqu'un modèle est complexe, la conséquence peut être un sur-apprentissage. Cette notion fait référence au fait que le modèle cherche à trop s'appuyer sur les données d'apprentissage au risque de mal prédire toute nouvelle donnée n'ayant pas participé à son estimation.

Dans le cas contraire, c'est-à-dire un modèle trop simple on risque un sous-apprentissage. Cette notion se traduit par le fait que pour sa construction, le modèle va s'appuyer sur peu

de données. Cela entraine une mauvaise adaptation du modèle aux données d'apprentissage avec une faible variance et un fort biais. Dans ce cas aussi la prédiction sera problématique.

Ainsi lorsqu'on applique une méthode de classification, l'objectif sera de minimiser l'erreur de prédiction. Le classificateur idéal est celui qui offre un bon compromis entre variance et biais. Les graphiques ci-dessous illustrent les situations de sur-apprentissage et de sous-apprentissage pour un classificateur donné.



Sur-apprentissage sous-apprentissage idéal

Compte tenu du risque de sur-apprentissage ou de sous-apprentissage, lors du découpage de la base en base d'apprentissage et en base de test, les proportions de découpage doivent permettre de disposer de suffisamment de données pour l'apprentissage sans pour autant pénaliser la base test.

- **L'erreur de validation croisée**

L'erreur de validation croisée consiste à séparer la base en N groupes puis à prédire le $k^{ème}$ groupe à partir de l'apprentissage effectué sur les N-1 groupes restants. Ce processus est répété pour les N groupes et pour chaque groupe l'erreur de prédiction est calculée. L'erreur de validation croisée s'obtient en faisant la moyenne des erreurs sur les N groupes.

Cette méthode est particulièrement intéressante lorsqu'on dispose de peu de données pour faire une partition apprentissage et test. Elle permet aussi d'avoir une estimation de l'erreur plus pertinente car le modèle est testé sur plusieurs bases.

2. Les méthodes d'agrégations des classificateurs

Ce sont des méthodes dites ensemblistes qui s'appuient sur des stratégies adaptatives ou aléatoires permettant d'améliorer l'ajustement et les performances de l'arbre de classification. Les algorithmes d'agrégation cherchent à améliorer les résultats du classificateur en jouant sur l'équilibre biais variance.

2.1 Le boosting

Afin d'améliorer les performances d'un classificateur, il est possible d'utiliser un algorithme de *boosting*. C'est l'une des méthodes proposées par WEISS [10] pour pallier aux faibles performances du classificateur en cas de déséquilibre des classes.

Le *boosting* comme son nom l'indique vise à améliorer le classificateur. C'est une méthode qui a fait ses preuves au fil des années et selon certains auteurs ce serait l'une des meilleures techniques introduite en apprentissage. Elle consiste à s'appuyer au fil de l'apprentissage, sur les bonnes hypothèses construites et à se focaliser sur les hypothèses moins bonnes en vue de les corriger. Le premier algorithme de *boosting* a été introduit par SCHAPIRE en 1989. Depuis il y a eu d'autres améliorations et l'algorithme le plus souvent utilisé est l'algorithme ADABOOST développé par FREUND et SCHAPIRE en 1996.

- ✓ **1er algorithme : boosting par sous-ensembles**

En notant S, l'échantillon d'étude et S_1, S_2, S_3 des sous-échantillons de S, l'objectif de cet algorithme est d'appliquer un algorithme d'apprentissage en l'occurrence un arbre de décision noté A sur les trois sous-ensembles de l'échantillon d'origine. Le processus de construction de cet algorithme est le suivant :

1. L'arbre A est appliqué sur l'échantillon S_1 et on obtient une première règle de décision (hypothèse) notée H_1
2. Ensuite on génère un deuxième échantillon S_2 dont les éléments sont choisis dans S- S_1 et tel que la moitié des éléments de S_2 sont mal classés par la règle H_1. On obtient donc une nouvelle règle de décision notée H_2.
3. Un troisième échantillon S_3 est construit tels que ses éléments sont choisis dans S-S_1-S_2 et présentent un désaccord de classement pour les règles H_1 et H_2. On obtient alors une troisième règle de décision notée H_3.

La règle finale H est obtenue par vote majoritaire des trois règles apprises.

H=vote majoritaire (H_1, H_2, H_3).

Cette méthode peut être utilisée de façon récursive et on peut faire varier le nombre de sous-ensembles par exemple 9,10,… Sur le plan théorique, les gains de cette méthode ont été prouvés. L'hypothèse H déduite des 3 hypothèses a une meilleure performance comparée à celle de l'hypothèse qui aurait été apprise directement sur l'échantillon S. cependant sur le plan pratique, les gains sont faibles ce qui explique les diverses améliorations qui ont été proposées.

- ✓ **ADABOOST : boosting probabiliste**

ADABOOST est un algorithme de *boosting* probabiliste contrairement à l'algorithme présenté dans le paragraphe précédent qui est de type ensembliste. L'idée à la base des algorithmes de *boosting* probabilistes est l'utilisation d'un paramètre de mise à jour adaptatif permettant de surpondérer (de donner plus d'importance) aux éléments

difficiles à prédire c'est-à-dire mal classés aux étapes précédentes de façon à forcer le classificateur à se focaliser sur ces derniers. L'algorithme est le suivant.

En entrée :

- un échantillon S de taille m
- un nombre d'itérations T

En sortie :

- une hypothèse globale H

Début

Soit

- $(x_1, y_1), \dots, (x_m, y_m)$ la base d'exemple de l'échantillon S
- $y_i \in \{-1, +1\}$ est la variable de classification, i=1,…, m
- $x_i \in X$ est le vecteur des attributs, i=1,…, m

Initialisation

Pour i=1,…, m

$$P_0(x_i) = \frac{1}{m}$$

Fin pour

t ← 0

Fin initialisation

Pour t=1,…,T

- Tirer un échantillon d'apprentissage S_t dans S selon les probabilités p_t
- Apprendre une règle de classification h_t sur S_t par l'arbre A
- Calculer ε_t l'erreur apparente de h_t sur S_t donc $\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} p_t(x_i)$
- Calculer $\alpha_t \leftarrow \frac{1}{2} log \frac{1-\varepsilon_t}{\varepsilon_t}$

Pour ∀ i = 1,…, m

$p_{t+1}(x_i) \leftarrow \frac{p_t(x_i)}{z_t} e^{-\alpha_t}$ si $h_t(x_i) = y_i$ c'est –à-dire exemple bien classé par h_t

$p_{t+1}(x_i) \leftarrow \frac{p_t(x_i)}{z_t} e^{+\alpha_t}$ si $h_t(x_i) \neq y_i$ c'est –à-dire exemple mal classé par h_t

(Z_t est une constante de normalisation telle que $\sum_{i=1}^{m} p_t(x_i) = 1$)

Fin pour

t ← t + 1

Fin pour

En sortie $H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$ où $\text{sign}(x)=\begin{cases} -1 \; si \; x < 0 \\ 0 \; si \; x = 0 \\ 1 \; si \; x > 0 \end{cases}$

Fin

Au cours des T étapes d'itérations, l'algorithme construit un échantillon S_t puis applique l'arbre à cet échantillon de façon à déduire une règle de décision h_t. Initialement, les exemples ont tous le même poids mais au fil des itérations, les poids sont augmentés ou réduits selon que l'exemple a été ou non bien classé par la règle h_t. Une nouvelle distribution de probabilité est alors formée en fonction des résultats de l'algorithme à l'étape précédente. Aussi à chaque t, le poids α_t de la règle h_t est calculé en fonction de l'erreur d'apprentissage liée à la règle de décision h_t. Cette erreur est en fait la somme des poids des exemples d'apprentissage mal classés. Le poids α_t est calculé de façon à donner plus de crédit aux règles ayant un faible taux d'erreur. Le principe général est donc de donner toujours plus de poids aux exemples mal classés pour qu'on améliore le classement par rapport à eux et de donner plus de poids aux règles ayant une bonne performance (faible erreur).

La règle finale est alors obtenue en sommant les règles h_t pondérées de leur poids α_t selon la formule $H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$.

L'avantage de cet algorithme est qu'il est rapide et peut être utilisé avec n'importe quel algorithme d'apprentissage mais l'inconvénient est que en mettant l'accent sur les mal classés on peut avoir des valeurs aberrantes dues à la croissance exponentielle des poids.

2.2 Le bagging

C'est une méthode qui est basée sur le principe du bootstrap. Elle consiste à sous échantillonner la base initiale et à appliquer un algorithme de classification pour chaque sous échantillon. Les sous échantillons sont tirés par tirage aléatoire uniforme des individus. Par rapport à l'algorithme choisi, on obtient alors un modèle différent pour chaque sous échantillon. Dans le cas d'un arbre de décision par exemple, on obtient des règles de décision différentes.

Comme pour le *boosting* vu précédemment, les règles obtenues pour chaque sous échantillons sont combinées de façon à définir une règle globale. L'hypothèse finale est alors calculée dans le cas d'une variable à prédire binaire selon la formule suivante :

$$H(x) = \text{décision majoritaire parmi les } h_b(x)$$

Où :

- B est le nombre de tirage
- h_b est l'hypothèse obtenue pour le sous-échantillon b

- H(x) est l'hypothèse globale

Lorsque le classificateur est l'arbre CART, alors il existe alors 3 possibilités pour la construction des arbres :

- Construire des arbres complets sans élagage
- Construire un arbre constitué d'au plus q feuilles
- Construire des arbres complets et élaguer par validation croisée

En pratique, c'est la première stratégie qui est souvent appliquée car elle est plus facile à mettre en place et ne nécessite pas le choix d'un paramètre contrairement à la deuxième méthode. Les étapes de l'algorithme du bagging sont décrites ci-dessous :

En entrée :

- Un échantillon S de taille m
- Un nombre B représentant le nombre d'itération du bootstrap

En sortie :

- Une hypothèse globale H

Début

Soit

- y la variable à prédire
- x les attributs ou variables explicatives
- $(x_1, y_1), \dots, (x_m, y_m)$ la base d'exemple de l'échantillon S
- h_b l'hypothèse apprise pour le sous-échantillon b

Pour b=1,..., B

- Tirer dans la base d'exemple aléatoirement avec remise un échantillon bootstrap s_b de taille m
- Apprendre une règle de classification h_b sur S_b par le classificateur A

Fin pour

Calculer la règle H(x) pour prédire y en agrégeant les règles h_b sur chacun des échantillons bootstrap.

- H(x) = décision majoritaire parmi les $h_b(x)$ si y est qualitative
- $H(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x)$ si y est quantitative

Retourner H(x)

Fin

Le principe général de cette méthode est d'agréger une collection de classificateurs pour obtenir un meilleur classificateur. Cette méthode permet d'avoir un modèle moins sensible aux données. En effet, dans le cas où l'algorithme appliqué est un arbre par exemple, cet arbre fournit un résultat en fonction de l'unique échantillon considéré et si le modèle est sensible à l'échantillon alors les résultats ne seront pas robustes. L'idée est d'introduire l'aléa sur l'échantillon à travers le bootstrap pour améliorer la prédiction du classificateur. De plus, si les hypothèses calculées pour chaque sous échantillon ont une forte variance (sensible à l'échantillon), la variance de l'hypothèse globale est réduite puisqu'on prend la moyenne des hypothèses h_b. En effet, en supposant que tous les modèles sont indépendants de variance σ^2alors la variance globale sera

$$V(H) = V(\frac{1}{B}\sum_{b=1}^{B} h_b(x))$$

$$= \frac{1}{B^2}\sum_{b=1}^{B} \sigma^2$$

$$= \frac{1}{B} * \sigma^2$$

Cette variance diminue lorsque le nombre d'échantillon bootstrap augmente. En cas de corrélation la variance globale est alors :

$$V(H) = V(\frac{1}{B}\sum_{b=1}^{B} h_b(x))$$

$$= \frac{1}{B^2}\sum_{b=1}^{B}(\sigma^2 + \sum_{b'=1;b'\neq b}^{B} \rho * \sigma^2)$$

$$= \frac{\sigma^2}{B^2}\sum_{b=1}^{B}(1 + \sum_{b'=1;b'\neq b}^{B} \rho)$$

$$= \frac{\sigma^2}{B^2}(B + B * (B-1) * \rho)$$

$$= \frac{\sigma^2}{B}(1 + (B-1)\rho)$$

$$= \frac{\sigma^2(1-\rho)}{B} + \sigma^2\rho$$

Avec ρ la corrélation entre chaque paire. Cette expression est composée de deux termes. Il est possible de réduire le premier terme en augmentant le nombre d'échantillon bootstrap B. Cependant, le second terme ne peut être réduit qu'en jouant sur le coefficient de corrélation.

L'avantage du bagging est qu'il est adapté aux algorithmes à forte variance tels que les réseaux de neurones et les arbres de décisions mais l'inconvénient est qu'il pourrait dégrader les performances des algorithmes plus stables tels que les K-ppv. De plus, les résultats ne sont pas faciles à interpréter. A cela s'ajoute le fait que les classificateurs des B sous-échantillons ne sont pas indépendants et se retrouvent même parfois trop corrélés car la méthode de construction des sous échantillons ne conduit pas toujours à créer des règles suffisamment diverses.

2.3 Les forêts aléatoires

Le principe des forêts aléatoires est de construire plusieurs arbres puis de les agréger. Cette méthode est introduite en 2001 par Breimann. Les forêts aléatoires sont des algorithmes de bagging adaptés aux arbres de décisions. L'objectif de cette méthode est de corriger une possible corrélation et de rendre les différents arbres CART construits par le bagging indépendants en introduisant un critère de décorrélation.

Comme présenté dans le paragraphe sur le bagging, la variance globale dont l'expression est la suivante,

$$V(H) = \frac{\sigma^2(1 - \rho)}{B} + \sigma^2\rho$$

peut être réduite en jouant sur le coefficient de corrélation ρ. Réduire la corrélation va donc diminuer la variance et rendre le modèle plus efficace. Pour cela, un aléa de plus est introduit dans la construction de chaque arbre. En effet, lors du choix des variables permettant de former la décision associée à un nœud, au lieu de tester toutes les variables explicatives admissibles comme dans un arbre CART classique, on tire uniformément q variables parmi les p variables. En général un choix optimal pour q est $q = \sqrt{p}$ lorsqu'on est face à un problème de classification. Ainsi l'algorithme est le suivant :

En entrée :

- Un échantillon S de taille m
- Un nombre B représentant le nombre d'itérations du bootstrap

En sortie :

- Une hypothèse globale H

Début

Soit

- y la variable à prédire
- x les attributs ou variables explicatives
- $(x_1, y_1), \dots, (x_m, y_m)$ la base d'exemple de l'échantillon S

- h_b l'hypothèse apprise pour le sous-échantillon b
- p le nombre de variables explicatives

Pour b=1,…, B

- Tirer dans la base d'exemple aléatoirement avec remise un échantillon bootstrap s_b de taille m
- Apprendre une règle de classification h_b sur S_b par le classificateur A avec une sélection aléatoire de $\sqrt{p}$ variables explicatives parmi p variables

Fin pour

Calculer la règle H(x) pour prédire y en agrégeant les règles h_b sur chacun des échantillons bootstrap.

- $H(x) =$ décision majoritaire parmi les $h_b(x)$ si y est qualitative
- $H(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x)$ si y est quantitative

Retourner H(x)

Fin

On remarque que cet algorithme est similaire à celui du bagging mais l'avantage principal de cette méthode par rapport au bagging est que le tirage aléatoire des variables explicatives à chaque nœud aboutit à des arbres non corrélés. Chacun des arbres est moins performant mais une fois regroupés, la performance est optimisée.

2.4 L'erreur OOB des méthodes ensemblistes

L'erreur OOB (out-of-bag error estimation) est une erreur calculée lorsqu'on applique les méthodes d'agrégations. L'objectif est de mesurer l'erreur directement durant l'apprentissage, sans avoir à passer par un échantillon test. Si l'on considère par exemple une forêt aléatoire constituée de 4 arbres, pour calculer l'erreur OOB de l'individu i, on considère les échantillons ne contenant pas cet individu et à partir des arbres construits sur ces échantillons on attribue à l'individu i, le vote majoritaire parmi les arbres. Cette opération est répétée pour tous les individus et l'erreur OOB est calculée par la proportion d'individus bien prédits par les arbres construits sur des échantillons ne les contenant pas.

3. Instruments d'évaluation de la prédiction

En apprentissage, les indicateurs les plus utilisés pour évaluer la prédiction sont le taux d'erreur et le taux de bonnes réponses. Dans le cas particulier d'une base présentant un déséquilibre des classes, ces outils ne sont pas toujours adaptés pour évaluer les performances. Lorsqu'ils sont utilisés, ils doivent être accompagnés d'autres indicateurs. Sur la base de données des ADBAI, en classant tous les individus en sains, le taux d'erreur

obtenu serait de 0,02% alors qu'aucun sinistre n'a été bien prédit. Ce taux d'erreur bien que faible reflète alors uniquement le bon classement des négatifs.

Le but étant de prédire les positifs, la courbe ROC, l'indice AUC et certains indicateurs de la matrice de confusion sont bien plus informatifs sur les performances du modèle. Dans la suite, nous présenterons les indicateurs utilisés en apprentissage en cas de déséquilibre des classes.

Les indicateurs sont calculés à partir de la matrice de confusion. Soit la matrice de confusion suivante :

	Sinistre ($\widetilde{y_i}$=1)	Non-sinistre ($\widetilde{y_i}$=0)
Prédit sinistre ($\widetilde{y_i}$=1)	VP	FP
Prédit non sinistre ($\widetilde{y_i}$=0)	FN	VN

Où :

- VP représente les vrais positifs
- VN représente les vrais négatifs
- FP représente les faux positifs
- FN représente les faux négatifs

Considérons les deux indicateurs suivants :

- Le taux de bonne prédiction

  Il représente la proportion d'observations bien prédites et se calcule par la formule $\frac{VP+VN}{VP+VN+FP+FN}$

- Le taux d'erreur

  Le taux d'erreur représente la proportion d'observations mal prédites. Il se calcule par la formule $\frac{FP+FN}{VP+VN+FP+FN}$

On remarque aisément que dans le cas où les individus positifs sont en infériorité numérique par rapport aux individus négatifs supposons respectivement 5% et 95%, en choisissant un classificateur naïf qui place toutes les observations en positifs, le taux d'erreur est seulement de 5% et le taux de bonne prédiction est de 95% mais pas un seul positif n'est bien prédit. Cela peut s'avérer problématique pour une comparaison de classificateurs. On dit alors que ces indicateurs sont sensibles au déséquilibre des classes [5].

Alors les indicateurs suivants sont utilisés en complément au taux de bonnes prédictions en situation de déséquilibre des classes :

- La précision

  Cet indicateur représente la proportion d'éléments positifs bien prédits parmi tous les prédits positifs et se calcule par la formule $\frac{VP}{VP+FP}$

- Le rappel

Cet indicateur représente la proportion d'éléments positifs bien prédits parmi tous les positifs observés et se calcule par la formule $\frac{VP}{VP+FN}$

- La spécificité

Cet indicateur représente la proportion d'éléments négatifs bien prédits parmi tous les négatifs observés et se calcule par la formule $\frac{VN}{VN+FP}$

- La F-mesure

Cet indicateur combine les deux précédents et se calcule par la formule $\frac{(1+\beta)^2 \times rappel \times precision}{\beta^2 \times rappel + precision}$

Où β est l'importance relative du rappel par rapport à la précision mais elle est en général fixée à 1. Cet indicateur atteint sa meilleure valeur en 1 et sa pire valeur en 0.

- La G-means

La G-means est la moyenne géométrique du rappel et de la précision. Plus connue sous le nom de Fowlkes–Mallows index, elle est utilisée pour déterminer la similarité entre deux groupes (les prédictions et les observations). Sa formule est la suivante :

$$\sqrt{\frac{VP}{VP+FN} \times \frac{VP}{VP+FP}} = \sqrt{rappel \times précision}$$

Plus la valeur de cet indicateur est grande, plus grande est la similarité entre les deux groupes considérés et donc meilleure est la prédiction.

- Le MCC

Le coefficient de corrélation de Matthew est un indicateur introduit par le biochimiste Bryan Matthew. Il prend en compte les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs. Il est considéré comme un indicateur de performance adapté au cas de déséquilibre des classes. Sa formule de calcul est la suivante :

$$\frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Ce coefficient retourne des valeurs comprises entre -1 et 1. Plus la valeur est proche de 1 et meilleure est la qualité de prédiction. Une valeur de -1 indique une totale discordance entre les prédictions et les observations.

- La courbe ROC

Il s'agit de la courbe représentant le taux de vrais positifs $\frac{VP}{VP+FN}$ en fonction du taux de faux positifs $\frac{FP}{FP+VN}$. Comme dans le cas de la régression logistique, cette courbe permet d'obtenir l'AUC et le Gini.

Chapitre 4: Comparaison et sélection des meilleurs modèles

Dans ce chapitre, nous présentons les résultats obtenus sur les données pour les différents modèles construits. La procédure de modélisation utilisée est la suivante :

Protocole de modélisation

1. Séparer la base en apprentissage et en test
2. Sur la base d'apprentissage on applique les 5 techniques de rééchantillonnage présentées ci-dessus
3. Application de la régression logistique classique à chacune des bases d'apprentissage obtenue

Pour une approche par régression logistique,

4. Appliquer les méthodes de correction de la régression logistique sur la base d'apprentissage
    - Prior correction
    - Weighting method
5. Appliquer le modèle obtenu sur la base test et calculer l'AUC, le GINI et le R^2 de Mcfadden
6. Calculer les indicateurs de performances MCC, G-means, F-mesure…

Pour une approche par apprentissage,

4. Appliquer les techniques d'apprentissage à la base d'apprentissage
    - arbre
    - arbre élagué
    - forêt aléatoire
    - bagging
    - boosting
5. Appliquer le modèle obtenu sur la base test et calculer l'AUC, le GINI
6. Appliquer le modèle sur la base test et calculer les indicateurs de performances MCC, G-means, F-mesure…

Afin de valider les différentes méthodes utilisées et de calculer leur performances nous divisons la base en apprentissage et en test. La première base est composée de 75% des observations et servira de base à l'apprentissage. Ce sont ces données qui permettront aux classificateurs d'apprendre des règles. La deuxième base contient le reste des données soit 25%. Cette base sera utilisée pour tester les performances prédictives des classificateurs.

Cette séparation a été réalisée de façon stratifiée par rapport à la variable de sinistralité par la proc Survey select de SAS. Cela implique que la proportion sains/sinistres observée dans la base initiale est conservée dans les bases d'apprentissage et de test. Cela permet d'éviter de se retrouver avec un modèle appris sur une base ne contenant quasiment pas de sinistres auquel cas le modèle construit ne serait pas pertinent. A des fins de comparaison, nous utilisons le même découpage pour l'approche par régression logistique et l'approche par apprentissage. Aussi les indicateurs introduits en apprentissage pour évaluer les performances prédictives des modèles seront calculés pour les modèles de régression logistique.

Après séparation de la base initiale, les bases obtenues se composent de façon suivante :

Apprentissage		**Test**	
sinistres	sains	sinistres	sains
83	40927	28	13643

Table 5: Effectif des base d'apprentissage et base test

A partir de la base d'apprentissage, chacune des techniques de rééchantillonnage présentée ci-dessus a été appliquée. Chaque rééchantillonnage a été réalisé de façon à obtenir une base équilibrée c'est-à-dire que les proportions de dossiers en sinistres et de dossiers sains sont équivalentes. Dans son mémoire «conversion modeling in direct motor insurance » Zhe Li a analysé les performances des modèles lorsque la proportion de rééquilibrage varie et il arrive à la conclusion qu'à partir du moment où les classes sont suffisamment représentées, la proportion de rééquilibrage n'impacte pas les performances des modèle. Nous avons donc opté pour un équilibrage 50%-50%. Pour la réalisation du rééchantillonnage de la base nous avons utilisé les packages DMwR et ROSE de R. La répartition en termes d'effectifs des nouvelles bases sont présentées ci-dessous :

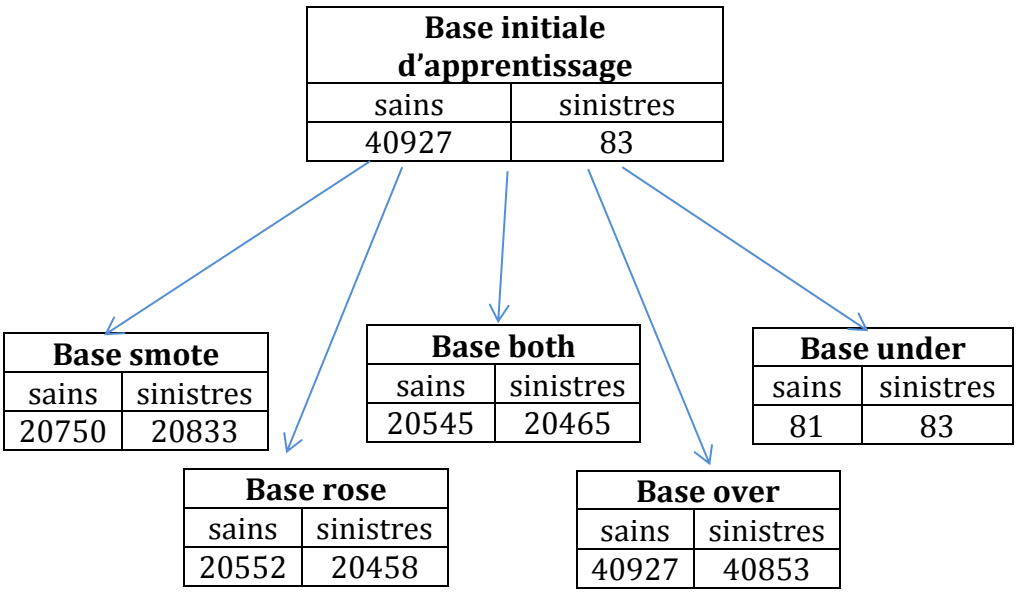


Table 6: Répartition des effectifs des bases rééchantillonnées

1. Approche par régression logistique

1.1 Recherche du modèle optimal de régression logistique

- **Proposition de modèles : choix des variables explicatives**

Dans ce paragraphe, nous recherchons le modèle optimal et pour cela il est important de définir les variables explicatives à prendre en compte dans notre modèle. Dans le premier modèle considéré, les individus dont le score n'est pas renseigné (Score_NC=1) sont séparés selon leur segment en Score_AI_NC et Score_ADB_NC. Les individus de référence pour ce modèle sont ceux ayant le score vert ou orange. Dans le second modèle, la variable Score_AI_NC est supprimée et donc les individus de référence sont les individus vert ou orange ou de classe Score_AI_NC. Il parait plus pertinent de considérer la variable de segment ADB sans l'associer à la variable de Score_NC car les tests et études réalisés dans le chapitre 3 de la partie 1 montre que le segment explique la survenance du sinistre. Le dernier modèle proposé répond à cette préoccupation. Les individus de référence sont alors les agents immobiliers classés vert.

Modèles	Variables explicatives
Modèle 1	Score_AI_NC Score_ADB_NC Score_R Score_Z
Modèle 2	Score_ADB_NC Score_O Score_R Score_Z
Modèle 3	Score_NC ADB Score_O Score_R Score_Z

Table 7: Variables explicatives par modèle

- **Les statistiques d'ajustements**

Afin de tester la qualité d'ajustement de ces modèles, les statistiques utilisées sont le critère AIC, le critère BIC, le critère -2log, la déviance et la statistique de Pearson. Ces critères permettent de comparer les modèles et de trouver le modèle optimal. Dans ce paragraphe, ces critères seront présentés brièvement ainsi que les résultats obtenus pour les modèles.

- **Le critère AIC**

Encore appelé critère d'information d'Akaiké, il permet d'évaluer la bonne adéquation du modèle et de le comparer à d'autres. Ce critère s'utilise pour les modèles dont l'estimation des paramètres est basée sur la méthode du maximum de vraisemblance. Il se calcule avec la formule suivante :

$$AIC = -2logL(\Theta) + 2k$$

Où :

- $L(\Theta)$ représente la vraisemblance maximisée
- k le nombre de paramètres dans le modèle

L'AIC pénalise les modèles permettant un grand nombre de variables. Le risque d'un modèle sur paramétré étant une mauvaise estimation des paramètres par le maximum de vraisemblance. L'AIC offre un bon compromis entre qualité d'ajustement des données au modèle et la parcimonie qui se définit comme le fait de décrire les données avec le plus petit nombre de paramètres possibles. Le modèle optimal est alors celui qui présente la plus petite valeur du critère AIC.

- **Le critère BIC**

Le critère d'information bayésien (BIC) comme le critère AIC, permet de comparer des modèles en se basant sur leur qualité d'ajustement et s'utilise pour les modèles dont les paramètres sont estimés par la méthode du maximum de vraisemblance. Cependant, en plus du nombre de paramètres du modèle, il se base sur le nombre d'observations. Il se calcule par la formule suivante :

$$BIC = -2logL(\Theta) + klog(n)$$

Où :

- $L(\Theta)$ représente la vraisemblance maximisée
- k le nombre de paramètres dans le modèle
- n le nombre d'observations

Le BIC est plus parcimonieux que l'AIC car il pénalise encore plus le nombre de variables présentes dans le modèle. Pour ce critère aussi, le meilleur modèle est celui ayant la plus petite valeur.

- **Le critère $-2logL(\Theta)$**

Pour ce critère, comme pour les deux précédents, on recherche les modèles permettant d'obtenir des petites valeurs. Contrairement à l'AIC ou le BIC, il ne tient compte ni du nombre de paramètres ni du nombre d'observations. Il ne pénalise donc pas les modèles ayant un grand nombre de variables.

Il se calcule par la formule :

$$-2\boldsymbol{logL}(\Theta)$$

Où :

- $L(\Theta)$ représente la vraisemblance maximisée

Les bases théoriques qui sous-tendent deux approches AIC et BIC sont différentes. L'AIC a été introduit pour retenir des variables jugées pertinentes lors de prévisions, tandis que le critère BIC vise la sélection de variables statistiquement significatives dans le modèle. Selon le professeur Brian Ripley, l'AIC est un critère prédictif tandis que le BIC est un critère explicatif. Les résultats obtenus sous SAS pour ces critères sont regroupés dans le tableau suivant.

statistiques	Modèle 1	Modèle 2	Modèle 3
AIC	1499 ,727	1496,104	1488,597
BIC	1544,273	1540,650	1542,052
-2log	1489,727	1486,104	1476,597

Table 8: Statistiques AIC, BIC, -2log des modèles construits

Pour ces 3 critères l'interprétation est la même. Plus le critère est petit et meilleur est le modèle. On voit alors qu'en s'appuyant sur le critère AIC,-2log, et BIC le meilleur modèle est le modèle 3.

- **Significativité des variables**

- **Significativité individuelle**

Les tests de significativité permettent d'identifier la contribution d'une ou de plusieurs variables explicatives dans l'estimation de la variable à expliquer. Il s'agira donc de tester la significativité des variables utilisées dans la régression pour chacun des 3 modèles spécifiés ci-dessus.

Il s'agit en fait d'un test de nullité du coefficient de la variable concernée. L'idée est de tester deux modèles : celui contenant toutes les variables explicatives sauf la variable dont on teste la significativité contre le modèle contenant toutes les variables explicatives. Pour un test sur la variable X_i , les hypothèses testées sont les suivantes :

H_0: $Logit\ P(\ Y\ =\ 1)\ =\ \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j$

H_1: $Logit\ P(\ Y\ =\ 1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_p x_p$

La statistique du test est $U = (\frac{\widehat{\beta_i}}{\hat{\sigma}(\beta_i)})^2$ Où : $\hat{\sigma}(\beta_i)$ est l'écart-type de l'estimateur du coefficient β_j. Sous l'hypothèse H_0 cette statistique suit une loi du khi-deux à 1 degré de liberté. Ainsi, l'interprétation du test se fait en comparant la statistique obtenue à une loi du khi-deux.

Si $U > X^2(1)$ alors H_0 est rejetée auquel cas le modèle contenant la variable x_i est préféré au modèle ne la contenant pas. En terme de p-value, l'hypothèse H_0 ne peut être rejetée

lorsque la valeur de la probabilité est supérieure au seuil fixé (ici α =5%). Dans le cas contraire, on refuse l'hypothèse nulle et on admet que la variable x_i a une influence sur la probabilité de survenance du sinistre. Les résultats obtenus sont les suivants :

Modèle 1

paramètre	DDL	estimateurs	Variance des estimateurs	Statistique du test	Pr >Chisq
Intercept	1	-6,9600	0,2042	1161,4801	<0,0001
Score_AI_NC	1	0,2150	0,2920	0,5422	***0,4615***
Score_ADB_NC	1	1,6800	0,2562	43,0011	<0,0001
Score_R	1	2,4141	0,3455	48,8135	<0,0001
Score_Z	1	3,5737	0,3957	81,5653	<0,0001

Modèle 2

paramètre	DDL	estimateurs	Variance des estimateurs	Statistique du test	Pr >Chisq
Intercept	1	-6,9828	0,1645	1802,9550	<0,0001
Score_ADB_NC	1	1,7039	0,2258	56,9448	<0,0001
Score_O	1	0,7880	0,3567	4,8793	0,0272
Score_R	1	2,4380	0,3236	56,7508	<0,0001
Score_Z	1	3,5975	0,3767	91,1931	<0,0001

Modèle 3

paramètre	DDL	estimateurs	Variance des estimateurs	Statistique du test	Pr >Chisq
Intercept	1	-8,5011	0,3442	609,9191	<0,0001
Score _NC	1	1,8021	0,3033	35,2970	<0,0001
ADB	1	1,3927	0,2314	36,2161	<0,0001
Score_O	1	1,1266	0,4144	7,3903	0,0066
Score_R	1	2,8188	0,3866	53,1670	<0,0001
Score_Z	1	4,0080	0,4328	85,7635	<0,0001

Table 9: Caractéristiques des estimateurs obtenus par la régression logistique

Après l'étude de ces tableaux, on conclut que pour les modèles 2 et 3, toutes les variables sont significatives. En revanche pour le modèle 1, la p-value obtenue pour la variable Score_AI_NC (0,4615) indique qu'elle ne contribue pas à l'estimation de la probabilité de survenance du sinistre.

- **Significativité globale**

Les tests de significativité globale consistent à tester la contribution de l'ensemble des variables explicatives du modèle à l'estimation de la variable d'intérêt. Il existe plusieurs tests. Ce sont entre autres le test du rapport de vraisemblance, le test du score et le test de Wald. Le modèle trivial ne contenant aucune variable explicative est testé contre le modèle contenant toutes les variables explicatives. Les hypothèses testées sont les suivantes :

H_0: $Logit\ P(\ Y\ =\ 1)\ =\ \beta_0$

H_1: $Logit\ P(\ Y\ =\ 1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_p x_p$

- En ce qui concerne le test du rapport de vraisemblance, la statistique du test est : $\boldsymbol{T_R = 2 * (\mathbf{Log}(\mathbf{L}\left(\mathbf{x}, \widehat{\boldsymbol{\beta}}\right) - \mathbf{Log}(\mathbf{L}(\mathbf{x}, \boldsymbol{\beta_0}))}$
  Où : $L\left(x, \hat{\beta}\right)$ et $L(x, \beta_0)$ sont respectivement la vraisemblance du modèle sous H1 et H0.
- Le test de Wald est basé sur la différence entre l'estimateur de vraisemblance et le coefficient β_0 . En pratique, cette différence est rapportée à la variance des estimateurs et le carré de ce rapport est comparé à une loi du khi-deux. Un estimateur raisonnable de la variance étant la matrice d'information de Fisher, il est alors possible d'utiliser la statistique suivante :
$$\boldsymbol{T_W = (\widehat{\beta} - \beta_0)' * I(\widehat{\beta}) * (\widehat{\beta} - \beta_0)}$$
- Le test du score est plutôt basé sur la dérivée de la log-vraisemblance en β. La statistique de ce test est la suivante :
$$\boldsymbol{T_S = (\frac{\delta \mathbf{Log}(\mathbf{L}(\mathbf{x}, \beta_0))}{\delta \beta})' * I(\beta_0)^{-1} * (\frac{\delta \mathbf{Log}(\mathbf{L}(\mathbf{x}, \beta_0))}{\delta \beta})}$$

  Pour ce test aussi, on peut montrer asymptotiquement que cette statistique tend, vers une loi du khi-deux.

Sous l'hypothèse H_0, ces statistiques suivent une loi du khi-deux à p degrés de liberté. Ainsi l'interprétation du test se fait en comparant la statistique obtenue à une loi du khi-deux.

Si $T > X^2(p)$ alors H_0 est rejetée auquel cas le modèle contenant toutes les variables est préféré au modèle trivial. En terme de p-value, l'hypothèse H_0 est acceptée lorsque la valeur de la probabilité est supérieure au seuil fixé (ici α =5%). Dans le cas contraire, on refuse l'hypothèse nulle et on admet que les variables explicatives ont simultanément une influence sur la probabilité de survenance du sinistre. Les résultats obtenus sont les suivants :

	Modèle 1		Modèle 2		Modèle 3	
Test	Chi-square	Pr >Chisq	Chi-square	Pr >Chisq	Chi-square	Pr >Chisq
Vraisemblance	108,39	<0,0001	112,01	<0,0001	121,52	<0,0001
Score	231,13	<0,0001	233,50	<0,0001	233,48	<0,0001
Wald	134,25	<0,0001	134,50	<0,0001	142,28	<0,0001

Table 10: Résultats par modèle des tests de significativité

On en déduit que pour tous les modèles, les variables explicatives ont simultanément un effet sur la variable réponse et ceci pour les tests de vraisemblance, de score et de Wald

- **Adéquation du modèle : test d'Hosmer et Lemeshow**

Le test de Hosmer-Lemeshow est basé sur une statistique qui mesure la qualité d'ajustement des modèles. Ce test est décrit en annexe du mémoire. Pour effectuer ce test, nous avons construits 5 groupes basés sur les 5 classes de scores. Les résultats pour les 3 modèles sont récapitulés dans le tableau suivant :

			Modèle 1	Modèle 2	Modèle 3
Groupe	Effectif	Observé	Théorique	Théorique	Théorique
Score_NC	27857	65	65,00003	60,11810	65,00000
Score_V	20393	14	19,33750	18,88194	14,00000
Score_O	4917	10	4,66251	10,00000	10,00000
Score_R	1239	13	13,00915	13,00946	13,00001
Score_Z	275	9	9,00002	9,00002	9,00000

	Chi-square	DF	Pr > ChiSq
Modèle 1	7,591	3	0,055
Modèle 2	1,661	3	0,646
Modèle 3	5,60608 E-12	3	0,999

Table 11: Résultats par modèle du test d'adéquation d'Hosmer-Lemeshow

Le test d'Hosmer-Lemeshow désigne aussi le modèle 3 comme meilleur modèle.

Au regard des statistiques d'ajustement et de prédiction, le modèle retenu est le 3ème modèle à savoir :

$$I\_sinistre \sim ADB + score\_NC + score\_O + score\_R + score\_Z$$

Une fois le modèle optimal de régression logistique défini et le rééchantillonnage de la base réalisé, les méthodes de correction de la régression logistique à savoir l'ajustement

préalable et la régression logistique pondérée sont appliquées à chacune des nouvelles bases.

Dans un premier temps nous appliquons la méthode des ajustements préalables puis nous appliquons celle des pondérations.

1.2 Ajustement préalable (Prior correction)

Les coefficients $\hat{\beta}_0$ obtenus par la régression logistique classique sont inscrits dans la 4ème colonne et les ajustements sont inscrits dans la dernière colonne. τ représente la proportion de 1 (sinistres) dans la population initiale. $\bar{y}$ représente la proportion de 1 (sinistres) dans la base rééchantillonnée.

Base	$\bar{y}$	τ	$\hat{\beta}_0$	Correction intercept $\hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$
smote	0,50099	0,2030 %	-1,2215	-7,42320
rose	0,49885	0,2030 %	-1,6596	-7,85272
both	0,49902	0,2030 %	1,2832	-7,85361
over	0,49955	0,2030 %	-1,6406	-7,83649
under	0,50609	0,2030 %	-1,4165	-7,13561

Table 12: Valeur des coefficients obtenus par la méthode d'ajustement préalable

Afin d'avoir une première idée de la qualité des modèles de correction construits, les indicateurs AUC, GINI et R^2 McFadden sont calculés sur la base d'apprentissage. Pour comparer les résultats, nous appliquons la régression logistique classique sur la base d'apprentissage et nous calculons les 3 indicateurs.

Modèles	R^2 McFadden	AUC	Gini
Smote corrigé	0,20534	0,68517	0,37035
Rose corrigé	0,25369	0,69812	0,39624
Both corrigé	*0,25369*	0,69966	0,39933
Over corrigé	0,25264	*0,70145*	*0,40291*
Under corrigé	0,14847	0,64056	0,28112

Table 13: Indice AUC, GINI et R^2 McFadden des modèles d'ajustement préalable sur la base d'apprentissage

Le meilleur coefficient de R^2 McFadden s'observe pour le modèle *Both corrigé* tandis que les meilleures valeurs du GINI et de l'AUC sont obtenues avec le modèle *over corrigé*. De façon générale, 3 modèles fournissent d'assez bon résultats. Il s'agit du *rose corrigé*, du *both corrigé* et d'*over corrigé*. Pour confirmer ces résultats les mêmes indicateurs sont recalculés mais cette fois sur la base de test.

Modèles	R^2 McFadden	AUC	Gini
Smote corrigé	0,17345	0,79720	0,59440
Rose corrigé	0,21749	0,85224	0,70449
Both corrigé	0,21761	0,85442	0,70884
Over corrigé	*0,218180*	*0,85983*	*0,71967*
Under corrigé	0,079185	0,71593	0,43186

Table 14: Indice AUC, GINI et R^2 McFadden des modèles d'ajustement préalable sur la base test

Les résultats sur la base test confirment le choix des trois modèles *both corrigé*, *rose corrigé* et *over corrigé*. Les trois indicateurs désignent l'over corrigé comme le meilleur modèle.

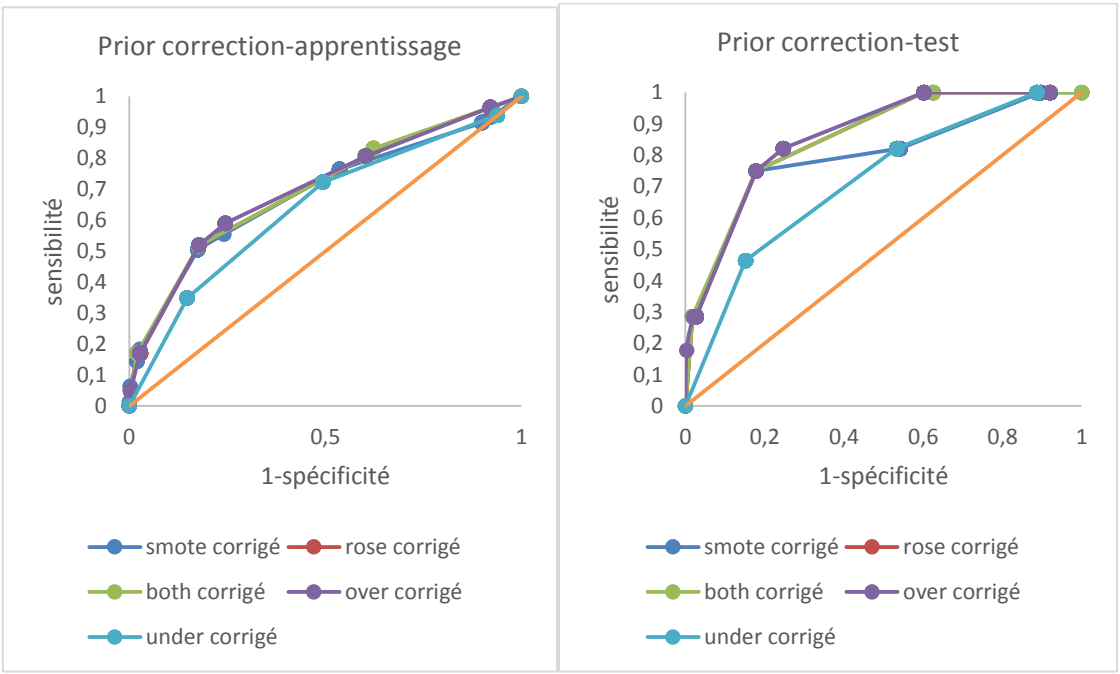


Figure 19 : Courbe ROC des modèles de correction par ajustement préalable sur la base test

Ce jeu de courbe ROC, permet de voir que les courbes des 3 modèles *both corrigé*, *rose corrigé* et *over corrigé* sont au-dessus des autres courbes ce qui implique qu'ils ont une meilleure qualité d'ajustement et un meilleur pouvoir discriminant.

1.3 Méthode par pondération (weighting method)

L'application de cette méthode est simple, il suffit de pondérer chaque observation d'un certain poids. Pour les observations ayant connu un sinistre, le poids appliqué est $W_1 = \frac{\tau}{\bar{y}}$

Pour les observations dont la variable sinistre a pour modalité 0, le poids est $W_0 = \frac{1-\tau}{1-\bar{y}}$

Avec :

- τ représente la proportion de 1 (sinistres) dans la population initiale
- $\bar{y}$ représente la proportion de 1 (sinistres) dans la base rééchantillonnée.

Modèles	τ	$\bar{y}$	W_0	W_1
Smote pondéré	0,00203	0,50102	0,00405175	2,00001335
Rose pondéré	0,00203	0,50028	0,00405770	1,99707219
Both pondéré	0,00203	0,50070	0,00405429	1,99875458
Over pondéré	0,00203	0,49931	0,00406563	1,99317853
Under pondéré	0,00203	0,49554	0,00409658	1,97827681

Table 15: Valeur des poids obtenus pour la méthode de pondération

Ensuite, la régression logistique est appliquée sur la base pondérée. Comme pour la méthode d'ajustement préalable, les indicateurs R^2 McFadden, AUC et GINI sont calculés sur la base d'apprentissage et sur la base test. Les résultats obtenus sont présentés ci-dessous :

- **Sur la base apprentissage**

Modèles	R^2 McFadden	AUC	Gini
Smote pondéré	0,26789	0,67363	0,34725
Rose pondéré	*0,27852*	0,69526	0,39053
Both pondéré	0,27835	*0,69592*	*0,39184*
Over pondéré	0,27625	0,69296	0,38592
Under pondéré	0,16997	0,64726	0,29451

Table 16: Indice AUC, GINI et R^2 McFadden des modèles par pondération sur la base d'apprentissage

- **Sur la base test**

Modèles	R^2 McFadden	AUC	Gini
Smote pondéré	0,23490	0,83703	0,67407
Rose pondéré	*0,24287*	0,83653	0,67306
Both pondéré	0,24278	*0,85129*	*0,70258*
Over pondéré	0,24147	0,83513	0,67026
Under pondéré	0,09639	0,70584	0,41168

Table 17: Indice AUC, GINI et R^2 McFadden des modèles par pondération sur la base de test

Les valeurs des indicateurs sur la base d'apprentissage et la base test permettent d'identifier deux modèles : le *rose pondéré* et le *both pondéré*. Le premier possède le meilleur coefficient R^2 McFadden et le second le meilleur indice de GINI.

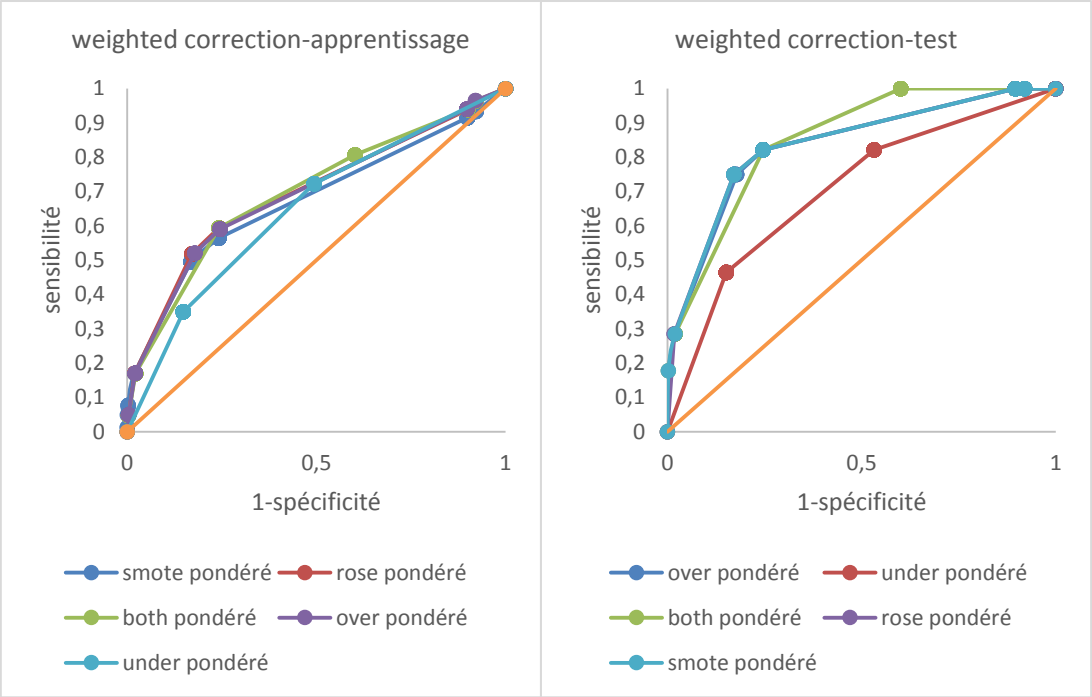


Figure 20 : Courbe ROC des modèles de correction par pondération sur la base test

De façon générale on remarque que pour les deux méthodes les meilleurs coefficients sont obtenus pour les bases rééchantillonnées par les algorithmes Rose et Both. Aussi les modèles par pondération présentent à l'exception de l'Under pondéré une meilleure qualité d'ajustement par rapport aux modèles d'ajustement préalable. A contrario ces derniers présentent de meilleures performances car les valeurs de l'indice de GINI sont plus élevées.

1.4 Performance prédictive des modèles

Pour évaluer le pouvoir prédictif des modèles, nous considérons les indicateurs suivant construits à partir de la matrice de confusion. Pour obtenir la matrice de confusion optimale, nous avons utilisé l'indice de Youden. Cet indice nous a permis de déterminer la probabilité seuil et de prédire l'appartenance des observations à l'une ou l'autre des classes de la variable de sinistralité.

Modèles	Erreur	Rappel	Spécificité	Précision	G-means	F-measure	MCC
smote corrigé	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
rose corrigé	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
both corrigé	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
over corrigé	0,2467	*0,8214*	0,7531	0,0068	0,7865	0,0135	0,0601
under corrigé	0,1522	0,4643	*0,8486*	0,0063	0,6277	0,0123	0,0394
smote pondéré	0,1716	0,7500	0,8286	*0,0089*	*0,7883*	*0,0176*	*0,0692*
rose pondéré	0,1716	0,7500	0,8286	*0,0089*	*0,7883*	*0,0176*	*0,0692*
both pondéré	0,2467	*0,8214*	0,7531	0,0068	0,7865	0,0135	0,0601
over pondéré	0,2467	*0,8214*	0,7531	0,0068	0,7865	0,0135	0,0601
under pondéré	*0,1522*	0,4643	0,8486	0,0063	0,6277	0,0123	0,0394

Table 18: Indicateurs de la qualité de prédiction des modèles par pondération et d'ajustement préalable

A la lecture de ces résultats les meilleurs modèles en termes de prédiction sont le *smote pondéré* et le *rose pondéré*. Les indicateurs MCC, G-means et F-mesure sont plutôt bons et le taux d'erreur de prédiction est l'un des plus faibles. Ces modèles offrent un bon compromis entre spécificité et précision c'est à dire bon équilibre entre le taux de bonnes prédictions dans la classe de positifs et de négatifs. Les performances de ces modèles étant identiques, le choix s'est porté sur celui qui est resté le plus constant en termes de performances d'ajustement à savoir le *rose pondéré.*

1.5 Stabilité du modèle

Dans un premier temps nous faisons le test d'Hosmer et Lemeshow pour tester l'adéquation du modèle aux données. Le résultat est satisfaisant car la p-value est supérieur au seuil 0,05.

Groupe	Effectif	Attendu	Observé
Score_NC	27857	65	65,07177
Score_V	20393	14	14,14886
Score_O	4917	10	10,02340
Score_R	1239	13	13,68953
Score_Z	275	9	10,57400

Chi-square	DF	Pr > ChiSq
7,591	3	0,964

Table 19: Résultats du test d'Hosmer-Lemeshow pour le modèle rose pondéré

Nous voulons tester la stabilité du modèle dans le temps. On observe donc l'évolution des estimateurs et de leur variance/covariance lorsqu'on rajoute une année supplémentaire à l'historique. Pour cette étude nous avons considéré l'historique 2005-2015. Partant des contrats en portefeuille en 2005, nous avons réalisé une régression logistique. Ensuite, nous rajoutons une année d'historique et nous recalculons la régression logistique sur le nouvel historique construit et ainsi de suite. Le graphique ci-dessous représente l'évolution des estimateurs pour chaque année supplémentaire :

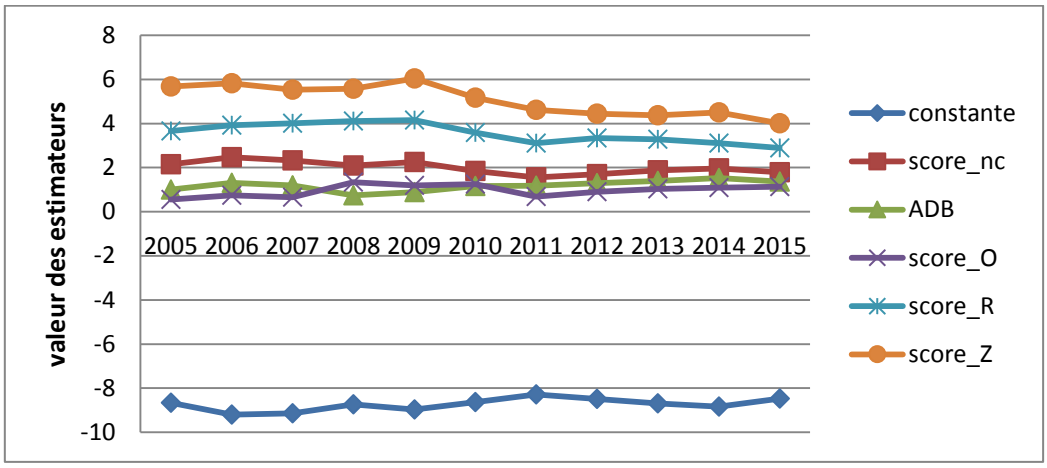


Figure 21: Evolution des estimateurs du modèle rose pondéré en fonction des années

On remarque que les estimateurs se stabilisent à partir de l'année 2011.L'évolution des variances des estimateurs entre 2005 et 2015 à chaque année supplémentaire est la suivante :

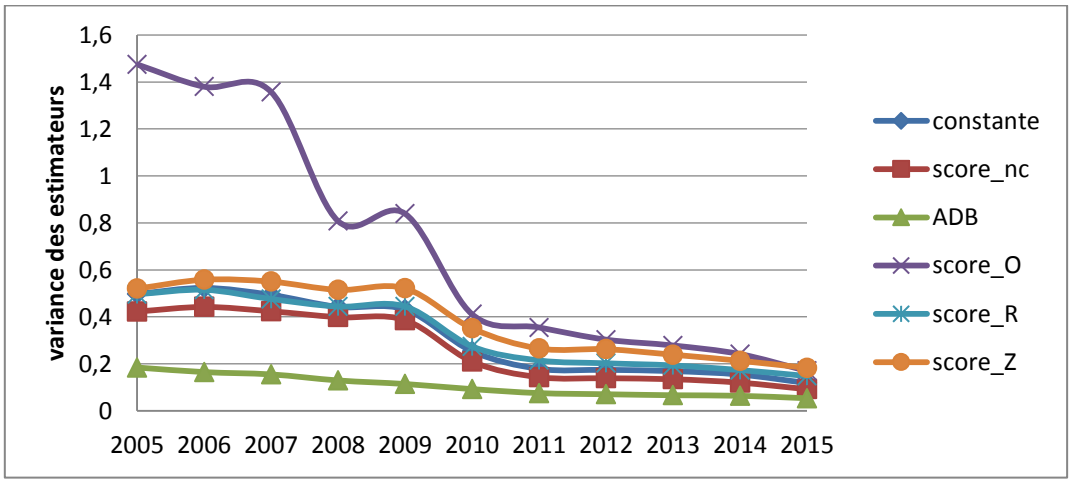


Figure 22 : Evolution de la variance des estimateurs du modèle rose pondéré en fonction des années

La variance des estimateurs évolue à la baisse au fur et à mesure qu'on ajoute une année d'historique. Ce qui implique que les estimateurs sont de plus en plus robustes au fil du temps. Le modèle sélectionné présente de bonne qualité d'ajustement, de prédiction et de robustesse.

2. Approche par apprentissage

Comme pour la régression logistique, la base initiale est séparée en base d'apprentissage et en base test dans l'optique de tester les performances des modèles sur des observations n'ayant pas servi à la construire. Les deux nouvelles bases sont construites en conservant dans chacune des bases la proportion de sinistralité dans la base initiale

Afin de voir le fonctionnement de l'arbre CART en cas de déséquilibre des classes, nous avons implémenté l'algorithme de construction de l'arbre sur la base d'apprentissage. En conséquence les arbres implémentés contenaient un unique nœud ce qui implique une absence totale de classification. Pour pallier cela, les 5 méthodes de rééchantillonnage précédemment introduites ont été appliquées à la base d'apprentissage selon une répartition équilibrée entre les classes.

2.1 Application de l'arbre CART

Le modèle retenu pour l'application de l'arbre CART est le même qu'en régression logistique :

$$I\_sinistre \sim ADB + score\_NC + score\_O + score\_R + score\_Z$$

Pour chacune de ces bases de données, nous avons appliqué l'arbre CART. Pour cela, la construction s'est faite en deux étapes. D'abord la construction de l'arbre le plus profond possible, ce qui permet de réduire le biais mais il existe un risque de sur-apprentissage.

Pour cela, dans une seconde phase, les arbres ont été élagués c'est-à-dire que certaines feuilles ont été supprimées selon le critère de minimisation de l'erreur de validation de façon à éviter un sur-apprentissage et à conserver de bonnes capacités prédictives. Aucun des arbres n'a eu besoin de phase d'élagage (l'arbre maximal est optimal) à l'exception de celui obtenu pour la base Under pour lequel, on est passé de 6 feuilles à 5 comme on peut le voir sur le tableau suivant, fourni par R :

Nb_feuille	CP	erreur
0	0,11111	1,20988
2	0,08642	1,06713
3	0,04938	0,76543
4	0,01235	0,65432
5	0,00000	0,71605

Table 20 : Erreur du modèle en fonction du coefficient de complexité

Lorsque le coefficient de pénalisation de la complexité augmente, le nombre de feuille de l'arbre est réduit. L'erreur de validation est plus importante lorsqu'on passe de 4 à 5 divisions. On réduit donc l'arbre pour conserver une erreur plus faible. Le graphique suivant représente l'arbre Under maximal et l'arbre Under élagué obtenu avec le package rpart de R.

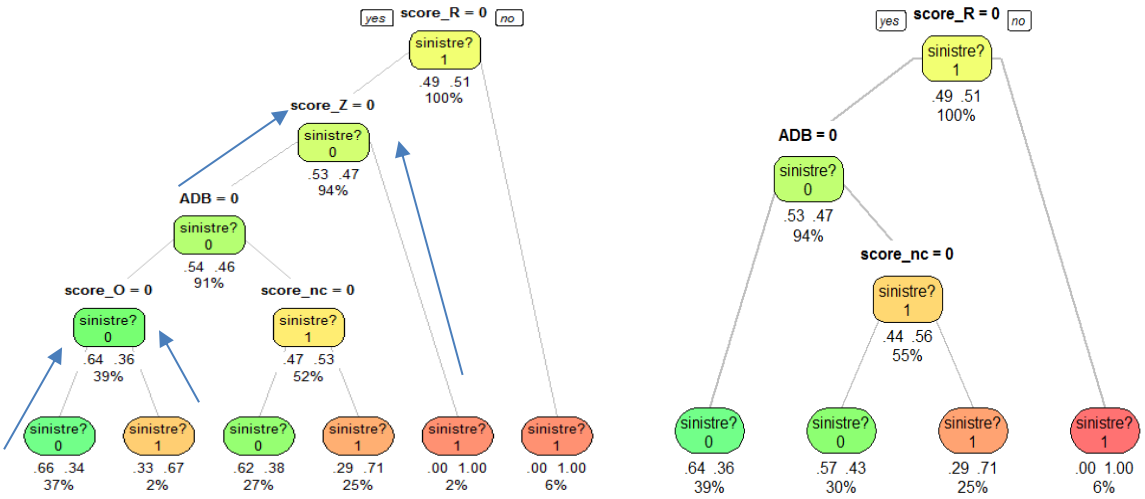


Figure 23 : Les arbres Under et Under élagué

Nous souhaitons tester et comparer les performances des arbres. Pour cela nous calculons l'indice de Gini, l'AUC et la courbe ROC.

Arbre	AUC	Indice de Gini
Arbre smote	0,7741	0,5482
Arbre rose	0,8528	0,7056
Arbre both	0,8528	0,7056
Arbre over	0,8545	0,709
Arbre Under	0,7812	0,5624
Arbre Under élagué	0,7269	0,4538

Table 21: Indice AUC, GINI et R^2 McFadden des modèles CART

L'indice de Gini permet de mesurer le pouvoir discriminant des modèles. Les meilleurs coefficients sont obtenus pour l'arbre rose et l'arbre both. La courbe ROC représentant les vrais positifs en fonction des faux positifs permet de mieux visualiser l'écart de performances entre les différents arbres.

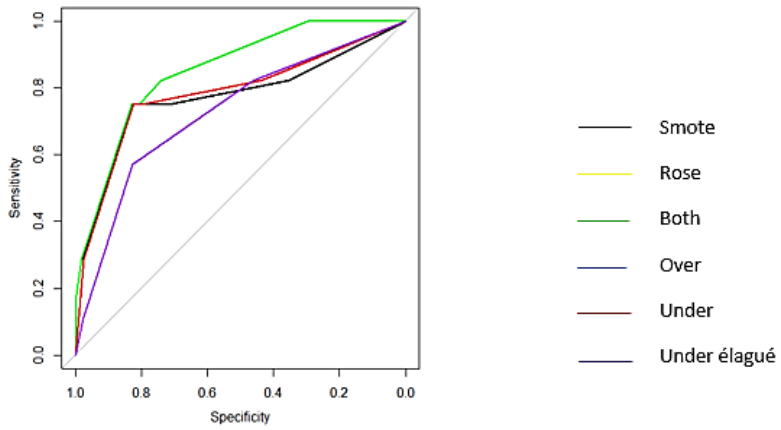


Figure 24 : Courbe ROC des arbres CART

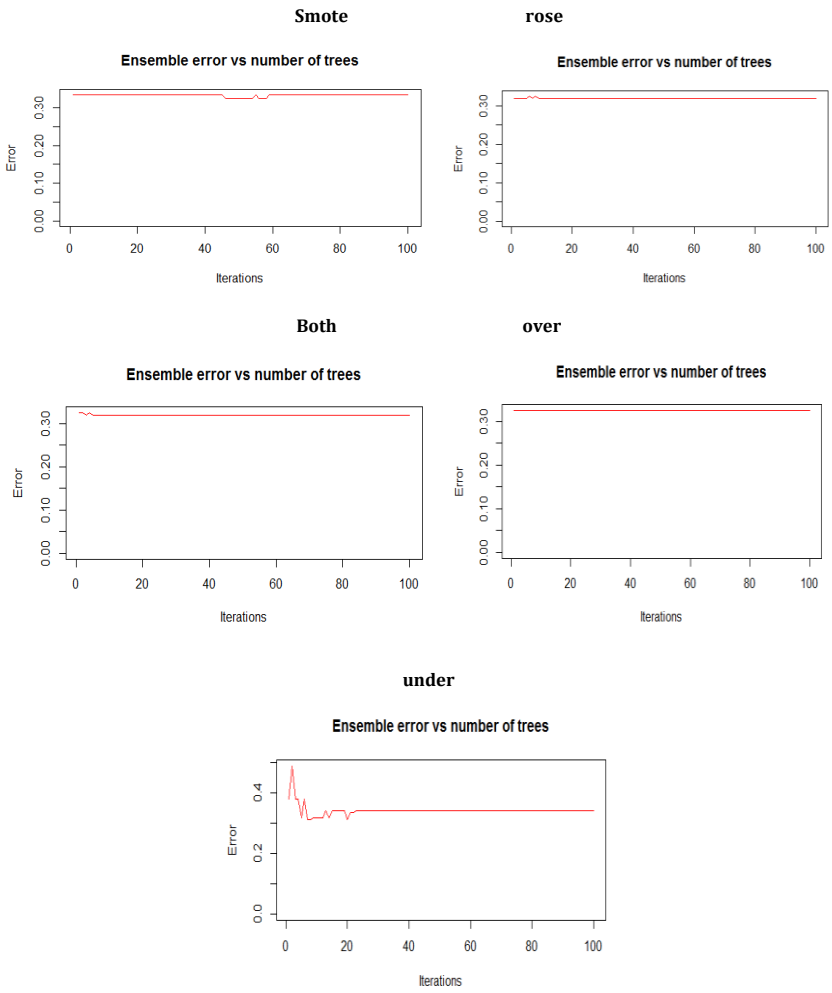
Plus la courbe est proche du point (0;1) et meilleure est la classification des données. On distingue clairement que les arbres *rose*, *both* et *over* ont un meilleur pouvoir discriminant. Dans la suite, les méthodes ensemblistes sont appliquées aux arbres et les modèles obtenus sont étudiés.

2.2 Les méthodes d'agrégations

Afin de définir le nombre d'itérations optimal nous utilisons l'erreur OOB. Nous considérons que si l'erreur OOB est stable à partir d'un certain nombre d'itérations alors nous pouvons retenir ce nombre d'itérations pour les modèles. Ainsi, en réduisant le nombre d'itérations, nous réduisons le temps des calculs. Pour construire les modèles de forêts aléatoires, nous avons utilisé le package randomForest de R et pour les modèles de *bagging* et de *boosting* nous avons utilisé le package Adabag.

- Bagging

Comme présenté plus haut, le bagging permet par un processus de multiplication des arbres de diminuer la variance des modèles. Nous choisissons de construire 100 arbres pour le *bagging*. Comme pour les arbres, le *bagging* est appliqué à toutes les bases d'apprentissage rééchantillonnées puis les modèles construits sont testés sur la base test. Au cours de la phase d'apprentissage, l'erreur out-of-bag est calculée en fonction du nombre d'itérations et ceci pour chacun des modèles.



<u>Figure 25</u> : Erreur OOB des modèles de bagging

A partir de ces graphiques, on détermine le nombre d'arbres à partir duquel l'erreur OOB est stable. Le choix du nombre d'itérations est réalisé de façon à réduire le temps mis par les programmes mais à avoir suffisamment d'arbres pour améliorer les performances des arbres. Nous retenons 80 itérations.

- <u>Les forêts aléatoires</u>

Pour les forêts aléatoires, nous choisissons de construire 1000 arbres. Afin de vérifier la stabilité de l'erreur au bout de 1000 itérations, l'erreur out-of-bag est calculée en fonction du nombre d'itérations et ceci pour chacun des modèles, au cours de la phase d'apprentissage. Nous retenons 800 arbres.

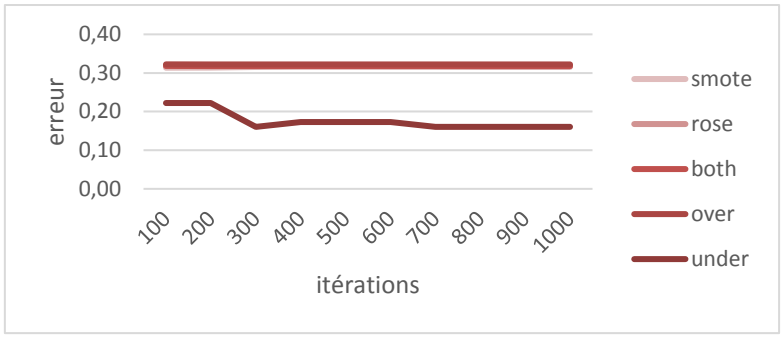


Figure 26 : Erreur OOB des modèles de forêts aléatoires

- Le boosting

Nous avons implémenté le *boosting* probabiliste ADABOOST sur la base d'apprentissage. L'erreur out-of-bag est calculée en fonction du nombre d'itérations et ceci pour chacun des modèles, au cours de la phase d'apprentissage. L'erreur OOB par itérations est représentée sur les graphiques suivants :

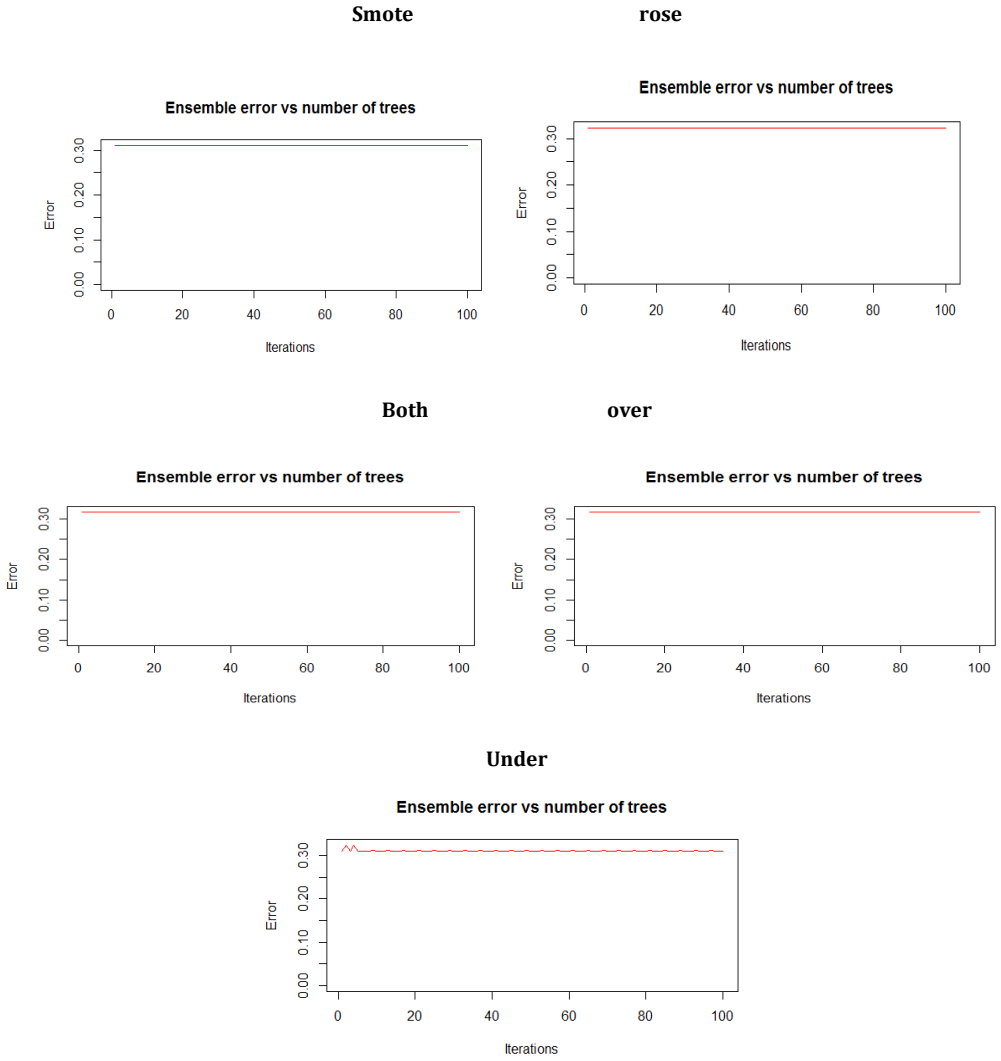


Figure 27 : Erreur OOB des modèles de boosting

Pour le *boosting*, nous choisissons de construire 50 arbres car cet algorithme est coûteux en temps. Le choix du nombre d'itérations est réalisé de façon à réduire le temps mis par les programmes tout en ayant suffisamment d'arbres pour améliorer les performances des modèles. Ces courbes montrent que le choix du nombre d'itérations est optimal puisque pour tous les modèles l'erreur OOB se stabilise au bout du nombre d'itérations choisis.

Pouvoir de discrimination des modèles

L'indice de Gini et l'AUC sont calculés pour tester le pouvoir discriminant des modèles :

Modèles	AUC	Indice de Gini
Bagging smote	0,7895	0,579
Bagging rose	0,8109	0,6218
Bagging both	0,8109	0,6218
Bagging over	07895	0,579
Bagging under	0,7162	0,4324
Forêt smote	0,7691	0,5382
Forêt rose	0,749	0,498
Forêt both	0,7691	0,5382
Forêt over	0,7691	0,5382
Forêt under	0,7891	0,5782
Boosting smote	0,7847	0,5694
Boosting rose	0,8608	0,7216
Boosting both	0,8608	0,7216
Boosting over	0,8596	0,7192
Boosting under	0,7849	0,5698

Table 22: Indice AUC, GINI et R^2 McFadden des modèles ensemblistes

Contrairement à nos attentes, l'indice de Gini montre que les forêts sont moins performantes que les arbres. Le pouvoir discriminant des modèles de *bagging* sont similaires à ceux des arbres. Les résultats du *boosting* montrent une amélioration des performances par rapport aux arbres. Ces résultats sont confirmés par les courbes ROC suivantes :

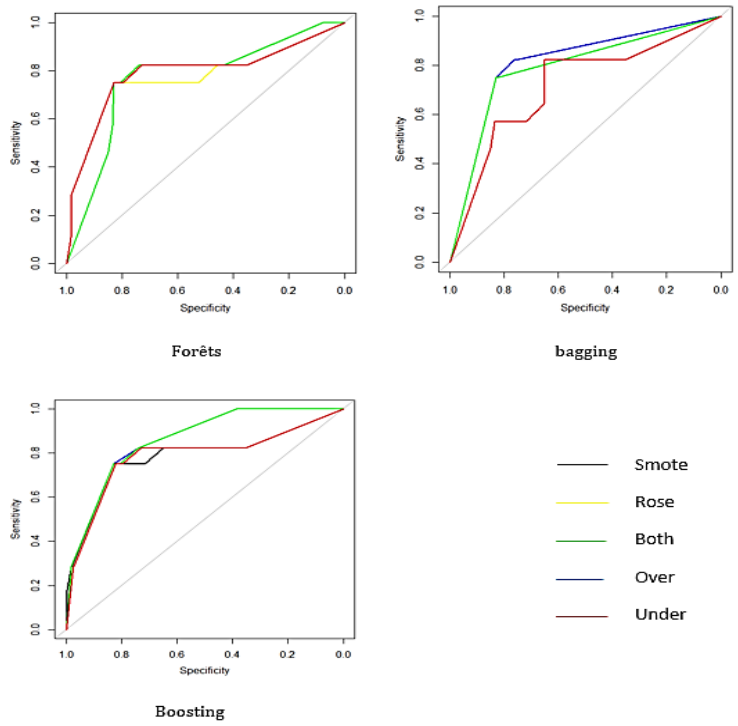


Figure 28: Courbes ROC des modèles d'agrégation des arbres CART

2.3 Analyse des performances prédictives des modèles

Pour la suite de l'étude et pour valider les modèles et leurs performances, il est intéressant de les tester sur la base test n'ayant pas servie à l'apprentissage et de regarder leurs performances en termes de prédiction. A partir de ces matrices de confusions nous calculons le taux d'erreur, la précision, le rappel, la F-mesure, le G-means, le MCC

- Pour les arbres nous obtenons les résultats suivants :

Modèles	Erreur	Rappel	Spécificité	Précision	G-means	F-measure	MCC
arbre smote	0,1787	0,7500	0,8214	*0,0085*	*0,7849*	*0,0169*	*0,0673*
arbre rose	0,2608	*0,8214*	0,7390	0,0064	0,7791	0,0127	0,0576
arbre both	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
arbre over	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
arbre under	0,2022	0,7500	0,7979	0,0076	0,7736	0,0150	0,0616
arbre under élagué	*0,1738*	0,5714	*0,8267*	0,0067	0,6873	0,0133	0,0475

Table 23: Indicateurs de la qualité de prédiction des arbres CART

L'arbre possédant la meilleure précision est le modèle « arbre Smote ». Rappelons que la précision est la part de sinistres bien prédits par le modèle parmi ses prédictions de sinistres. Ce taux doit être le plus important lorsqu'on veut réduire l'erreur liée aux faux positifs.

Le rappel mesure la capacité du modèle à bien prédire le sinistre et lorsqu'on recherche un modèle qui prédit bien le sinistre malgré le nombre d'erreur potentielle c'est l'indicateur à prioriser. Le modèle « arbre Under » a une bonne spécificité mais manque de précision. Ceci est probablement dû à la technique de rééchantillonnage qui réduit les données de la base initiale. Quant aux modèles « arbre rose », « arbre both » et « arbre over », ils ont un très bon pouvoir discriminant mais les performances prédictives sont plus faibles.

- le bagging, les forêts aléatoires et le boosting

Comme pour les arbres, les modèles d'agrégation sont appliqués à la base test puis les matrices de confusions sont calculées ainsi que les indicateurs de performance cités plus haut. Les valeurs obtenues pour les indicateurs permettent d'évaluer la performance prédictive des modèles.

Modèles	Erreur	Rappel	Spécificité	Précision	G-means	F-measure	MCC
bagging smote	*0,1693*	0,7500	*0,8308*	*0,0090*	0,7894	*0,0178*	*0,0698*
bagging rose	0,2374	*0,8214*	0,7625	0,0070	*0,7914*	0,0140	0,0619
bagging both	0,2374	0,8214	0,7625	0,0070	0,7914	0,0140	0,0619
bagging over	0,1693	0,7500	0,8308	0,0090	0,7894	*0,0178*	*0,0698*
bagging under	0,1995	0,5714	0,8009	0,0059	0,6765	0,0116	0,0421
forêt smote	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
forêt rose	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
forêt both	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
forêt over	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
forêt under	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
boosting smote	0,1787	0,7500	0,8214	0,0085	0,7849	0,0169	0,0673
boosting rose	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
boosting both	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
boosting over	0,2608	0,8214	0,7390	0,0064	0,7791	0,0127	0,0576
boosting under	0,2022	0,7500	0,7979	0,0076	0,7736	0,0150	0,0616

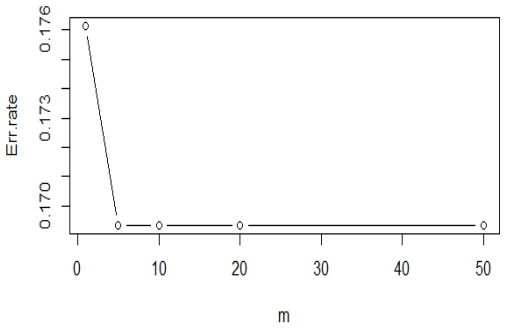
Table 24: Indicateurs de la qualité de prédiction des modèles ensemblistes

Ici le choix du meilleur modèle est délicat car les indicateurs désignent des modèles différents. Les modèles de *boosting* ont de très bonnes performances d'ajustement mais en prédiction les résultats ne sont pas satisfaisants. Nous avons retenu *le bagging smote* et le *bagging over* car ils présentent les meilleurs indice F-mesure et MCC et offrent un bon compromis entre performance d'ajustement et performance prédictive.

2.4 Stabilité des modèles

Pour tester la stabilité de l'erreur en test sur les 2 modèles retenus, nous construisons un *bagging* ayant pour nombre d'itération 1, 5, 10, 20 puis 50. Pour chacune de ces itérations le *bagging* est réalisé 10 fois. Afin de voir si les résultats sont stables l'erreur en test moyenne est calculée à l'issue des itérations.

Bagging over



Bagging Smote

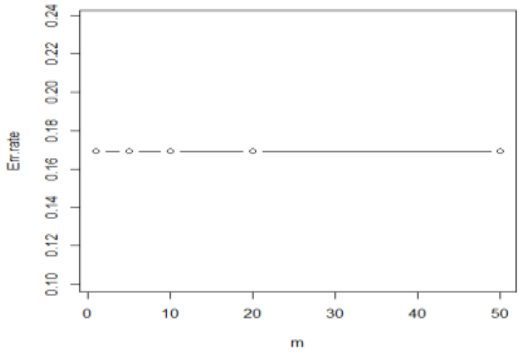


Figure 29 : Stabilité en erreur en test des modèles bagging smote et bagging over

Pour les deux modèles, l'erreur se stabilise très vite. L'erreur du bagging over, se stabilise à partir d'un nombre d'itération du *bagging* égal à 5. L'erreur de *bagging smote* se stabilise dès le départ à une erreur égale à 0,17. Cette erreur bien que stable est réalisée à chaque itération du *bagging* en utilisant la même base d'apprentissage et de test. Nous souhaitons voir l'évolution de l'erreur lorsque la base d'apprentissage et de test varient. Pour cela nous réalisons une validation croisée.

Validation croisée :

L'idée de la validation croisée est de considérer la base de données, de la séparer en 10 groupes. Dans le cas de notre base de données présentant un déséquilibre des classes, les 10 groupes construits sont stratifiés selon la variable de sinistralité de façon à ce que cette variable soit représentée de la même façon dans tous les groupes. Pour k allant de 1 à 10 on apprend sur l'ensemble des groupes hormis le groupe k. le modèle appris est testé sur le groupe k et on calcule l'erreur de prédiction. L'erreur de validation croisée est alors la moyenne des erreurs en test obtenues sur les k groupes.

Pour k=10 nous obtenons les erreurs de validation croisée suivantes :

modèles	**Erreur de validation croisée**
Bagging over	0,3359
Bagging Smote	0,1469

Table 25: Erreur de validation croisée des modèles bagging smote et bagging over

La plus faible erreur de validation croisée est obtenue pour le modèle *bagging smote*. Nous retenons donc ce modèle.

En définitive nous retenons deux modèles : le modèle *rose pondéré* pour l'approche logistique et le *bagging smote* pour l'approche par apprentissage.

PARTIE 3 : CALCUL DU SCR

L'objectif du calcul d'un SCR en modèle interne est de permettre à la compagnie de mieux modéliser ses risques. Cette démarche sera d'autant plus pertinente si elle permet d'obtenir un niveau de capital de solvabilité requis plus faible qu'en formule standard. Sur le marché des ADBAI certains travaux de modélisation des risques de souscription avaient déjà été réalisés par l'équipe modèle interne en vue de la mise en place d'un modèle interne partiel. Nous avons conservés certaines techniques de modélisation en l'occurrence celle du risque de réserves car le modèle de provisionnement proposé nous semble pertinent. Aussi avons-nous fait le choix de nous focaliser sur la modélisation des autres risques de souscription et de proposer en nous appuyant sur les études réalisés en partie 2, un modèle interne partiel pouvant challenger la formule standard. La première étape de notre démarche de modélisation a consisté en une étude du fonctionnement du marché des ADBAI. En tenant compte des particularités du marché, certains risques de souscription ne feront pas l'objet d'une modélisation. Dans une seconde étape, les facteurs de risques ont été identifiés sur les périmètres de risques modélisés. Dans une troisième étape, une modélisation des facteurs de risques est proposée. Enfin le SCR est calculé et les résultats sont comparés à ceux de la formule standard.

Chapitre 1: présentation du modèle interne partiel

1. Préliminaires

1.1 Les particularités du cycle d'assurance sur le marché des ADBAI

Un modèle adéquat, doit prendre en compte les particularités du marché. Pour ce faire nous nous sommes renseignés auprès des équipes métiers sur les points suivants :

- **La constitution du portefeuille d'assurance**

Nous avons tenté de comprendre la manière dont s'effectue le renouvellement du portefeuille d'assurance sur ce marché à savoir la formation du new business et la sortie des contrats de garanties. Le portefeuille d'assurance au 31/12 de l'année, est obtenu à partir des contrats ayant fait l'objet d'une campagne de renouvellement (et qui n'ont pas souhaité résilier leur contrat) appelés tacite reconduction et des nouvelles souscriptions dont on a connaissance. Au cours de l'année, l'effectif des nouveaux entrants peut augmenter, car les ADBAI ont la possibilité de souscrire tout au long de l'année. Aussi certains des dossiers ayant fait l'objet d'une campagne de renouvellement et dont on a supposé le renouvellement tacite, peuvent être résiliés sans paiement de prime. A la fin de la période de garantie, on peut également avoir des cas de résiliation du contrat. Elle peut être à l'initiative de l'ADBAI ou de la compagnie.

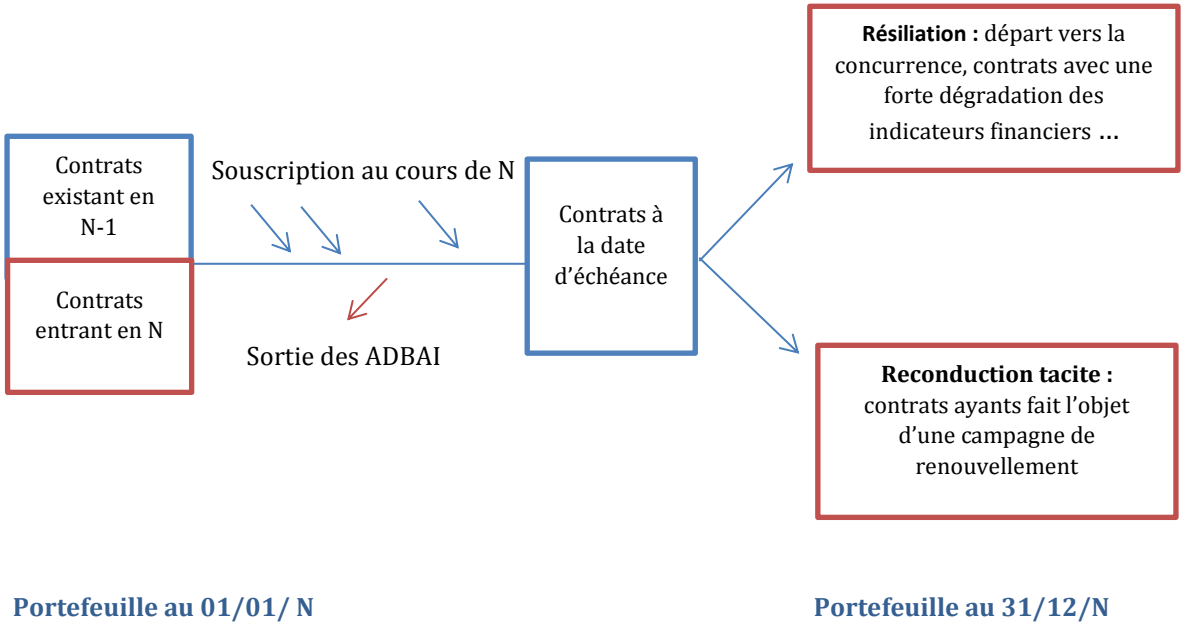


Figure 30: Schéma récapitulatif de la vie du portefeuille

- **La sinistralité**

Ensuite, nous nous sommes intéressés à la sinistralité. Le but étant de comprendre dans quels cas de figure la garantie est engagée et comment se déroule le sinistre. Il existe deux types de sinistres. D'une part, les sinistres liés à une non représentation des fonds mandants et donc faisant directement appel à la garantie financière nommés « vrais sinistres » et d'autre part les sinistres qui font suite à des réclamations qui ne touchent pas directement la garantie financière nommés « faux sinistres » ou « contentieux ». Il est nécessaire de faire cette distinction car les montants de sinistre dépendent du type de sinistre. Sur les contentieux, les charges de sinistres sont plus faibles que sur les vrais sinistres. Les échanges ont permis de confirmer certains éléments qui ressortent de l'étude statistiques de la base de données. Notamment le fait que les ADB sont plus sujets aux sinistres que les AI ou le fait que l'ancienneté et l'âge de l'entreprise n'ont pas d'impact sur la sinistralité.

- **Les frais d'activité**

L'activité sur le marché des ADBAI, génèrent certains frais. On distingue 5 principaux frais :

Frais d'acquisition : ce sont les frais générés par l'acquisition des contrats souscrits la première année de projection c'est-à-dire les tacites reconductions et les affaires nouvelles.

Frais d'administration : ce sont des frais liés à la gestion des contrats sains.

Frais de gestion : il s'agit de frais de gestion des sinistres survenus.

Frais financiers : ce sont des charges financières liées au placement d'actif en représentation des engagements de l'entreprise.

Autres charges techniques : ce sont des frais liés à la gestion courante qui ne sont pas pris en compte dans les précédents modules tels que les frais indirects de fonctionnement du marché (charge des fonctions supports, des outils informatiques, ...)

- **Le risque de rachat**

Nous nous sommes aussi intéressés au rachat du contrat. Le rachat peut être défini comme la possibilité pour l'ADBAI ayant souscrit à l'une des garanties proposées, de mettre fin de façon anticipée c'est-à-dire avant le 31/12 au contrat. Sur le marché des ADBAI, le rachat total ou partiel du contrat n'est pas possible. En effet, les contrats sont souscrits pour une durée annuelle. Cette durée étant relativement courte il n'est pas possible pour l'ADBAI d'opter pour un rachat avant terme de la garantie. Cela implique qu'il n'y a pas de possibilité de remboursement de la prime acquise par la compagnie. Il n'y a donc pas de risque de rachat sur ce marché et en définitive il ne serait pas pertinent d'en faire un objet de modélisation.

- **L'impact de la récession économique**

Le risque de récession peut découler d'une situation de crise entrainant une baisse d'activité importante pour les agents immobiliers et les administrateurs de biens. Par exemple une crise immobilière entrainant la fermeture de plusieurs agences immobilières et l'arrêt de l'activité par plusieurs professionnels. Nous avons cherché à connaître l'impact d'une crise immobilière sur l'activité de la compagnie. Il en ressort que l'activité des ADBAI est une activité assez stable dans le temps. Il y a une certaine récurrence et donc ce marché ne se retrouve pas affecté en cas de récession économique. Lors de la crise de 2008, ce sont les ADBAI ayant une activité de transaction et donc essentiellement les agents immobiliers qui ont été affectés puisque leur activité est directement impactée par le nombre de transactions immobilières. De plus, le portefeuille étant majoritairement constitué des administrateurs de biens, l'impact est d'autant plus atténué. En réalité, l'impact ne se fait sentir qu'à partir de la deuxième année. On peut constater ce retard sur le graphique 8 où l'impact de la crise de 2008 n'a été observé qu'à partir de 2009. On observe donc un effet retard d'un an sur le marché des ADBAI qui fait que l'impact de la récession si elle se produisait serait nulle sur le calcul du SCR.

- **La possibilité de recouvrement**

Nous nous sommes renseignés sur la possibilité pour la compagnie de recouvrer ses fonds en cas de sinistres. Le recours est la démarche que réalise la compagnie afin d'obtenir de l'ADBAI qu'il rembourse les frais déboursés en raison de la non représentation des fonds qui lui ont été confiés par ses mandants. D'après les dires des équipes métiers, c'est un processus coûteux et rare sur ce marché. Le remboursement n'a été observé que dans très peu de cas. Il ne semble pas nécessaire d'en tenir compte dans notre modélisation.

Ces échanges ont permis de prendre une décision quant à la modélisation ou non de certains risques de souscription. En nous basant sur les réponses apportées par les équipes métiers à nos questions, et après étude du fonctionnement du marché des ADBAI nous avons donc choisi de ne pas modéliser le risque de rachat et le risque catastrophe de récession. Dans ce mémoire, le SCR modèle interne pour ces périmètres de risque est nul.

Les risques de souscription qui feront l'objet d'une modélisation, sont le risque de primes, le risque de réserves et le risque catastrophe individuel. La modélisation de ces risques nécessite d'identifier les facteurs donnant lieu à l'aléa.

1.2 Les facteurs de risques identifiés

Le risque de prime

C'est un risque qui concerne les futurs sinistres sur les affaires nouvelles et les tacites reconductions. Il découle d'une potentielle inégalité entre le montant de primes acquises par la compagnie et celui des paiements générés par les sinistres futurs. En effet, l'étude du renouvellement du portefeuille en fin d'année montre qu'au 31/12/N, on a un aléa sur plusieurs éléments. Le nombre d'affaires nouvelles n'est pas connu et il peut évoluer puisque les ADBAI peuvent souscrire tout au long de l'année. De plus, le montant d'encours sur ces nouveaux ADBAI ne peut pas être déterminé avec certitude. Aussi certains des ADBAI reconduits, ne paieront au final pas leur prime. Ainsi, lorsqu'on se place au 31/12/N, il est impossible de prévoir avec certitude les sommes que la compagnie déboursera en règlement des sinistres qui surviendront au cours de l'année N+1. Par conséquent, on ne peut dire si le montant de primes acquises permettra de faire face à la sinistralité. Les facteurs de risques identifiés en lien avec les risques de primes sont énumérés ci-dessous. Ils concernent la survenance de sinistres futurs et les paiements qu'ils pourraient engendrer.

- **La probabilité de survenance :** Elle se définit comme la probabilité qu'un sinistre se produise au cours de l'année considérée. On cherche à estimer pour chaque ADBAI encore sain c'est à dire n'ayant pas de sinistre en cours, la probabilité qu'il connaisse un sinistre au cours de l'année étudiée. C'est le facteur de risque qui va induire l'existence d'un sinistre et donc déterminer la présence ou non de paiements liés aux sinistres.
- **Le montant d'encours :** Lors de la souscription, à chaque ADBAI correspond un montant d'encours donné. L'encours peut être défini comme le montant sur lequel la compagnie s'engage lors de la souscription du contrat. Les paiements de la compagnie liés à la garantie financière en cas de sinistre ne peuvent excéder le montant d'encours. C'est à partir de ce montant que sera déterminé le montant d'exposition de la compagnie
- **L'exposition au défaut :** L'exposition au défaut est en général un taux qui sera appliqué au montant d'encours afin d'anticiper le montant sur lequel la compagnie sera effectivement engagée en cas de sinistres futurs. En effet, lors de la survenance d'un sinistre avec mise en cause de la garantie financière, la compagnie

se retrouve exposée selon les cas sur tout ou partie du montant d'encours. L'exposition au défaut représente donc la part de montant d'encours sur laquelle la compagnie se retrouve exposée.

- **Montant et cadences des paiements :** Le montant des paiements correspond à la somme globale que la compagnie devra régler aux mandants des ADBAI en sinistre. Ce règlement peut s'étaler sur plusieurs années. On a donc une incertitude quant au déroulement des paiements des sinistres futurs à la fois sur leurs montants et sur leurs cadences.
- **Montant et cadences des frais :** Comme indiqué dans le chapitre 2, l'activité de la compagnie génère un certain nombre de frais. Dans le cadre du risque de prime et du risque cat individuel, les frais concernés sont : les frais d'acquisition, les frais d'administration, les frais de gestion, les frais financiers. Comme pour le montant des paiements, on observe un aléa sur le montant total de frais qui pourrait être généré par les sinistres futurs. Il en est de même pour leur cadence de règlement.

Le risque catastrophe individuel

Comme le risque de prime, il concerne les affaires nouvelles et les tacites reconductions. Il peut être défini comme le risque de prime sur des sinistres dits catastrophe. Il s'agit de sinistre ayant une forte sévérité. Les facteurs de risques sur ce périmètre sont :

- **Le nombre de sinistre et la charge de sinistre:** Ce sont des sinistres dont l'occurrence est faible et la charge importante, comparée aux sinistres dits attritionnel. Il y a un aléa quant au nombre et aux coûts des sinistres de type catastrophe individuel qui pourraient survenir au cours d'une année.

Dans les paragraphes suivants, nous proposons une méthode de modélisation des facteurs de risques identifiés, puis nous présentons aussi la méthodologie de calcul des BE des flux d'engagement et de calcul des SCR de primes, catastrophe individuel et réserves.

Le risque de réserves

Ce risque est déjà modélisé sur le marché des ADBAI. Nous reprenons les résultats du modèle actuel. Toutefois, rappelons que ce risque est lié aux sinistres survenus sur les exercices précédents et qui sont encore en cours. Ils sont appelés sinistres en stock. Il découle du fait que les montants mis en réserves lors des exercices précédents ne sont pas à la hauteur des règlements générés par ces sinistres. L'aléa ici vient du fait que lorsqu'un sinistre est déclaré, le processus de règlement est long et peut durer plusieurs années surtout lorsqu'il nécessite des procédures judiciaires ou l'intervention d'un expert. Les facteurs de risques sont donc les montants et cadences de frais générés par les sinistres en stock et non clos.

2. Modélisation des facteurs de risques identifiés

2.1 Le risque de primes

Le portefeuille d'assurance

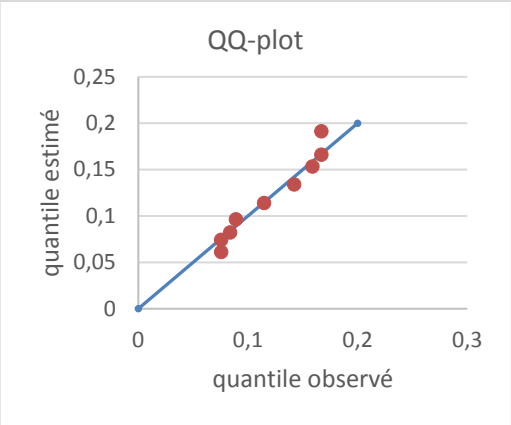
En central, le portefeuille d'assurance est composé de l'ensemble des contrats sur lesquels la compagnie s'engage pour l'année d'étude. Le portefeuille d'assurance est constitué de deux types de contrats. Il s'agit des affaires nouvelles (ADBAI qui souscrivent pour la première fois à la garantie) et les tacites reconductions (ADBAI déjà présents dans le portefeuille d'assurance l'année précédente).

- Afin de modéliser les ADBAI en tacite reconduction, nous faisons l'hypothèse que l'ensemble des contrats ayant fait l'objet au 31 décembre d'une campagne de renouvellement pour l'année suivante, fera l'objet d'un renouvellement. Cette hypothèse a été faite car sur le marché des ADBAI, l'activité est assez stable. Il s'en suit que les montants de primes attendues et de garanties octroyées pour ces dossiers sont déterministes.

- Pour la modélisation des affaires nouvelles, nous faisons l'hypothèse que les caractéristiques des nouveaux entrants sont identiques d'une année sur l'autre. Les nouveaux entrants dans le portefeuille d'assurance sont déterminés par un tirage aléatoire avec remise au sein de la base constituée des affaires nouvelles de l'année précédente. Le nombre d'admissions pour l'année d'étude est déterminé à partir d'un taux d'entrée moyen calibré sur un historique remontant à 1999 appliqué à la base des tacites reconductions.

- Le montant d'encours total lié aux affaires nouvelles est obtenu en faisant la différence entre le business plan de la compagnie et le montant de garantie global des dossiers qui seront renouvelés. Ce montant de garantie est ventilé par dossier à partir du poids des dossiers entrants de l'année précédente. Le montant des primes est obtenu en appliquant un taux de tarification moyen (basé sur l'historique) au montant d'encours obtenu pour chaque nouvel entrant.

- Les nouveaux dossiers sont sélectionnés aléatoirement dans la base des affaires nouvelles construite en central. Le nombre de dossiers entrants est obtenu en appliquant un taux d'entrée au nombre d'ADBAI de la base des tacites reconductions simulé à partir de loi calibrée sur l'historique. Pour déterminer la loi la mieux adéquate pour modéliser le taux d'entrée, plusieurs lois ont été testées notamment la loi lognormale, la loi weibull et la loi gamma. Nous avons choisi ces lois car elles sont à support positifs et pourraient donc convenir aux données de taux d'entrée. Nous retenons la loi gamma qui s'ajuste mieux aux données comme on peut le voir sur le graphique QQ-plot suivant :



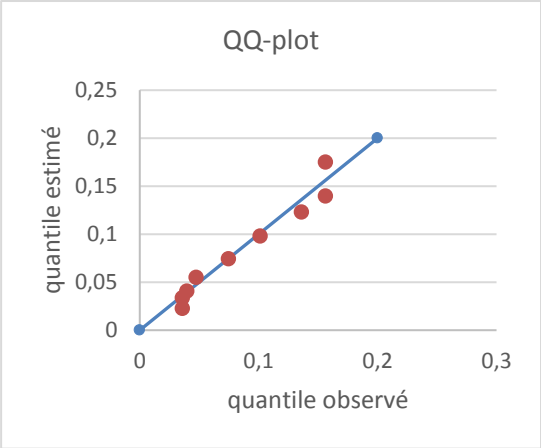
Paramètre estimés	
Sigma	0,00675
Alpha	17,27031

Figure 31: Ajustement de la loi gamma au taux d'entrée

Les tests d'ajustement du Khi-deux et de Kolmogorov Smirnov confirment ce résultat puisque nous obtenons des p_value supérieures au seuil 0,5%.

Test d'ajustement pour la distribution Gamma					
Test	Statistique		DDL	P-value	
Kolmogorov-Smirnov	D	0,16730		Pr > D	0,241
Chi-Square	Chi-Sq	5,83457	2	Pr > Chi-Sq	0,054

Les tacites reconductions sont composées de tous les contrats ayants fait l'objet d'une campagne de renouvellement en fin d'année N. A ces contrats renouvelés, on retranche un certain nombre de dossiers déterminé en appliquant un taux de sortie. Le taux de sortie est obtenu par une simulation de la loi calibrée sur l'historique des sorties de l'année précédente. Concrètement, le but ici est de retrancher les contrats dont on a supposé le renouvellement mais qui ne paieront pas leurs primes. Comme pour le taux d'entrée, nous avons testé quelques lois continues à support positifs comme la gamma, la weibull, la log-normale. La loi la mieux ajustée est la gamma dont l'adéquation est confirmée par le QQ-plot et les résultats des tests d'ajustement :



Paramètres estimés	
Sigma	0,01375
Alpha	5,76042

Figure 32 : Ajustement de la loi gamma au taux de sortie

Pour les tests d'ajustement, les p_value obtenues indiquent une adéquation de la loi gamma aux données :

Test d'ajustement pour la distribution Gamma					
Test	Statistique		DDL	P-value	
Kolmogorov-Smirnov	D	0,13316		Pr > D	>0,500
Chi-Square	Chi-Sq	4,17723	2	Pr > Chi-Sq	0,124

La probabilité de survenance du sinistre

Pour modéliser la probabilité de survenance du sinistre, plusieurs modèles ont été testés et présentés dans les chapitres précédents. Au regard des différents résultats obtenus, le choix s'est porté sur le *rose pondéré* pour les modèles testés dans l'approche logistique et sur le *bagging smote* pour les modèles testés en apprentissage. Nous utiliserons ces deux modèles pour le calcul du SCR. Ces modèles permettent d'obtenir pour chaque ADBAI, la probabilité que ce dernier connaisse un sinistre au cours de l'année d'étude. La survenance du sinistre est alors modélisée en simulant une loi de Bernoulli ayant pour paramètre, la probabilité construite par le modèle de survenance ici le *rose pondéré* ou le *bagging smote*.

L'exposition au défaut et la charge de sinistre

En cas de survenance d'un sinistre, l'intensité dépend de facteurs tels que le montant d'encours de l'ADBAI et la catégorisation du sinistre survenu. En effet, les paiements de la compagnie liés à la garantie financière en cas de sinistre sont déterminés par le montant d'encours. La charge dépend aussi du type de sinistre. Nous avons tenus compte de la distinction faite entre vrais sinistres et faux sinistres. En cas de vrai sinistre, la garantie est engagée et la charge de sinistre est plus élevée qu'en cas de faux sinistre où seuls des frais judiciaires et de gestion sont engagés. Pour cela sur les sinistres modélisés à l'étape précédente, le type de sinistre est modélisé en utilisant les réalisations d'une loi de Bernoulli ayant pour probabilité le taux moyen de vrai sinistre. Une fois les sinistres catégorisés, le montant d'exposition de la compagnie peut alors être déterminé.

L'exposition au défaut est en général un taux qui est appliqué au montant d'encours afin d'anticiper le montant sur lequel la compagnie sera engagée en cas de sinistres futurs. En effet, lors de la survenance d'un sinistre avec mise en cause de la garantie financière, la compagnie se retrouve exposée selon le cas sur tout ou partie du montant d'encours. L'exposition au défaut représente donc la part de montant d'encours sur laquelle la compagnie se retrouve exposée.

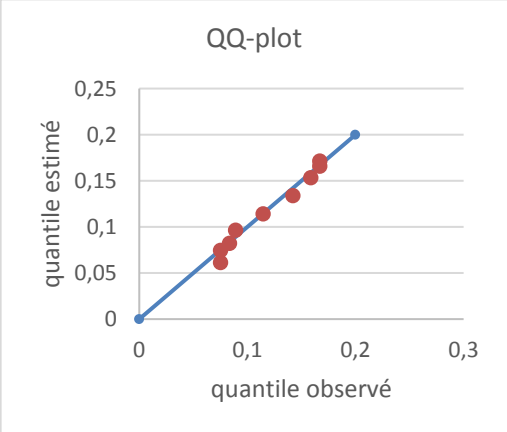
En cas de faux sinistre c'est-à-dire d'un litige, l'exposition au défaut est le montant forfaitaire de frais décaissés pour solutionner ce litige. Ce montant forfaitaire a été déterminé en se basant sur l'historique des coûts moyens unitaires des faux sinistres.

En cas de vrai sinistre, le montant d'exposition dépend du montant d'encours de l'ADBAI. A partir de l'historique des vrais sinistres et des dires d'experts, deux seuils d'encours s_1 et s_2 ont été défini tels que $S_1 < S_2$.

Ces seuils ont été retenus car les taux d'exposition varient selon que le montant d'encours considéré est en dessous ou dessus de ces seuils. Il n'était donc pas adéquat d'appliquer un taux moyen d'exposition. L'algorithme de modélisation retenu est le suivant :

- pour un sinistre associé à un montant d'exposition inférieur au seuil s_1, une simulation du taux d'exposition est réalisée à partir de la loi calibrée sur cette variable sur un historique remontant à 1999.
- pour un sinistre associé à un montant d'exposition compris entre s_1 et s_2, un taux d'exposition r_1 est appliqué.
- pour un sinistre associé à un montant d'exposition supérieur à s_2, un taux d'exposition r_2 est appliqué.

Pour modéliser le taux d'exposition nous avons testés plusieurs lois statistiques continues et à support positif. Ici aussi nous retenons la loi gamma car elle s'ajuste très bien aux données de taux d'exposition comme on peut le voir sur le QQ-plot :



Paramètres estimés	
Sigma	0,575222
Alpha	0,653015

Figure 33 : Ajustement de la loi gamma au taux d'exposition

Les tests d'ajustement confirment ce résultat :

Test d'ajustement pour la distribution Gamma					
Test	Statistique		DDL	P-value	
Kolmogorov-Smirnov	D	0,16074955		Pr > D	>0,150
Chi-Square	Chi-Sq	3,76366296	2	Pr > Chi-Sq	0,152

Le montant de paiement pour chaque sinistre est alors obtenu en appliquant le taux d'exposition au montant d'encours de l'ADBAI dont le sinistre est modélisé. Seuls les sinistres modélisés de charges inférieures à 0,5M€ sont considérés dans le cadre du risque de prime. Le choix de ce seuil est expliqué dans le paragraphe sur le risque catastrophe individuel.

Les frais d'activité

En plus des paiements en cas de sinistre, l'entreprise doit faire face à certains frais liés à son activité. Dans le cadre du risque de prime et du risque cat individuel, les frais concernés sont : les frais d'acquisition, les frais d'administration, les frais de gestion, les frais financiers. Comme pour le montant des paiements, on observe un aléa en ce qui concerne le montant total de frais qui pourrait être générés. Les frais sont modélisés selon la méthodologie suivante :

- **Frais d'acquisition** : ce sont les frais générés par l'acquisition des contrats souscrits la première année de projection c'est-à-dire les TR et les AN. Au global, Ils sont modélisés à partir d'un coût unitaire moyen observé en N-1 auquel on applique l'inflation et du nombre de contrats souscrits la première année de projection.

$$Frais\ d'acquisition_N = Coût\ unitaire\ inflaté_{N-1} \times Nombre\ de\ contrats\ souscrits_N$$

- **Frais d'administration** : ce sont des frais liés à la gestion des contrats n'ayant pas de sinistre en cours déclaré. Ils sont calculés sur les TR et AN. Comme pour les frais d'acquisition, le principe consiste à projeter le coût unitaire moyen observé en N-1 auquel on applique l'inflation. Ce coût unitaire inflaté est alors multiplié par le nombre de dossiers pour obtenir le montant de frais d'administration global.

$$Frais\ d'administration\ des\ contrats_N = Coût\ unitaire\ inflaté_{N-1} \times Nombre\ de\ contrats\ sains_N$$

- **Frais de gestion** : il s'agit de frais de gestion des sinistres survenus ou futurs. A partir d'une moyenne des frais observés sur les trois dernières années, un montant forfaitaire est défini. A ce montant, on applique l'inflation. Ce montant sera ventilé entre frais de gestion des sinistres en stock et frais de gestion des futurs sinistres selon la proportion de sinistres en stock et futurs sinistres.

$$Frais\ de\ gestion_{Primes} = Frais\ de\ gestion\ total \times \frac{N\_sin_N^{futur}}{N\_sin_N^{stock} + N\_sin_{N+1}^{futur}}$$

$$Frais\ de\ gestion_{réserve} = Frais\ de\ gestion\ total\ \times \frac{N\_sin_N^{stock}}{N\_sin_N^{stock} + N\_sin_{N+1}^{futur}}$$

- **Frais financiers :** ce sont des charges financières liées au placement d'actif en représentation des engagements de l'entreprise. Ils sont calculés à partir d'une assiette et d'un taux de frais financiers. Ce taux de frais de gestion financière est basé sur le montant des frais financiers et la valeur de marché du portefeuille observé. Cette méthode de calcul est la même pour les risques de réserve, de primes, et catastrophe individuel. Pour le risque de prime, l'assiette est définie à partir des flux de paiements, de frais, de primes.
- **Autres charges techniques (ACT):** ce sont des frais liés à la gestion courante et qui ne sont pas pris en compte dans les précédents modules de frais tels que les frais indirects de fonctionnement du marché (charge des fonctions supports, des outils informatiques, ..). ce montant ne dépend pas du volume de production. Il est modélisé par un coût forfaitaire observé sur l'année N-1 auquel on applique un facteur d'inflation pour le projeter en N. Seule la moitié de ce montant est prise en compte dans le calcul du SCR prime l'autre moitié étant prise en compte dans la modélisation du risque de réserve.

$$Frais\_ACT\_prime = {}^{1}/_{2} \times Frais\_ACT\_global$$

$$Frais\_ACT\_réserve = {}^{1}/_{2} \times Frais\_ACT\_global$$

Pour ces frais lorsqu'il est question de projeter un coût observé en N-1, la projection se fait sur une demi-année car on fait l'hypothèse que les frais sont réglés en milieu d'année. La courbe des taux sans risque utilisée est celle fournie par l'EIOPA.

Cadences de paiements de sinistres et frais

Les frais d'acquisition, les frais d'administration et les autres charges techniques n'impactent que la première année de projection et sont donc réglés en première année de projection. En ce qui concerne les frais financiers, les frais de gestions de sinistres et les paiements de sinistres, les règlements ont lieu tout au long des années de projection selon la cadence de paiement déterminée par le modèle de provisionnement.

2.2 Le risque catastrophe individuel

Il s'agit d'un risque atypique lié aux sinistres à forte intensité c'est-à-dire entrainant une importante charge de sinistralité. De façon générale, la notion de « sinistre atypique» est associée à un seuil. Tout sinistre dont le montant serait supérieur à ce seuil est alors considéré comme atypique. Le choix du seuil doit être effectué de façon à disposer de suffisamment de données de sinistres pour construire le modèle de risque catastrophe individuel. Généralement, ce risque est modélisé en se basant sur l'historique de sinistralité par une approche de type : $coût\ moyen\ *\ fréquence\ moyenne$

Dans le cadre de notre modélisation, nous avons choisi cette approche. Le seuil de risque atypique a été défini en se basant sur la théorie des valeurs extrêmes.

Choix du seuil

Dans la théorie des valeurs extrêmes, pour étudier la distribution d'une variable aléatoire au-delà d'un certain seuil, une famille de lois continues a été introduite. Il s'agit de la loi de Pareto généralisée. Considérons une variable aléatoire X qui admet une fonction de répartition F. Nous voulons estimer la fonction de répartition F_u de la variable X au-delà du seuil $u < x_F$.

On a :

$$F_u(y) = P(X - u \leq y | X > u) \; avec \; 0 \leq y \leq x_F - u$$

$$F_u(y) = \frac{P(X - u \leq y, X > u)}{P(X > u)} = \frac{P(u < X \leq y + u)}{P(X > u)}$$

$$F_u(y) = \frac{F(u + y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}$$

Pikands démontre que pour une large classe de distribution F, la fonction de répartition des excès au-dessus d'un certain seuil F_u peut être approchée par une loi de Pareto généralisée $G_{\varepsilon,\beta}^{p}$définie selon l'expression suivante :

$$G_{\varepsilon,\beta}^{p} = \begin{cases} \left(1 - \left(1 + \frac{\varepsilon y}{\beta}\right)^{-\frac{1}{\varepsilon}}\right) * 1_{x>0} & si \; \varepsilon > 0 \\ \left(1 - \left(1 + \frac{\varepsilon y}{\beta}\right)^{-\frac{1}{\varepsilon}}\right) * 1_{x \in [0, -\frac{\beta}{\varepsilon}]} & si \; \varepsilon > 0 \\ \left(1 - e^{-\frac{x}{\beta}}\right) & si \; \varepsilon = 0 \end{cases}$$

Un bon choix du seuil u est donc nécessaire à l'application des valeurs extrêmes. L'une des méthodes classiques utilisée est l'étude du graphique de la fonction de dépassement moyen. Elle est définie par l'expression suivante :

$$e(u) = E(X - u | X > u)$$

En considérant un échantillon de taille n, dont les observations ordonnées pour la variable aléatoire X sont $X_i, i = 1, \dots, n$ son expression empirique est la suivante :

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u) * 1_{X_i > u}}{\sum_{i=1}^{n} 1_{X_i > u}}$$

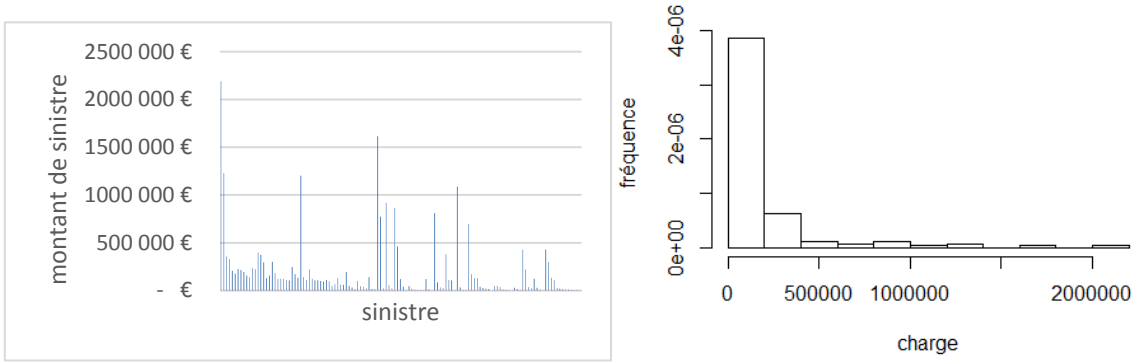
En notant $X_{(1)} \geq \cdots \geq X_{(n)}$la statistique d'ordre associée à notre échantillon, le graphique de dépassement moyen est obtenu par le tracé des points$\{X_{(i)}, e_n(X_{(i)})\}, i = 1, \dots, n$

Si l'ajustement des dépassements par une loi GPD (β, ε) est valide avec un seuil u, alors la fonction de dépassement moyen au-delà de u est :

$$E(X-u|X>u)=\frac{\beta+\varepsilon*u}{1-\varepsilon}$$

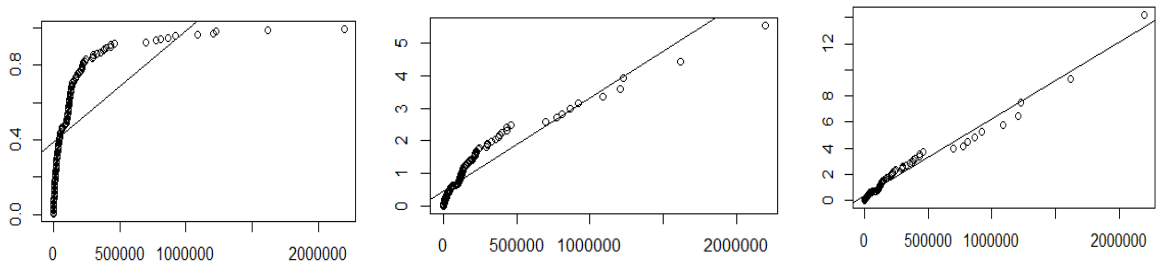
Cette fonction est donc linéaire en u avec une pente $\frac{\varepsilon}{1-\varepsilon}$. Le seuil à choisir est alors défini comme le seuil minimal pour lequel la propriété de stabilité est satisfaite.

Dans la suite nous appliquons cette technique à nos données pour déterminer le seuil u. Nous considérons les sinistres survenus sur le marché des ADBAI sur l'historique disponible.



<u>Figure 34 :</u> Distribution des montants de sinistres sur le marché des ADBAI

Pour examiner cette distribution, et voir si elle peut être associée à une distribution extrême, nous faisons le graphique des quantiles QQ-plot. Ce graphique représente les points $\{X_{(i)}, F^{-1}(1-i/n)\}, i=1,\dots,n$ et permet de comparer les distributions empiriques et théoriques. Nous comparons la distribution empirique des données avec différentes lois de Pareto généralisé.



$\varepsilon=-1$ $(loi\ uniforme)$ $\varepsilon=0$ $(loi\ exponentielle)$ $\varepsilon=0{,}3$

<u>Figure 35:</u> QQ-plot de la distribution empirique des donnés VS loi de Pareto généralisée

Le dernier QQ-plot indique une adéquation des données avec la loi de Pareto généralisée $\varepsilon=0{,}3$ car la courbe obtenue est à peu près linéaire. Pour déterminer le seuil u, nous recherchons sur le graphique de la fonction de dépassement moyen, le plus petit seuil à partir duquel la courbe semble stable.

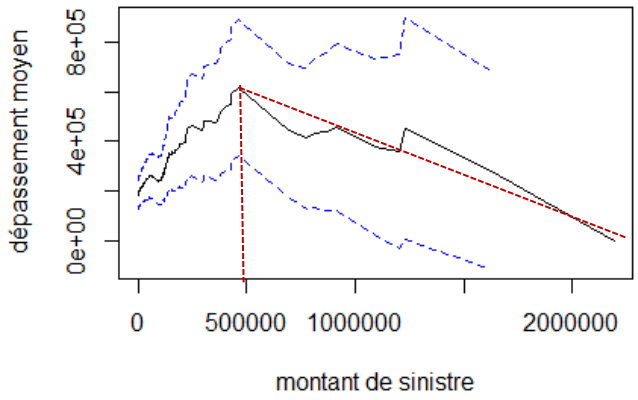


Figure 36: La fonction de dépassement moyen

Au vu de ce graphique, nous choisissons de retenir pour notre modélisation un seuil égal à 0,5 M€. Dans l'historique de sinistralité, les sinistres dont la charge est supérieure au seuil atypique sont classés en sinistre de type catastrophe individuel. En se basant sur les sinistres de montant supérieur au seuil atypique, la démarche de modélisation consiste à trouver les lois qui s'ajustent bien avec d'une part le nombre de sinistre et d'autre part les montants de sinistres. D'abord nous considérons les distributions empiriques des observations.

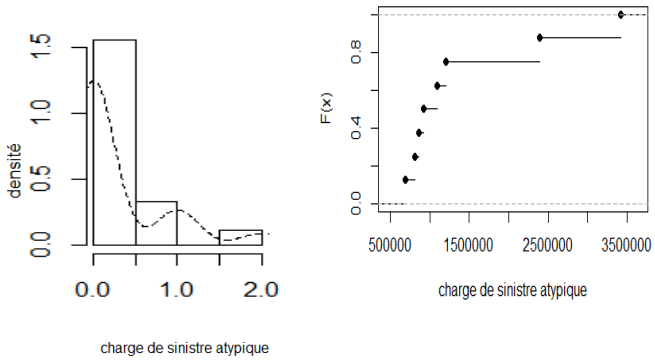


Figure 37: Distribution empirique de la charge de sinistre atypique

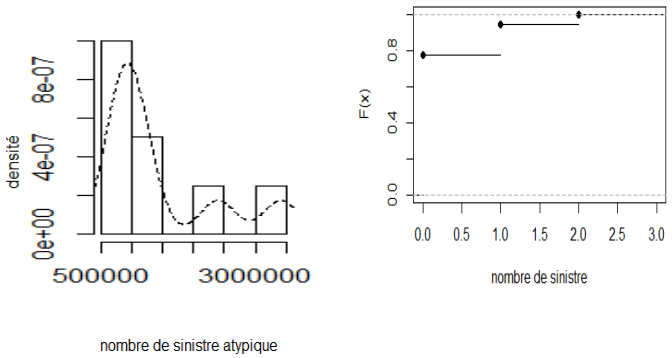


Figure 38 : Distribution empirique du nombre de sinistre atypique

Modélisation du nombre de sinistres de type catastrophe individuel

Le nombre de sinistres étant une donnée de comptage, nous avons testé la loi de poisson et la loi binomiale négative. Les paramètres des lois sont estimés par la méthode du maximum de vraisemblance et les résultats obtenus sont :

Loi	Paramètres	Valeurs estimées
Poisson	λ	0,28
Binomiale négative	(n ; p)	(1,68 ; 0,28)

Table 26: Paramètres estimés des lois poissons et binomiale négative

Pour tester l'ajustement nous avons comparé les distributions théorique et empirique pour la loi de binomiale négative et la loi de poisson. Les graphiques obtenus sont les suivants :

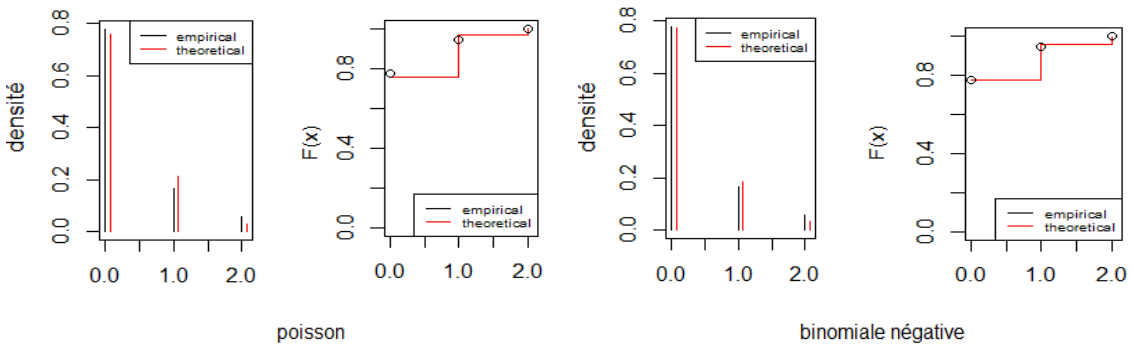


Figure 39: Ajustement des lois poisson et binomiale négative au nombre de sinistres atypiques

Les résultats sont assez similaires mais la loi binomiale négative semble mieux s'ajuster aux données. Le test d'adéquation du khi-deux confirme l'adéquation des lois aux données car la p-value obtenue pour les deux distributions est 0,000499. La loi retenue pour la modélisation du nombre de sinistre est la binomiale négative.

Modélisation de la charge de sinistres de type catastrophe individuel

Pour modéliser les montants de sinistre, nous avons testé quelques lois usuelles continues et à support positif à savoir la loi log normale, la loi gamma et la burr. Les paramètres des lois sont estimés par la méthode du maximum de vraisemblance et les résultats obtenus sont :

Loi	Paramètres	Valeurs estimées
Lognormale	$(\mu; \sigma)$	(14 ,01108; 0,1382185)
Gamma	$(\alpha; \beta)$	(3,330377 ; 2,339927e-06)
Burr	(a; b ; s)	(1,295601 ; 1,624433 ; 7,012675e-07)

Table 27: Paramètres estimés des lois lognormale, burr et gamma

Pour tester l'ajustement des lois aux données, nous considérons le graphique suivant :

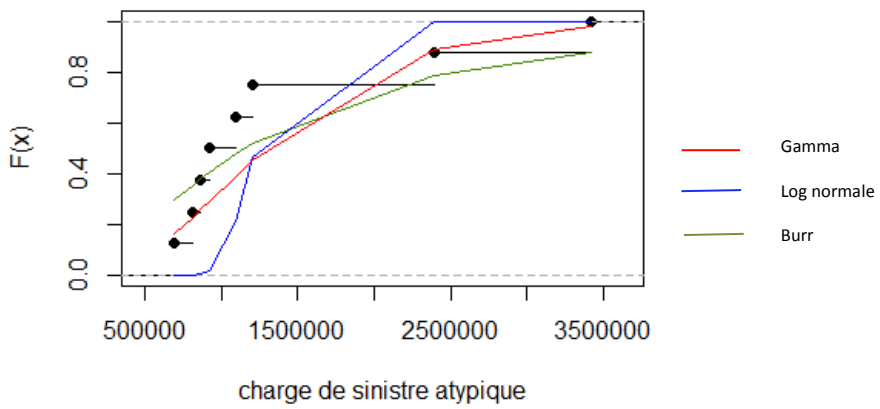


Figure 40: Ajustement des lois gamma, burr et log normale à la charge de sinistres atypiques

La loi gamma en rouge semble la plus adaptée. Pour confirmer ce résultat nous utilisons les tests du Khi-deux et de Kolmogorov Smirnov dont les résultats des p_value sont résumés ci-dessous :

Test	log normale	gamma	burr
khi-deux	0 ,995	1	1
Kolmogorov-smirnov	0,05207	0,4894	0,4863

Nous retenons donc la loi gamma pour modéliser les montants de sinistre.

En plus des montants de charge de sinistres, sur le périmètre du risque de catastrophe individuel, l'entreprise doit tenir compte des frais financiers. Comme pour le risque de primes, le calcul des frais financiers s'effectue via l'application d'un taux de frais financiers sur une assiette de flux moyens estimée en amont.

2.3 Le risque de réserves

Le risque de réserve est lié à l'incertitude des montants et cadences de paiement des sinistres en stock. Le risque est de sous-estimer le montant de provisions nécessaire pour couvrir ces sinistres. Dans le modèle construit par l'équipe Modèle Interne, la charge ultime et les cadences de paiement sont modélisées via un modèle de Chain-Ladder et un modèle de Mack. Notons que trois principaux frais viennent s'ajouter aux montants d'engagements de la compagnie sur le périmètre du risque de réserve à savoir les frais de gestion des sinistres en stock, les frais financiers associés aux sinistres en stock et les autres charges techniques. La modélisation de ces frais est celle décrite dans le risque de prime. On fait l'hypothèse que ces frais sont réglés dans le temps de la même façon que les montants de paiement. Ce qui implique que pour obtenir les flux de frais à actualisés, les cadences de règlement issue du modèle de provisionnement sont appliquées. Le modèle de provisionnement est présenté en annexe du mémoire.

3. Méthodologie de calcul des SCR de primes, catastrophe individuelle et de réserves

Pour le calcul du SCR, nous faisons les hypothèses suivantes :

- Il n'y a pas de réassurance sur ce marché
- Le nombre d'année de projection est égale au nombre d'année de développement
- Les cadences de paiement des flux liés au risque de primes et au risque catastrophe individuel sont identiques car dans les triangles de paiements dont nous disposons, il n'y a pas de distinction entre sinistres attritionnel et sinistre atypique.

Le calcul du SCR est basé sur une projection à un an de la situation nette de l'entité. On introduit une fonction de perte L qui se calcule selon l'expression suivante :

$$L = SN(0) - D(1) * SN(1)$$

Cette fonction permet de mesurer la perte potentielle après introduction d'une année. Le SCR étant le niveau minimal de fonds propres permettant de faire face à une ruine à 1 an avec un risque de 0,5%, alors il vérifie l'expression suivante :

$$plus\ petit\ X\ tel\ que\ Proba(\ L > X) \leq 0{,}5\%$$

Alors le SCR est obtenu en appliquant la Value At Risk à la fonction de perte :

$$SCR = VaR_{0{,}5\%}[SN(0) - D(1) * SN(1)]$$

$$SCR = SN(0) - D(1) * \ VaR_{0.5\%}SN(1)$$

Où :

- *D(t)* correspond au facteur d'actualisation entre les dates 0 et t : il s'agit du prix en 0 d'un zéro coupon de maturité un an.
- *SN(t) = A(t)-BE(t),* correspond à la situation nette de l'entreprise en t; il s'agit des capitaux propres correspondant à la différence entre les valeurs économiques de l'actif *A(t)* et la meilleure estimation du passif *BE(t)* à la date t.

Puisque $SN(t) = A(t) - BE(t)$, alors la formule du SCR s'écrit aussi :

$$SCR = A(0) - BE(0) - D(1) * \ VaR_{0.5\%}(A(1) - BE(1))$$

Le calcul du SCR s'est fait selon le processus suivant

- Construire un scénario en central permettant d'obtenir le portefeuille en entrée pour les tirages stochastiques
- Définir un certain nombre n de simulations

- A chaque simulation, simuler les facteurs de risques sur la base de tirage pour obtenir la distribution en t=1 de la situation nette de la compagnie
- Calculer la VaR en appliquant un quantile à 0,5% sur l'échantillon des n simulations de la situation nette en t=1

Le graphique ci-dessous résume la procédure de calcul.

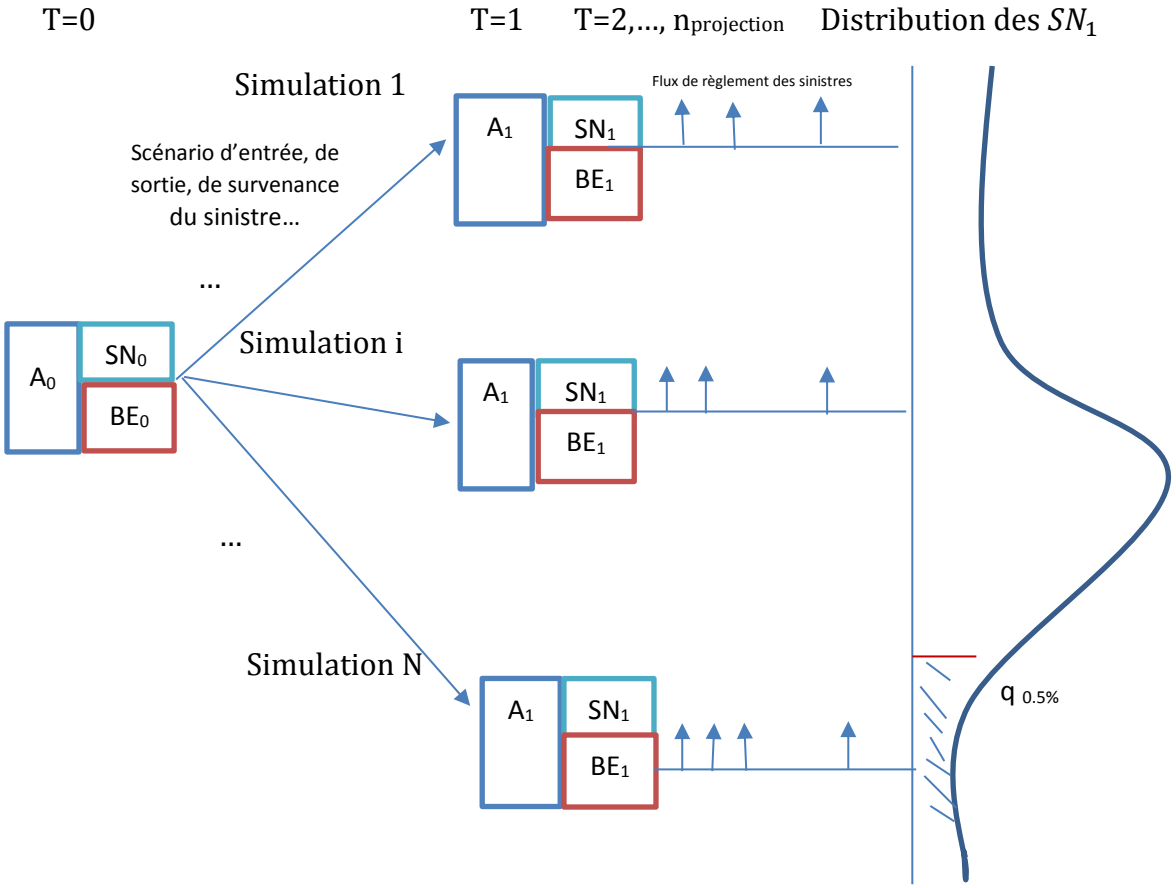


Figure 41: Schéma de simulation de la situation nette en 1

La méthodologie décrite ci-dessus pour le calcul du SCR nécessite de connaitre la valeur des BE et des actifs en t=0 et en t=1. Dans les paragraphes suivants nous décrirons la méthodologie d'évaluation des BE pour les SCR de primes, catastrophe individuel et de réserves.

La simulation des facteurs de risques présentés ci-dessus entre t=0 et t=1 permet de prendre en compte l'aléa relatif aux risques de souscription. La modélisation de ces facteurs aboutit pour chaque simulation considérée, à un montant de sinistralité, à un montant de frais de gestion et de frais financiers représentant l'engagement de l'entreprise. Ces montants sont écoulés selon les cadences de paiements du modèle de provisionnement. A ces flux de paiements, s'ajoutent les primes perçues par la compagnie en contrepartie de la garantie offerte et d'autres frais (acquisition et administration) versés uniquement en première année de projection. Les BE sont alors calculés en prenant la moyenne des simulations de flux futurs de paiements. Sous l'hypothèse d'absence de variabilité des actifs dans le calcul du risque de souscription, la différence de la valeur de

marché des actifs entre la date t=0 et la date t=1 est uniquement liée aux flux de trésorerie de l'année : primes reçues liées à la nouvelle production, aux sinistres et aux frais payés en 1ère année. Donc l'actif A(t) varie, entre t=0 et t=1, uniquement des flux de première période. Le calcul du SCR repose alors essentiellement sur l'évaluation des BE.

3.1 Evaluation des BE du risque de primes

Pour le risque de primes les flux modélisés sont les suivants :

- Les paiements relatifs aux sinistres modélisés entre t=0 et t=1 (hors sinistres de type catastrophe individuel)
- Les frais de gestion des sinistres de ces sinistres et les frais de placements financiers des actifs en représentation de ces engagements.
- Les primes perçues sur l'ensemble des contrats souscrits pour l'année
- Les frais d'administration et les frais d'acquisition

Les deux premiers sont versés de façon étalée sur le nombre d'année de développement des sinistres. Les deux derniers éléments sont versés ou perçus en une seule fois et concernent à la fois le risque lié aux sinistres de faible intensité et ceux de type catastrophe individuelle. Dans ce modèle, nous faisons le choix d'inclure ces éléments dans le périmètre du risque de prime. En appliquant le taux d'actualisation adéquat, les BE sont estimés par les formules ci-dessous :

$$\mathrm{BE}^{primes}(0) = -Primes + \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n_{projection}} (Flux_{paiement} * 1_{paiement<seuil} + Frais) * D(i)$$

$$\mathrm{BE}^{primes}(1) = -Primes + \frac{1}{n}\sum_{k=1}^{n}\sum_{i=2}^{n_{projection}} (Flux_{paiement} * 1_{paiement<seuil} + Frais) * D(i)$$

Le SCR est alors obtenu en appliquant la formule présentée ci-dessus

3.2 Evaluation des BE et SCR du risque catastrophe individuel

Les paiements relatifs aux sinistres modélisés entre t=0 et t=1 et dont le montant de charge dépasse le seuil atypique (sinistre catastrophe individuel). Pour calculer les montants de paiements, nous réalisons des simulations du nombre de sinistres et des montants de sinistres. Les montants de sinistre sont modélisés par une loi gamma et le nombre de sinistres par une binomiale négative calibrées sur l'historique. La charge de sinistre par simulation s'obtient en faisant le produit des coûts et du nombre de sinistres. Ce montant est projeté et actualisé en utilisant les cadences de paiement des sinistres attritionnels. Le BE de charges pour chaque simulation est obtenu avec la formule suivante :

$$\mathrm{BE}^{cat} = \sum_{i=1}^{n_{projection}} (paiement\ (i)) * D(i)$$

Le SCR est ensuite calculé en prenant le quantile à 99,5% de la distribution du BE de charges :

$$\mathrm{SCR}^{cat} = q_{99,5\%}(distribution\ des\ \mathrm{BE}^{cat})$$

3.3 Agrégation des risques

Nous obtenons, le SCR pour chacun des périmètres de risque de souscription. Les SCR obtenus doivent être agrégés de façon à tenir compte de la corrélation qui pourrait exister entre ces risques.

Les capitaux élémentaires du module risque de souscription sont agrégés, en tenant compte de la diversification existante au sein de ce module de risque. Pour calculer le SCR total de souscription nous agrégeons les risques primes, réserves et catastrophe individuel en utilisant la matrice de corrélation utilisée en formule standard. Une corrélation de 0.5 est appliquée entre risques de primes et réserves et de 0.25 avec le risque catastrophe.

Chapitre 2: Modèle interne partiel VS formule standard

1. Méthode de calcul des SCR souscription en formule standard

L'EIOPA dans le QIS[6] 5 propose le calcul du SCR, par une formule dite formule standard. Cette formule nécessite au préalable le calcul de différents SCR par sous module. Nous nous intéressons au SCR souscription non-vie qui est constitué des trois sous-modules suivant :

- Le sous module « risque de primes et de réserves non-vie »
- Le sous module « risque de catastrophe en non-vie »
- Le sous module « risque de rachat en non-vie »

La formule de calcul retenue est la suivante :

$$SCR_{non-vie} = \sqrt{\sum_{i,j} Corr(i,j) * SCR_i * SCR_j}$$

Où : $Corr(i,j)$est le coefficient de corrélation entre les sous modules i et j ;

SCR_i et SCR_j sont respectivement le capital de solvabilité requis pour les sous modules i et j.

Les coefficients de corrélation correspondant aux différents couples de sous-modules sont indiqués dans le tableau suivant :

i \ j	Primes et réserves	Catastrophe	rachat
Primes et réserves	1	0.25	0
Catastrophe	0.25	1	0
Rachat	0	0	1

Table 28: Corrélation des risques de souscription en formule standard

Conformément aux risques de souscription cités plus haut, les SCR à calculer par sous-modules de risques seraient : le $SCR_{prim-res}$, le SCR_{cat}, et le SCR_{rachat}.

Module risque de primes et de réserves non-vie

Calcul du $SCR_{prim-res}$:

Le $SCR_{prim-res}$ se calcule en faisant le produit entre l'écart type et le volume du risque de primes et de réserves selon la formule suivante :

$SCR_{prim-res} = 3 * \sigma * V$,

- σ *est l'écart type du risque de primes et de réserves*

[6] Quantitative impact study

- V *est le volume pour le risque de primes et de réserves*

➢ **Le volume de prime et réserve se calcule selon la formule suivante :**

$V = (V_{prim} + V_{res}) * (0.75 + 0.25 * DIV_s)$,

- DIV_s *est le coefficient de diversification du segment* 6
- V_{prim} *et* V_{res} *respectivement les volumes de primes et de réserves*

Le volume de primes se calcule à partir des volumes de primes acquis l'année précédente et l'estimation des primes à acquérir l'année suivante.

$V_{prim} = \max(p_s , p_{last,s}) + FP_{existant,s} + FP_{future,s}$

- p_s: *estimation des primes* à *acquérir l'année suivante*
- $p_{last,s}$: *primes acquises l'année passée*
- $FP_{existant,s}$: *valeur actuelle probable des primes attendues aprè*s *l'année* à *venir pour les contrats existant*
- $FP_{future,s}$: *valeur actuelle probable des primes* à *acquérir pour les contrats futurs comptabilisés l'année* à *venir*

Le volume de réserves est la meilleure estimation des provisions pour les sinistres en stock.

$V_{res} = BE$ *de provisions pour sinistres* à *payer net de réassurance*

Les montants de provisions peuvent être calculés en utilisant la méthode de Chain-Ladder pondérée. A partir des projections effectuées sur les triangles de paiements cumulés, les montants de réserves sont calculés. Les provisions sont alors estimées en sommant les montants actualisés de réserve d'une année à l'autre.

$$BE\ provisions = \sum_{i=0}^{n_{dev}} \left(\sum_{j=i}^{n_{dev}} (\hat{C}_{i,n_{dev}-j+i} - \hat{C}_{i,n_{dev}-j+i-1}) + Frais \right) * D(i)$$

Où n_{dev} est le nombre d'années de développement et $\hat{C}_{i,j}$ l'estimation par la méthode de Chain-Ladder des paiements cumulés en année de développement j pour les sinistres survenus en i et $D(i)$ le facteur d'actualisation pour l'année i.

➢ **L'écart type de primes et réserves se calcule selon la formule suivante :**

$$\sigma = \frac{\sqrt{\sigma_{prim}^2 * V_{prim}^2 + \sigma_{prim}^2 * V_{prim} * \sigma_{res}^2 * V_{res} + \sigma_{res}^2 * V_{res}^2}}{(V_{prim} + V_{res})}$$

En intégrant les particularités de l'activité de cautionnement qui correspond au segment d'activité n°6 dans la réglementation et les spécificités des paramètres de la formule

standard pour cette branche (12% pour l'écart type du risque de prime et 19% pour l'écart type du risque de réserves), on obtient les formules suivantes :

$$\sigma = \frac{\sqrt{12\%^2 * V_{prim}^2 + 12\% * V_{prim} * 19\% * V_{res} + 19\%^2 * V_{res}^2}}{(V_{prim} + V_{res})}$$

Module risque de catastrophe en non-vie

Calcul du SCR_{cat} :

Il est constitué du $SCR_{cat\ récession}$ et du $SCR_{cat\ individuel}$ et se calcule selon la formule suivante :

$$SCR_{cat} = \sqrt{SCR_{cat\ récession}{}^2 + SCR_{cat\ individuel}{}^2}$$

$SCR_{cat\ récession}$ se calcule en prenant 100% du montant de primes acquises au cours de l'année d'étude brut de réassurance.

$SCR_{cat\ individuel}$ se calcule par la perte nette de réassurance liée aux défauts des 2 plus grosses expositions avec une hypothèse de perte de 10% de la somme assurée

Module risque de rachat en non-vie

Calcul du SCR_{rachat} : ce calcul dépend du modèle choisi par l'entreprise. Pour ce risque, la compagnie recherche les fonds propres en représentation d'un évènement soudain tels que la cessation de 40% de ses contrats ou la baisse de 40% du nombre de ces futurs contrats. Le risque de rachat dépend de l'activité de l'entreprise. Le rachat n'étant pas possible sur le marché des ADBAI, le capital de solvabilité requis pour ce risque est nul.

2. Les résultats

Suite à l'étude réalisé en partie 2, nous avions retenus deux modèles pour la survenance du sinistre à savoir le modèle *rose pondéré* et le *bagging smote*. Pour calculer le risque de prime, nous avons implémenté ces deux modèles. En utilisant les probabilités de sinistre construites par ces modèles, il est possible de simuler la réalisation du sinistre et donc de déterminer le nombre de sinistre futurs projetés.

Nous souhaitons nous assurer que les modèles construits sont en adéquation avec les données réelles. Pour cela nous avons calculé pour chaque modèle, le ratio moyenne/écart-type que nous comparons aux données réelles. Nous avons réalisé 1000 simulations au total. Les valeurs des ratios obtenus par modèle sont résumées dans le tableau suivant :

	Moyenne	Ecart-type	Moyenne/ Ecart-type
rose pondéré	7,302	2,838	2,573
bagging smote	387,740	45,232	8,572
Réel	6,529	3,220	2,028

Table 29: Ratio moyenne sur écart-type du nombre projeté de sinistres attritionnels

On remarque que le ratio obtenu pour le modèle *rose pondéré* est proche du ratio calculé sur les données réelles. Une surestimation du nombre de sinistre en fait de plus une approche prudente. Des ajustements marginaux pourront éventuellement être effectués dans le futur pour améliorer un peu plus l'adéquation de ce modèle aux données réelles. Ce modèle semble bien toutefois bien s'ajuster aux données disponibles dans notre historique. En revanche, le ratio du *bagging smote* est 4 fois plus grand que le ratio réel ce qui implique un mauvais ajustement du modèle aux données. Nous considérons à nouveau ce ratio pour comparer les nombre de vrais sinistres et de faux sinistres projetés aux données réelles. Les résultats obtenus sont récapitulés dans les tableaux suivants :

- Vrais sinistres modélisés

	Moyenne	Ecart-type	Moyenne/ Ecart-type
rose pondéré	5,550	2,388	2,324
bagging smote	298,359	35,985	8,291
Réel	4,941	2,531	1,952

Table 30: Ratio moyenne sur écart-type du nombre projeté de vrais sinistres attritionnels

- Faux sinistres

	Moyenne	Ecart-type	Moyenne/ Ecart-type
rose pondéré	1,752	1,320	1,327
bagging smote	89,385	13,288	6,727
Réel	1,588	1,458	1,090

Table 31: Ratio moyenne sur écart-type du nombre projeté de faux sinistres attritionnels

On remarque que les résultats du modèle *rose pondéré* sont plutôt stables. Les écarts de ratio par rapport aux données réelles ne sont pas importants. Pour le *bagging smote* en revanche, les écarts sont encore plus importants sur le nombre de faux sinistres. Plus de 6 fois le ratio réel.

Pour évaluer la modélisation retenue pour la charge de sinistre, nous considérons les montants de sinistralités simulés par les deux modèles.

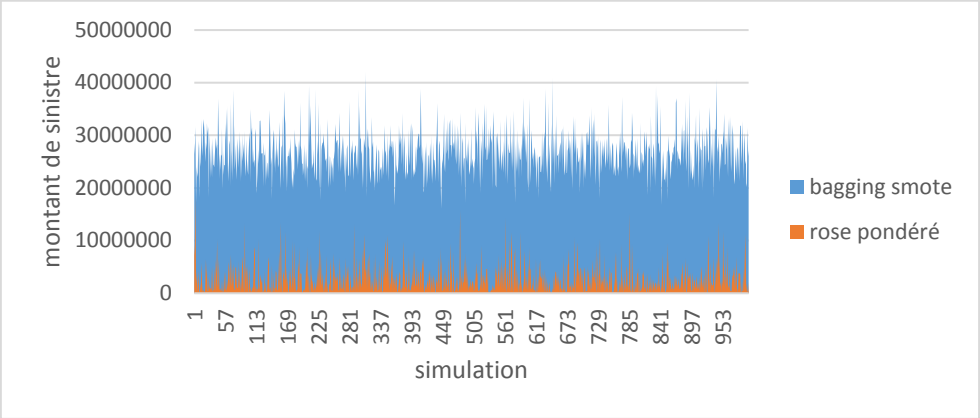


Figure 42: Distribution de la charge totale projetée de sinistres attritionnels

Comme on peut le voir sur ce graphique le modèle de *bagging smote* fait exploser la charge de sinistralité totale. Ce qui est une conséquence du nombre important de sinistres modélisés. Pour regarder au plus près, nous considérons à nouveau le ratio moyenne/ écart-type. Cette fois nous le calculons sur la distribution de charge projetée et nous la comparons aux données réelles du montant de sinistres. Les résultats obtenus sont récapitulés dans le tableau suivant :

	Moyenne	Ecart-type	Moyenne/ Ecart-type
rose pondéré	2613341,048	2662590,104	0,982
bagging smote	26791551,530	6162784,798	4,347
Réel	2673490,249	2142811,520	1,248

Table 32: Ratio moyenne sur écart-type du nombre projeté de faux sinistres attritionnels

On remarque que les résultats ne changent pas beaucoup pour le *bagging smote*, les écarts entre le ratio du modèle et le ratio réel sont conservés à peu près. A priori la modélisation retenue pour le taux d'exposition ne permet pas d'améliorer les performances au global de ce modèle. Si l'on veut utiliser ce modèle, il faudra affiner la modélisation

En définitive le modèle *rose pondéré* est celui qui s'ajuste le mieux aux données réelles et de ce fait est plus performant que le *bagging smote* lorsqu'on considère la survenance du sinistre. Néanmoins, on remarque que l'écart entre le ratio de charge totale modélisés et le ratio de charges réelles observées est négatifs ce qui pourrait laisser croire que le modèle *rose pondéré* sous-estime le montant de sinistre.

Considérons le graphique suivant :

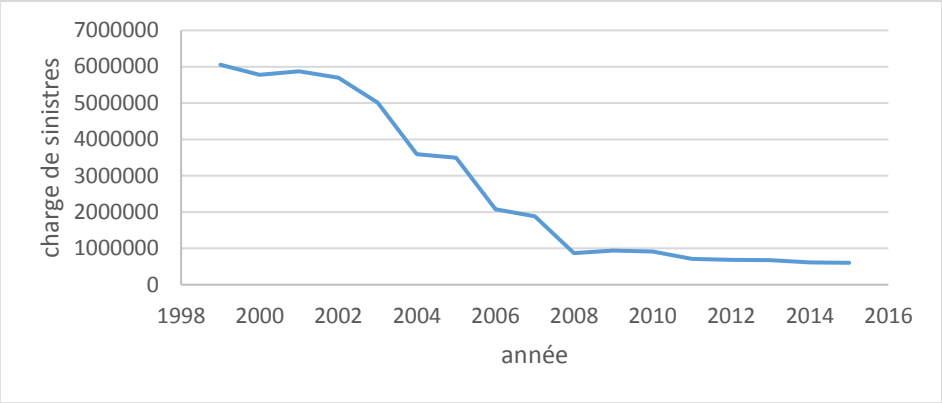


Figure 43: Evolution des montants de sinistralité annuelle

On remarque que la charge totale de sinistre a considérablement baissé ces dernières années. A partir de 2008, le montant moyen de la charge de sinistre est inférieur à 1 M€. On pourrait donc considérer qu'il n'y pas de sous-estimation du montant annuel de sinistralité par le modèle *rose pondéré* et que le résultat précédent est dû aux importantes sinistralités observées au début de l'historique. Toutefois ce modèle de prime pourra être davantage affiné pour arriver à une meilleure prédiction de la sinistralité.

Nous avons effectué les mêmes calculs pour le modèle du risque catastrophe individuel. Pour ce modèle, la comparaison réel et simulé des ratios moyenne/écart-type permet de vérifier que le modèle construit n'est pas en décalage avec les données réelles et ne sous-estime pas la survenance ou la charge lié à ce risque eu égard à l'historique de sinistralité. Le ratio moyenne/écart-type est calculé pour le nombre et la charge de sinistres simulés :

	Moyenne	Ecart-type	Moyenne/ Ecart-type
Simulé	4,373	3,980	1,099
Réel	0,278	0,558	0,497

Table 33: Ratio moyenne sur écart-type du nombre de sinistres catastrophe individuel simulés

	Moyenne	Ecart-type	Moyenne/ Ecart-type
Simulé	1424254,082	775564,202	1,836
Réel	331156,316	610936,886	0,542

Table 34: Ratio moyenne sur écart-type de la charge totale de sinistres catastrophe individuel simulée

Les ratios simulés sont plus grands que les ratios observés. Compte tenu de cela, on peut dire que le modèle construit pour le risque catastrophe individuel est plutôt en adéquation avec la réalité avec une légère surestimation de la charge totale de sinistres

du type catastrophe individuel, ce qui en soit ne constitue pas un problème puisque cela suppose que le modèle est prudent.

Afin d'estimer l'impact sur le SCR, nous avons appliqué les deux modèles et comparer les résultats obtenus aux résultats en formule standard. La formule standard sur le marché des ADBAI conduit aux résultats suivants :

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
6,82 M€	2,99 M€	119,23 M€	14,83 M€	-	120,98 M€

La mise en œuvre du modèle interne avec le modèle *rose pondéré* pour le calcul des probabilités de survenance aboutit aux résultats suivants :

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
11,22M€	0,87 M€	40,56 M€	-	-	44,96 M€
65%	-71%	-66%	-100%	-	-63%

La dernière ligne représente les variations des montants de SCR modèle interne par rapport au SCR en formule standard. Sur le SCR primes, on remarque une hausse de 65%. Sur ce périmètre, le modèle interne ne permet pas à la compagnie de réaliser des gains en termes de capital par rapport à la formule standard. En revanche, sur le risque de réserves et catastrophe individuel, l'utilisation du modèle interne permet à la compagnie de réduire considérablement le niveau de solvabilité requis. On observe respectivement des baisses de 71% et de 66% sur le SCR réserve et le SCR catastrophe individuel. Au global en mettant en place ce modèle interne partiel pour le SCR souscription non vie, la compagnie réaliserait un gain de 76 M€.

La mise en œuvre du modèle interne avec le modèle *bagging Smote* pour le calcul des probabilités de survenance aboutit aux résultats suivants :

SCR primes	SCR réserve	SCR cat individuel	SCR cat récession	SCR rachat	SCR souscription non vie
25,45 M€	0,87 M€	40,56 M€	-	-	53,30 M€
273%	-71%	-66%	-100%	-	-56%

La seule différence par rapport au modèle précédent se trouve au niveau du calcul du risque de prime. Le montant de SCR prime requis par ce modèle est 4 fois plus élevé que le montant requis en formule standard et 2 fois plus que le modèle précédent *rose pondéré.*

Cette perte est compensée par le gain réalisé sur le SCR catastrophe individuel ce qui fait qu'au global la mise en place de ce modèle permettrait à la compagnie de réaliser un gain de 68 M€ par rapport à la formule standard. En définitive, la compagnie gagnerait à mettre en place un modèle interne partiel pour le calcul du SCR de souscription.

Notons que toute chose égale par ailleurs, une modélisation de la survenance de sinistre par le modèle de *bagging smote* se trouve être plus coûteuse en terme de SCR pour la compagnie qu'une modélisation par le modèle *rose pondéré*. Pourtant lors de l'étude réalisée ce modèle présentait de très bonne qualité d'ajustement et les meilleures performances prédictives.

Ce résultat s'explique par le fait que les probabilités construites par le modèle de *bagging Smote* sont plus élevés que les probabilités obtenues par le modèle *rose pondéré.* En effet, lors de la simulation des sinistres, le modèle *bagging smote* génère un nombre de sinistre très important en raison des probabilités élevées. Nous pensons que tous les modèles construits en apprentissage autour de l'arbre CART auraient abouti à des résultats similaires car les probabilités générées sont élevées.

Lorsque l'objectif est la prédiction d'une observation, les modèles présentés en apprentissage sont performants mais lorsque le but recherché est de construire des probabilités les modèles en régression logistique sont plus adaptés car par construction l'arbre de décision cherche à répartir des observations en plusieurs groupes alors que le but de la régression logistique est d'estimer la probabilité pour un individu d'appartenir à une classe donnée.

Conclusion

Dans certaines bases de données, on observe des variables binaires présentant une asymétrie des classes. C'est le cas de la variable de sinistralité sur le marché des ADBAI. Ces dernières années, un certain nombre de mathématiciens se sont penchés sur la question et ont prouvé par leurs travaux que les méthodes telles que la régression logistique ou les arbres de décisions sont perturbés par ce problème d'asymétrie. Dans ce mémoire, nous avons pris en compte cette particularité et présenté quelques techniques de modélisations qui permettent de la traiter. Nous nous sommes intéressés aux techniques en régression logistique et en apprentissage à l'arbre de décisions CART. En régression logistique, la solution adoptée consiste à modifier la base pour retrouver un équilibre entre les classes puis à appliquer à la base équilibrée, les méthodes introduites par KING et ZENG à savoir l'ajustement préalable et la pondération des observations. En apprentissage, les méthodes utilisées consistent à rééquilibrer la base, et /ou à appliquer les techniques ensemblistes pour rendre l'arbre CART plus performant.

Pour l'application de ces méthodes nous avons séparé la base de données en deux bases : l'une pour l'apprentissage du modèle et l'autre pour les tests. La séparation a été faite de façon à conserver les mêmes proportions de classes que dans la base initiale. Ensuite, nous avons rééquilibré la base en utilisant 5 algorithmes de rééchantillonnage à savoir : *undersampling, oversampling, both,* SMOTE, ROSE. Pour chacune des nouvelles bases obtenues, nous avons appliqué en régression logistique, les méthodes d'ajustement préalable et les méthodes de pondération. En apprentissage, nous avons appliqué l'arbre CART, les méthodes d'agrégation telles que le *bagging*, les forêts aléatoires et le *boosting*.

Sur la base de test, nous avons évalué les qualités d'ajustement et de prédiction des modèles construits. Pour comparer ces modèles, nous avons présenté et utilisé un certain nombre d'indicateurs adaptés en cas d'asymétrie des classes. Les indicateurs n'évoluant pas toujours dans le même sens, il n'a pas été aisé de définir un modèle optimal. En effet, les modèles présentant les meilleures performances en termes de prédiction ne possèdent pas les meilleures qualités d'ajustement. Nous avons observé le meilleur ajustement pour les modèles *boosting* both et le *boosting rose* tandis que les meilleures prédictions sont pour les modèles *bagging smote* et *bagging over*. En définitive, les meilleures performances sont observées sur les modèles construits en apprentissage avec le classificateur CART.

Nous souhaitions un modèle qui offre un bon compromis entre ajustement et prédiction, nous avons donc retenu deux modèles : le *rose pondéré* en régression logistique et le *bagging smote* en apprentissage. Pour vérifier la robustesse du *rose pondéré* nous avons considéré les estimateurs et leurs variances pour chaque année supplémentaire d'historique. Pour le *bagging smote*, nous souhaitions voir si les résultats obtenus sont stables et nous avons donc calculé l'erreur de validation croisée.

Forts de ces modèles qui prennent en compte les particularités des données, nous avons proposé un modèle interne partiel pour le calcul du SCR de souscription. Le risque de

prime a été modélisé deux fois car nous avons considéré les deux modèles construits pour la survenance du sinistre attritionnel. Le SCR de prime obtenu pour le *bagging smote* est quasiment le double du SCR prime obtenu pour le *Rose pondéré*. Le modèle de prime basé sur le *bagging smote* est encore plus coûteux pour la compagnie. Cela est dû au fait que les probabilités de sinistre construites par les modèles en apprentissage sont plus élevées. Sur ce périmètre de risque la modélisation interne ne permet pas à la compagnie de réaliser des gains par rapport à la formule standard mais lui offre néanmoins une meilleure appréhension de ces risques. En revanche, sur le risque catastrophe individuel et le risque de réserve on observe un gain important par rapport à la formule standard. Ceci qui fait qu'au global le SCR souscription en modèle interne est inférieur au SCR souscription en formule standard. La compagnie aurait donc tout à gagner en mettant en place un modèle interne que ce soit en termes de connaissance de ses risques ou en termes de coût en capital.

Pour traiter le déséquilibre des classes en apprentissage l'une des méthodes consiste à introduire une matrice de coût de mauvaise classification dans l'algorithme de classification (Cost Sensitive Learning). Dans le cadre de ce mémoire nous n'avons pas considéré cette approche. Il serait donc intéressant d'appliquer cette méthode et d'en étudier les performances.

En ce qui concerne le modèle interne partiel présenté, deux aspects peuvent faire l'objet de travaux complémentaires. Pour le risque catastrophe récession, il serait intéressant de faire des études complémentaires pour décider si sa modélisation est pertinente sur ce marché. Aussi nous avions choisis de corréler les risques en utilisant la matrice de corrélation de la formule standard. Introduire la théorie des copules pour mesurer la dépendance entre les risques permettrait une meilleure appréciation de la corrélation entre les risques de souscription étudiés.

Table des illustrations

Bibliographie

[1] BEE W. et *al.* [2014] «An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets»

[2] BELLINA R. [2014] « Méthodes d'apprentissage appliquée à la tarification non vie »

[3] CHAWLA V. et *al*. [2002], «SMOTE: Synthetic Minority Over-sampling Technique»

[4] DECUPERE S., [2011] « Agrégation des risques et allocation de capital sous Solvabilité II »

[5] HAIBO H., AND GARCIA E. [2009] «Learning from Imbalanced Data»

[6] KING G., ZENG L., [2001] «Logistic Regression in Rare Events Data»

[7] LAILY R. [2014] « Construction d'un Modèle Interne Partiel en Assurance non-vie »

[8] LOPEZ V. et *al.* [2013] « An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics»

[9] LUNARDON N. et *al.* [2014] «ROSE: A Package for Binary Imbalanced Learning»

[10] WEISS G. [2004] «Mining with Rarity: A Unifying Framework»

[11] WILLIAMS R. [2016] «Analysing Rare Events with Logistic Regression»

[12] YUN-CHUNG LIU A. [2004] «The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets »

[13] ZHE LI [2010] « Conversion modeling in direct motor insurance and study of some related rare events Issues »

[14] SCHISTERMAN F. [2005] «Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples »

[15] SMITH T. & MCKENNA C. [2013] «A Comparison of Logistic Regression Pseudo R 2 Indices»

Annexe

Les tests statistiques

Dans ce paragraphe nous présentons les tests statistiques utilisés pour valider nos hypothèses.

Le test d'indépendance du Khi-deux

Le test d'indépendance du khi-deux est un test statistique introduit par Karl Pearson pour tester l'indépendance de deux caractères d'une population donnée. Ce test peut être utilisé pour tester l'indépendance de variables aussi bien quantitatives que qualitatives. Il renseigne sur l'existence d'une possible dépendance entre ces deux variables mais pas sur le sens de la dépendance. En considérant deux variables explicatives X et Y, l'objectif du test du khi-deux est de tester l'hypothèse nulle (H0) d'indépendance de X et Y contre l'hypothèse alternative (H1) de dépendance. Dans ce test, les variables sont supposées avoir un nombre fini de modalités dans le cas de variable qualitatives et un nombre fini de valeurs dans le cas de variables quantitatives (ceci implique un regroupement des valeurs). Les variables X et Y prennent respectivement n et m modalités ou valeurs. Ainsi pour chaque observations, les possibilités seront $X_1, \dots, X_n$ pour la variable X et $Y_1, \dots, Y_m$ pour la variable Y. En considérant simultanément les deux variables, on a au total n*m possibilités réparties dans le tableau de contingence suivant :

	$\mathbf{Y_1}$	$\mathbf{Y_2}$	. . .	$\mathbf{Y_j}$	. . .	$\mathbf{Y_m}$	**Total**
$\mathbf{X_1}$	O_{11}	O_{12}		O_{1j}		O_{1m}	L_1
$\mathbf{X_2}$	O_{21}						L_2
. . .							
$\mathbf{X_i}$	O_{i1}			O_{ij}			L_i
. . .							
$\mathbf{X_n}$	O_{n1}			O_{nj}		O_{nm}	L_n
Total	C_1			C_j		C_m	K

Où :

- O_{ij} représente le nombre d'observations ayant la modalité i pour la variable X et la modalité j pour la variable Y.
- L_i représente le nombre d'observations ayant i comme modalité pour la variable X
- C_j représente le nombre d'observations ayant j comme modalité pour la variable Y
- K est le nombre d'observations total

Sous l'hypothèse d'indépendance des variables, l'effectif espéré pour la modalité ij, est : $\mathbf{E_{ij}} = \frac{\mathbf{L_i * C_j}}{\mathbf{K}}$.

Pour comparer l'effectif espéré en cas d'indépendance à l'effectif observé en réalité, on calcule la statistique suivante **:** $\mathbf{T} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Cette statistique représente une erreur qui permettra de dire dans quelle mesure la différence observée entre l'effectif espéré et l'effectif observé n'est pas due au hasard de l'échantillon choisi. T suit asymptotiquement une loi du X^2 à (n-1)(m-1) degrés de liberté. En choisissant, la fiabilité du test à 5% c'est-à-dire 5 chances sur 100 de se tromper, on compare la valeur du khi-deux théorique en 0,05 ($X^2_{0.05}$) avec la valeur du test T.

La règle est la suivante :

- Si la valeur du test est inférieure au khi-deux théorique alors on ne rejette pas l'hypothèse nulle d'indépendance entres les variables au seuil de significativité fixé
- Si la valeur du test est supérieure au khi-deux théorique alors on rejette l'hypothèse nulle d'indépendance entre les variables au seuil de significativité fixé

Le test du khi-deux sous SAS permet d'obtenir la p-value qui s'interprète comme la probabilité d'obtenir sous H0 une statistique aussi grande que la statistique du khi-deux observée sur l'échantillon. L'objectif ici étant de voir la si la différence observée est réelle ou due à l'échantillon étudié.

- Si la p-value est élevée, cela signifie que dans ce test, il y avait de fortes chances d'observer ces différences sous l'hypothèse nulle. Cela conforte dans le choix de ne pas rejeter H0.
- Si la p-value est faible cela veut dire qu'il y a de faible chances d'observer ces différences sous H0. Dans ce cas H0 est peu crédible et on aura donc tendance à rejeter cette hypothèse.

Ainsi l'interprétation retenue est la suivante, si la p-value calculée est inférieure à alpha, on rejette l'hypothèse nulle au profit de l'hypothèse alternative et la statistique calculée est alors significative au risque alpha. Rappelons qu'alpha représente le taux d'erreur de première espèce, c'est-à-dire l'erreur liée au rejet de l'hypothèse H0 alors qu'elle est vraie.

Test d'adéquation d'Hosmer et Lemeshow

Le test de Hosmer-Lemeshow est basé sur une statistique qui mesure la qualité d'ajustement du modèle. Les données sont réparties en g groupes et la statistique est calculée à partir des effectifs observés et théoriques. La statistique à la base du test d'Hosmer-lemeshow est la suivante :

$$X^2_{hosmer-lemeshow} = \sum_{j=1}^{g} \frac{(N_i \Pi_i - \theta_i)^2}{N_i \Pi_i (1 - \Pi_i)}$$

Où :

- N_i représente le nombre d'observations du i-*ème* groupe,
- θ_i représente le nombre de réponses positives du i-*ème* groupe
- Π_i représente la moyenne des probabilités prévues dans le i-*ème* groupe.

En fait pour chaque groupe, on observe l'écart entre les valeurs prédites et observées. L'importance de la distance entre ces valeurs va déterminer la qualité d'ajustement du modèle. Afin de savoir si le modèle spécifié est bon ou mauvais en termes d'ajustement on compare la statistique du test avec une statistique du khi-deux avec g-2 degré de liberté. Les hypothèses à la base de ce test sont les suivantes :

- H_0 : ajustement bon (petite distance entre valeurs prédites et observées)
- H_1 : ajustement mauvais (grande distance entre valeurs prédites et observées)

L'hypothèse H_0 n'est pas rejetée lorsque la valeur de la probabilité (p-value) est supérieure au seuil fixé (ici α =5%). Dans le cas contraire, on refuse l'hypothèse nulle.

Test d'adéquation de Kolmogorov-Smirnov

Pour tester l'ajustement d'une loi à des données l'un des tests les plus utilisé est le test de Kolmogorov-Smirnov. Le test de Kolmogorov-Smirnov vise à détecter toute forme de différenciation entre deux distributions. Il repose sur le calcul de l'écart maximum entre les fonctions de répartition empirique des deux distributions. Ce test consiste à tester l'hypothèse H_0 selon laquelle les données étudiées ayant pour fonction de répartition empirique F suivent une loi de probabilité théorique donnée dont la fonction de répartition est notée F_0 contre l'hypothèse H_1 de différentiation des deux fonctions de répartitions :

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

L'écart maximal entre les valeurs théoriques et observées pour un échantillon $(x_1, \dots, x_n)$ se calcule par la statistique suivante :

$$D_n = \max_x |F(x) - F_0(x)|$$

Soit $U_n = \sqrt{n}\, D_n$. Il a été démontré que lorsque l'hypothèse nulle est vraie et la fonction de répartition est continue, U_n converge vers une distribution de Kolmogorov dont la fonction de répartition est la suivante :

$$\lim_{n\to\infty} P\left(\sqrt{n}\, D_n \leq x\right) = L(x) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2x^2}$$

Le test d'adéquation de Kolmogorov est donc construit en utilisant les valeurs critiques de la distribution de Kolmogorov. On dispose d'une table permettant de déterminer la valeur K_α de la loi de Kolmogorov tel que $P(K \leq K_\alpha) = 1 - \alpha$

Lorsque$\sqrt{n}\, D_n > K_\alpha$, l'hypothèse H_0 est rejetée et dans le cas contraire H_1est rejetée. Une p-value supérieure au seuil α signifie que la statistique D_n est inférieur à $K_\alpha/\sqrt{n}$ et donc que l'hypothèse d'adéquation des deux fonctions de répartition n'est pas rejetée.

Test d'adéquation du Khi-deux :

Le test d'adéquation du Khi-deux est aussi un test utilisé pour évaluer l'ajustement de données à une loi théorique. A la différence du test d'indépendance du khi deux on ne dispose pas de deux séries d'observation mais d'une seule série d'observation qui est comparée à la distribution théorique dont on souhaite tester l'ajustement.
Le principe consiste à découper le domaine de la distribution en intervalles. Dans chaque intervalle, on calcule à partir de la loi spécifiée sous l'hypothèse nulle la fréquence théorique attendue. On compte ensuite combien d'observations l'on retrouve dans chaque intervalle. Le nombre de classe est choisi de facon arbitraire mais pour que l'approximation soit bonne, il est nécessaire que les effectifs théoriques dans chaque classe soit supérieur à 5.
Il suffit alors de comparer les effectifs observés aux effectifs théoriques. L'écart entre les effectifs peut être mesuré par la statistique D :

$$D = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$$

Où :
$k\ est\ le\ nombre\ de\ classes$
$n\ est\ l'effectif\ total\ observé$
$O_i\ est\ l'effectif\ observé\ dans\ la\ classe\ i$
$p_i\ est\ la\ probabilité\ d'obtenir\ une\ observation\ de\ la\ loi\ de\ probabilité\ théorique\ dans\ la\ classe\ i$
$np_i\ est\ l'effectif\ théorique\ dans\ la\ classe\ i$

Le test du khi deux d'ajustement oppose les hypothèses suivantes :

H_0 : il y a adéquation entre la distribution observée et la distribution théorique
H_1 : la distribution observée est différente de la distribution théorique

Lorsque H_0 est vraie, la statistique D du test suit une Khi-deux à (k-r-1) avec r le nombre de paramètres estimés pour que la loi théorique soit entièrement déterminée. Si D > S on rejette H_0 au seuil α sinon on ne rejette pas H_0 avec S tel que :

$$\alpha = P(rejeter\ H_0\ quand\ H_0\ est\ vraie)$$
$$= P(D > S\ quand\ H_0\ est\ vraie)$$
$$= P(Khi - deux^2_{k-r-1} > S)$$

Une p-value supérieure au seuil de risqueα signifie que l'hypothèse d'adéquation des deux fonctions de répartition n'est pas rejetée.

Quelques lois de probabilités

Au cours de nos travaux nous avons utilisés certaines lois de probabilités. Le tableau ci-dessous récapitule leur densité ou fonction de probabilité :

Soit :

- la fonction gamma est définie pour a>0 par $\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{a-1} dx$

Nom de la loi	Support	Probabilités élémentaires $P(x = K)$ /Densité
Loi de Bernoulli B(p) $p \in]0,1[$	{0,1}	$P(x = 0) = 1 - p$ $P(x = 1) = p$
Loi binomiale négative $BN(n, p)$ $p \in]0,1[, n \in N^*$	N	$C_{k-1}^{n-1} p^n (1-p)^{k-n}$
Loi uniforme	$[a, b]$	$\frac{1}{b-a} 1_{]a,b[}(x)$
Loi de poisson $P(\lambda)$ $\lambda \in R^{+*}$		$e^{-\lambda} \frac{\lambda^k}{k!}$
Loi gamma $G(\alpha, \lambda)$	R^+	$\frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}$
Loi lognormale	R^+	$\frac{1}{x\sqrt{2\Pi s}} e^{-\frac{(lnx-u)^2}{2s^2}}$
Loi de Weibull $w(\eta, \beta)$ $\eta \in R^{+*}$ $\beta \in R^{+*}$	R^+	$\frac{\beta}{\eta} (\frac{x}{\eta})^{\beta-1} e^{-(\frac{x}{\eta})^\beta}$
Loi exponentielle	R^+	$\lambda e^{-\lambda x}$
Loi khi-deux	R^+	$G(\frac{n}{2}, \frac{1}{2})$
Loi Burr (a,b,s)	R^+	$\frac{as(\frac{x}{b})^s}{x(1+(x/b)^s)^{a+1}}$

Le modèle de provisionnement

En déterministe, le modèle utilisé pour déterminer la charge ultime est de type Chain Ladder. A partir de ce modèle, les coefficients de développements individuels sont calculés et les facteurs de développement sont estimés par une moyenne pondérée des coefficients individuels. Pour intégrer les connaissances et anticipations des experts, nous utilisons une matrice de pondération calibrée sur le triangle de charge. Cette matrice va permettre d'attribuer un poids à chaque coefficient de passage. A partir des facteurs de développement de Chain Ladder, la partie inférieure du triangle de charge est complétée ce qui permet d'obtenir une estimation de la charge ultime.

Considérons le triangle de paiement suivant :

	Année de développement						
Année d'origine	1	2	…	j	…	…	N
1	$X_{1,1}$	X_{12}	…	X_{1j}	…	…	X_{1N}
2	$X_{2,1}$	X_{22}	…	X_{2j}	…	…	
…	…	…	…	…	…		
i	…	…	…	$X_{i,j}$			
…	…	…	…				
…	…	…					
N	$X_{N,1}$						

On note :

- i l'indice de l'année de survenance du sinistre
- j l'indice de l'année de développement (ou de déroulement), comptée à partir de l'année de survenance
- N le nombre d'années de survenance
- Mle nombre d'années de développement
- X_{ij}le montant non cumulé de la variable de paiement ou charge à modéliser réglé à l'année comptable i+j, relatif aux sinistres en stock survenus au cours de l'année i vus j années après leur survenance
- C_{ij} Le montant cumulé de la variable de paiement ou de charge à modéliser $C_{ij} = \sum_{l=1}^{j} X_{ij}$

Le modèle de Chain Ladder est déterministe et permet de calculer les facteurs de développement en se basant sur les 3 hypothèses suivantes :

Considérons le triangle supérieur de la variable à modéliser cumulé : $\{C_{ij}\}_{1\leq i\leq N, 1\leq j\leq N-i+1}$,

- Il existe des facteurs de développement f_k tels que
$$E\left(C_{i,j+1}\middle|C_{i,j}\right) = f_j \times C_{i,j},\ \ i = 1, \dots, N\ j = 1, \dots, N-1$$
- Les paiements cumulés sont indépendants pour les années de survenance $i \neq k, i, k \in \{1, .. N\}$

Les facteurs de développement sont estimés par la formule suivante :

$$\forall \mathrm{j} = 1, \dots, \mathrm{M}-1; \hat{\mathrm{f}}_{\mathrm{j}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}-\mathrm{j}} \mathrm{C}_{\mathrm{i,j}} \times \mathrm{f}_{\mathrm{i,j}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}-\mathrm{j}} \mathrm{C}_{\mathrm{i,j}}}$$

Avec $f_{i,j}$ les facteurs de développement individuel tels que $\mathrm{f}_{\mathrm{i,j}} = \frac{C_{i,j+1}}{C_{i,j}}$

Dans le cas d'un Chain-Ladder pondéré par avis d'experts, la formule devient :

$$\forall \mathrm{j} = 1, \dots, \mathrm{M}-1; \hat{\mathrm{f}}_{\mathrm{j}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}-\mathrm{j}} \mathrm{C}_{\mathrm{i,j}} \times \mathrm{f}_{\mathrm{i,j}} \times P_{i,j}}{\sum_{\mathrm{i}=1}^{\mathrm{N}-\mathrm{j}} \mathrm{C}_{\mathrm{i,j}} \times P_{i,j}}$$

Avec $P_{i,j}$ les poids attribués aux facteurs de développement individuels $\mathrm{f}_{\mathrm{i,j}}$, en général 0 ou 1. La charge à l'ultime pour l'année de survenance i est alors estimée par la formule :

$\hat{C}_{i,N} = C_{i,N-i+1} \prod_{j=N-i+1}^{N-1} \hat{f}_j$.

Le montant de provision est alors déterminé par la différence entre la charge ultime et les paiements déjà effectués :

$$\hat{R}_i = \hat{C}_{i,N} - C_{i,N-i+1} = C_{i,N-i+1}\left(\prod_{j=N-i+1}^{N-1} \hat{f}_j - 1\right)$$

Et la provision globale est égale à $\hat{R} = \sum_{i=1}^{N} \hat{R}_i$

L'estimation de la cadence de paiements est effectuée en utilisant le triangle de paiements et la charge de sinistre ultime déterminée avec le triangle de charge. La méthodologie utilisée est celle du modèle de Mack. On considère le triangle des paiements non cumulés.

Dans une première étape, on calcule les cadences de paiements brutes par année de développement en faisant le rapport entre les deux valeurs suivantes :

- Le montant total des paiements pour toutes les années de survenance situées au-dessus de la diagonale dans le triangle des paiements non cumulés, pour l'année de développement considérée.
- Le montant total de charge ultime estimée pour toutes les années de survenance situées au-dessus de la diagonale pour l'année de développement considérée.

$$\mathrm{y_j} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}+1-\mathrm{j}} X_{\mathrm{i,j}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \hat{C}_{i,N}}, \mathrm{j} = 1, \dots, \mathrm{N}$$

Où :

$\mathrm{x_{i,j}}$ représente les paiements non cumulés (observés par survenance i et par développement j) et $\hat{C}_{i,N}$ représente l'estimation de la charge de sinistre à ultime estimée par le modèle en t=0

Une fois qu'on a obtenu les cadences de règlement par année de développement, on applique une méthode de lissage. En effet, à partir d'une certaine année de

développement, on considère que l'actuaire n'a pas suffisamment de visibilité et ne dispose pas d'assez de données pour obtenir une estimation de la cadence de paiements assez fiable.

A partir de l'horizon j_1 fixé par l'actuaire, on réalise un retraitement en utilisant un lissage à partir de la formule $ln(y_j) = \alpha - \beta j$ où :

- j: $l'année\ de\ développement$
- y_j : $cadence\ de\ règlement\ brute\ pour\ l'année\ de\ développement\ j$
- α, β: $deux\ paramètres\ réels$

On estime les paramètres $(\hat{\alpha}, \hat{\beta})$ par la méthode des moindres carrés ordinaires. A partir, de ces estimateurs et de l'expression $\tilde{y}_j = exp(\hat{\alpha} - \hat{\beta} j)$, $j = j_1 + 1, .., N$, on peut déduire les cadences de règlements pour toute année de développement supérieure à l'année choisie comme horizon.

Ces cadences sont calculées au-delà de l'année N, jusqu'à un horizon k_2 déterminé au préalable.

La dernière étape de la procédure consiste à normaliser les cadences obtenues en divisant chaque cadence par la somme totale des cadences pour toutes les années de développement.

Ainsi, les cadences de règlements obtenus sont :

$$\begin{cases} \tilde{y}_j = \frac{\tilde{y}_j}{s}, & j = 1, \dots, j_1 \\ \tilde{y}_j = \frac{\exp(\hat{\alpha} - \hat{\beta} j)}{s}, & j = j_1 + 1, .., N, \dots, j_2 \end{cases} \quad \text{Où } s = \sum_{i=1}^{j_2} \tilde{y}_j$$

<u>En stochastique</u>, la charge ultime est déterminée à partir d'un modèle basé sur un modèle de réplication par Bootstrap à 1 an. Ce modèle reprend les hypothèses de Mack à savoir :

- Il existe des facteurs de développement $f_j\ tels\ que\ E(C_{i,j+1} | C_{i,j}) = f_j \times C_{i,j}$, $i = 1, \dots, N\ j = 1, \dots, N-1$
- Il existe des variances $\sigma_j > 0\ tels\ que\ Var(C_{i,j+1} | C_{i,j}) = \sigma_j^2 \times C_{ij}, C_{i,j}$, $i = 1, \dots, N\ j = 1, \dots, N-1$

Les paramètres à estimer sont : les facteurs de développement $(f_j)_{1 \le j \le N-1}$ et des paramètres de volatilité associés $(\sigma_j^2)_{1 \le j \le N-1}$. L'estimation des facteurs de développement est la même que pour le modèle de Chain ladder. L'estimation des paramètres de volatilité associés, $\forall\ j = 1, \dots, N-1$ est :

$$\widehat{\sigma}_k^2 = \begin{cases} \frac{1}{N-k-1}\sum_{i=1}^{N-k} C_{i,k}\left(f_{ik} - \hat{f}_k\right)^2, \text{pour } k \in \{1, \dots, N-2\} \\ \min\left(\frac{\widehat{\sigma}_{N-2}^4}{\widehat{\sigma}_{N-3}^2}; \min\left(\widehat{\sigma}_{N-2}^2; \widehat{\sigma}_{N-3}^2\right)\right), \text{pour } k = N-1 \end{cases}$$

Pour avoir une vision de l'erreur de prédiction à 1 an, on considère la différence entre l'estimation de la charge ultime calculée en t = N avec les informations disponible à cette date et la prédiction de la charge ultime calculée en t = N + 1. La différence entre ces deux estimations est appelée Claim Developpement Result. La CDR observable est définie par :

$$\widehat{CDR}_i(N+1) = \hat{R}_i(N) - (X_{i,N-i+1} + \hat{R}_i(N+1))$$

Où :

$\hat{R}_i(N)$ est l'estimateur de la provision à la date t= N pour l'année de survenance i

$\hat{R}_i(N+1)$ est l'estimateur de la provision à la date t=N+1 pour l'année de survenance i

$X_{i,N-i+1}$ Correspond à l'incrément de paiement à un an

$$\widehat{CDR} = \sum_{i=1}^{N} \widehat{CDR}_i(N+1)$$

Pour calculer la charge ultime en t=1 et avoir une vision de l'erreur de prédiction à 1 an, la méthode utilisée est celle du bootstrap il s'agit de procéder à des tirages aléatoires avec remises, sur les résidus associés aux facteurs de développement, pour créer des pseudo-données. La méthode de projection et d'estimation présentée ci-dessus est ensuite appliquée aux pseudo-données. Afin d'obtenir une volatilité à un an, conformément à Solvabilité II, les simulations sont réalisées de telle sorte que la volatilité des distributions des $C_{i,j}$ à un an contienne l'erreur de processus et l'erreur d'estimation. Au-delà d'un an, seule l'erreur d'estimation est prise en compte dans la volatilité des distributions des $C_{i,j}$.

Dans un premier temps un triangle de résidus est construit autour des facteurs de développement estimés. A partir de ce triangle de résidus, et à chaque itération du bootstrap, Nous réalisons un tirage avec remise de résidus ce qui nous donne un nouveau triangle de résidus bootstrapés. Par une inversion de la formule précédente, nous calculons le triangle des facteurs de développement bootstrapés. A partir du triangle des facteurs de développement individuels bootstrapés, on calcule les facteurs de développement bootstrapés par année de développement en faisant une moyenne pondérée. En utilisant les facteurs de développement, on complète la diagonale inférieure du triangle des paiements cumulés. A partir du trapèze obtenu, on estime à nouveau les facteurs de développement par une moyenne pondérée des facteurs individuels. Enfin on réalise une projection à l'ultime du trapèze en utilisant les facteurs de développement obtenu à partie du trapèze et on obtient les montants de charges ultime pour chaque simulation.

Méthode de découpage d'une variable en classe

Classes équireparties

Supposons qu'on veut construire 10 classes ayant à peu près le même nombre d'observations. L'idée est donc de trier les observations par ordre croissant d'ancienneté puis de remplir chaque classe avec un nombre d'observations plus ou moins identique.

Le nombre de classes étant fixé, on peut calculer l'amplitude moyenne des classes en faisant le rapport entre le nombre total d'observations N et le nombre de classes souhaité. On obtient comme amplitude moyenne des classes $\frac{N}{10}$.

Pour obtenir des intervalles ayant le même effectif, on calcule les déciles ($q_1, q_2, \ldots, q_9$). Le 1er décile par exemple est la donnée pour laquelle le nombre d'observations inférieures représente 10% des données et le nombre d'observations supérieures représente 90% des données. Ainsi, le 5ème décile est la donnée qui sépare les observations en deux parties.

Les observations seront réparties en classe suivant le principe suivant : celles dont l'ancienneté est inférieure au premier décile seront regroupées dans la classe 1. Ensuite les observations dont l'ancienneté est comprise entre le 1er décile et le 2nd décile seront regroupées dans la classe 2. Puis les observations dont l'ancienneté est comprise entre le 2nd décile et le 3ème décile seront regroupées dans la classe 3 et ainsi de suite jusqu'à la classe 10. A partir des déciles obtenus, les bornes des intervalles des classes pourront être déterminées.

Classes homogènes et différenciés

L'objectif d'une analyse en cluster est de construire des classes homogènes c'est-à-dire dont les éléments d'une même classe sont similaires ayant une faible variance intra classe mais séparées ce qui implique une forte variance interclasse. Pour cela, nous utilisons pour la construction des clusters une méthode de classification mixte.

Le principe de la classification est de regrouper entre elles certaines observations que l'on juge similaire selon un critère défini. L'idée est de répartir les n observations dont on dispose et qui sont caractérisées par des variables, en p classes les plus homogènes et différenciées les unes des autres. De façon générale, il existe deux principales méthodes de classification :

- la classification hiérarchique
- le partitionnement

La classification hiérarchique comme son nom l'indique permet d'obtenir une hiérarchie de groupe d'observation. La classification ascendante hiérarchique est la plus couramment utilisée et l'expérience montre qu'elle fournit des résultats assez cohérents. Elle a la particularité de partir de l'ensemble des n données puis de créer au fil des itérations des sous-groupes ce qui aboutit schématiquement à obtenir un arbre. Pour la

classification descendante hiérarchique, le principe est contraire. On part d'un groupe contenant l'ensemble des n observations puis au fil des itérations ce groupe est divisé en deux parties qui seront à leur tour divisées en deux groupes et ceci jusqu'à ce que les sous-groupes soient réduits à une seule observation. L'un des inconvénients majeur de la classification descendante hiérarchique est le fait qu'avant chaque scission, il est nécessaire d'examiner toutes les séparations possibles et de retenir la meilleure selon le critère choisi. C'est ce qui explique la préférence pour la classification ascendante hiérarchique (CAH). La CAH comprend deux grandes étapes, une étape initiale au cours de laquelle s'effectue le choix de l'indice de dissimilarité et de l'indice d'agrégation et une deuxième étape qui est l'algorithme à proprement dit. Par souci de compréhension, l'algorithme sera présenté avant l'étape initiale.

- Algorithme

Soit l le nombre de classe. Initialement il est égal à n car on a au total n singletons contenant chacune des observations.

l=n

1. On dispose de n groupes contenant chacun une observation $\boldsymbol{C_1}$,…., $\boldsymbol{C_n}$
2. On calcule la matrice des distances entres les classes deux à deux
3. On identifie les deux classes C_i et C_j ayant la distance minimale telles que: $\boldsymbol{d(C_i, C_j) = \min_{l,k} d\,(C_l, C_k)}$
4. On regroupe ces deux éléments les plus proches en une seule classe
5. On met à jour de la matrice des distances en remplaçant les deux éléments regroupés par la classe qui en découle notée $\boldsymbol{C_{ij}}$

l=l -1 et on répète les étapes 2 à 6 de l'algorithme.

Tant que l >1, les itérations continuent et à la fin des itérations toutes les observations sont agrégées en une seule classe.

<u>Choix d'une distance entre les observations (indice de dissimilarité)</u>
La distance euclidienne qui est une distance géométrique dans un espace multidimensionnel qui se calcule par la formule : $\boldsymbol{d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$

<u>Choix d'un critère de regroupement des individus qui sera minimisé (indice d'agrégation)</u>
La méthode de WARD est la plus utilisée. L'objectif de cette méthode est de minimiser à chaque itération, le gain d'inertie intra classe et la perte d'inertie interclasse. La distance utilisée est la distance de Ward qui se calcule en utilisant les barycentres des deux classes au carré, pondérée par les poids des classes. Le barycentre d'une classe C_k est le point G_k tel que :$\boldsymbol{G_k = \sum_{i \in K} p_i x_i}$

p_i *est le poids de chaque observation* x_i *de la classe* C_k.
le poids souvent attribué à chaque observation est $\frac{1}{n}$
On en déduit la formule du barycentre global qui est la suivante : $\boldsymbol{G = \sum_{i=1}^{n} p_i x_i}$

Soit G_1,…, G_n les centres de gravité respectifs des classes C_1,…, C_n et p_{c_1} ,…, p_{c_n}le poids de chaque ces classes :

Méthode de Ward	$\boldsymbol{D(C_k, C_{ij}) = \frac{p_{c_k} * p_{c_{ij}}}{p_{c_k} + p_{c_{ij}}} d(G_k, G_{ij})^2}$

La méthode moyenne mobile : c'est l'une des méthodes de partitionnement la plus répandue. Plus connue sous le nom de K-means, cette méthode vise à classer les n observations en q classes. La particularité de cette méthode est qu'elle se fait de façon automatique. Il n'y a pas de lien hiérarchique entre les clusters contrairement à la méthode précédente. L'objectif de cette méthode est de trouver parmi un ensemble d'observation, la partition qui optimise un critère prédéfini notamment la distance. L'algorithme à la base des centres mobiles est le suivant :

1. La première étape consiste à sélectionner les centres. En fait il s'agit de choisir q points au hasard dans l'espace.
2. L'étape suivante consiste à calculer les matrices des distances entre tous les individus et les q centres.
3. L'étape suivante consiste à former q groupes en prenant chaque centre et en y associant les observations les plus proches en termes de distances. Ainsi pour chaque observation, l'idée est de comparer les distances aux q centres et de la regrouper avec le centre pour lequel on observe la plus petite distance. A l'issue de cette étape le n observations sont regroupées en q groupes.
4. La quatrième étape est le calcul du centre de gravité de chacun des q sous-groupes formés. Puis les centres de gravité calculés sont considérés comme les nouveaux centres puis on répète

Le problème de cette méthode est essentiellement dû à la multitude de partitions possibles, au choix assez arbitraire des q centres initiaux et au choix a priori du nombre de classes. La classification mixte consiste à réaliser une classification ascendante hiérarchique puis à appliquer la méthode des centres mobiles. Le fait de réaliser une CAH permet de définir le nombre de classe optimale et d'utiliser le barycentre de ces classes comme centres mobiles initiaux de la méthode de partitionnement. En appliquant la méthode des K-means, la répartition en classe obtenue est plus flexible et certaines observations peuvent être réaffectées contrairement aux classes obtenues par la méthode CAH. La mise en application de cette méthode a consisté dans un premier temps en la réalisation d'une classification ascendante hiérarchique avec la distance euclidienne comme indice de dissimilarité et la méthode de Ward comme critère d'agrégation en utilisant la Proc CLUSTER de SAS. Après le choix du nombre optimal de clusters, les barycentres

des classes obtenues, sont calculés. En considérant ces barycentres comme centroïd initiaux, la méthode des centres mobiles est appliquée par la Proc FASTCLUS de SAS.