Mémoire présenté le : 05/09/2019

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

Par : Julius QUIQUET

Titre : Méthode d'estimation de la charge ultime en RC corporelle automobile basée sur des données individuelles

Confidentialité : ☒ NON OUI ☐ (Durée : ☐ 1 an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires
.....
.....

.....

Membres présents du jury de l'ISFA

.....

.....

.....

Entreprise :
Nom : MMA

Signature :

Directeur de mémoire en entreprise :
Nom : Boris BADAULT

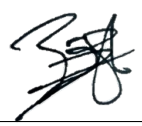
Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise



Signature du candidat

Résumé

Mots-clés : Méthode de projection individuelle, *Text Mining*, *Machine Learning*, forêts aléatoires, *Gradient Boosting*

L'objectif de ce mémoire est de proposer une méthode d'estimation de la charge ultime des sinistres en Responsabilité Civile Corporelle (RCC) automobile basée sur des données individuelles, c'est-à-dire propres aux sinistres étudiés. L'idée principale est donc d'intégrer un maximum d'informations sur les sinistres dans le but d'affiner la prédiction du coût ultime.

Cette méthode individuelle vise à challenger l'utilisation des méthodes agrégées (de type Chain Ladder) très largement plébiscitées par les assureurs pour des problématiques de projection de la charge sinistre à l'ultime au sens large.

A partir des sinistres issus du marché des professionnels de l'automobile et du périmètre des particuliers, la base d'étude est enrichie à l'aide de données comptables et des informations sur les victimes. De plus, une étude de *Text Mining* est réalisée afin d'intégrer des données non structurées relatives aux circonstances des sinistres.

Après avoir présenté la méthode de projection individuelle développée dans ce mémoire, deux algorithmes de *Machine Learning*, à savoir les forêts aléatoires *(Random Forest)* et le *Gradient Boosting Machine (GBM)*, sont proposés pour l'implémenter.

Les résultats obtenus sont ensuite comparés avec Chain Ladder, la méthode de référence pour estimer le coût ultime des sinistres. Enfin, un backtesting est réalisé en comparant ces deux méthodes aux coûts des sinistres réellement observés. Cet exercice nous permet de conclure sur les avantages et inconvénients relatifs à l'utilisation des méthodes individuelles.

D'un point de vue opérationnel, cette méthode vise à mieux appréhender la sinistralité corporelle en Responsabilité Civile automobile dans le cadre du pilotage des résultats et de la rentabilité du marché des professionnels de l'automobile.

Abstract

Keywords : Individual modeling, *Text Mining*, *Machine Learning*, *Random Forest*, *Gradient Boosting*

The purpose of this study is to propose a method for estimating the ultimate cost of bodily injury claims in Motor Third Party Liability (MTPL) based on individual data, i.e specific to the claims studied. The main idea is therefore to integrate a maximum of information on the claims in order to refine the prediction of the ultimate cost.

This individual method aims to challenge the use of aggregated methods (e.g Chain Ladder) very widely favored by insurers in the context of ultimate cost estimation in the broad sense.

Based on claims from the automotive professionals and personal automobile insurance market, the study data base is enriched with accounting data and information on victims. Moreover, a *Text Mining* study is performed to integrate unstructured data related to the circumstances of the claims.

After presenting the individual projection method developed in this study, two *Machine Learning* algorithms, namely *(Random Forest)* and *Gradient Boosting Machine (GBM)*, are proposed for the implementation.

Then, the results obtained are compared with Chain Ladder, the reference method for estimating the ultimate cost of claims. Finally, a backtesting is carried out by comparing these two methods to the costs of the claims actually observed. This exercise allows us to conclude on the advantages and disadvantages of using the individual methods.

From an operational point of view, this method aims to better understand the bodily injury claims in Motor Third Party Liability (MTPL) for steering the results and profitability of the automotive professionals market.

Remerciements

Tout d'abord, je tiens tout particulièrement à remercier mon tuteur en entreprise Boris Badault, responsable du service actuariat des marchés professionnels de l'automobile, CSPL, agricole, loisirs, associations et risques divers, pour son encadrement, son expertise dans le domaine de l'assurance des professionnels et ses conseils lors de la réalisation de ce mémoire.

Je remercie mon tuteur pédagogique, Nicolas Leboisne, pour sa disponibilité et ses précieuses remarques lors de la construction de ce mémoire.

Je remercie également toutes les personnes qui ont participé de près ou de loin à la réalisation de ce mémoire, que ce soit pour sa relecture ou bien pour répondre à mes différentes interrogations, en particulier : Alexandre Dias Lopes pour ta relecture et tes conseils sur l'application de la méthode Chain Ladder, Yoann Gouyen pour ta relecture et tes remarques pertinentes sur la partie *Machine Learning* et Jérémy Pougeon pour ces nombreux échanges très enrichissants et pour ton aide apportée sur la mise en forme de ce mémoire sous LaTeX.

Enfin je remercie l'ensemble du corps enseignant de l'Institut du Risque et de l'Assurance (IRA) du Mans m'ayant permis d'intégrer l'Institut de Science Financière et d'Assurances (ISFA).

Table des matières

Résumé **3**

Abstract **4**

Introduction **8**

1 Contexte de l'étude **9**
1.1 Le marché des professionnels de l'automobile 9
1.1.1 Présentation du marché 9
1.1.2 Le produit « MMA Pros de l'Auto » 10
1.1.3 Présentation du contexte et de la problématique 11
1.2 Généralités sur la garantie responsabilité civile automobile 12
1.2.1 Conventions applicables en responsabilité civile automobile 12
1.2.2 Projet de réforme de la garantie responsabilité civile automobile 14

2 Construction de la base d'étude **15**
2.1 La base sinistres 15
2.2 Enrichissement de la base sinistres 16
2.2.1 Intégration de la sinistralité issue du marché des particuliers 16
2.2.2 Vérification de l'homogénéité du coût des sinistres 17
2.3 La base comptable 18
2.4 La base victimes 19
2.5 Retraitement de la base d'étude 20
2.6 *Text Mining* sur les circonstances sinistres 22
2.6.1 Prétaitements des données et normalisation du texte 22
2.6.2 Analyse des résultats du *Text Mining* 24
2.6.3 Intégration des informations issues du *Text Mining* dans la base d'étude 27
2.7 Revalorisation des sinistres 33
2.7.1 Construction des statistiques *as-if* 35
2.8 Analyse descriptive de la base d'étude 37
2.8.1 Présentation générale 37
2.8.2 Analyse des variables explicatives 41

3 Méthode d'estimation de la charge ultime basée sur des données individuelles **44**
3.1 Motivations du développement d'une méthode individuelle 44
3.2 Généralités sur la vie d'un sinistre en assurance IARD 45
3.3 Présentation de la méthode de projection individuelle 47

3.3.1 Formalisation . . . 47
3.3.2 Suppression des sinistres « millionnaires » . . . 49
3.3.3 Construction des bases pour l'implémentation . . . 50

4 Implémentation **53**
4.1 Introduction aux méthodes ensemblistes . . . 53
4.2 Présentation des méthodes d'apprentissage automatique . . . 54
4.2.1 Les forêts aléatoires (*Random Forest*) . . . 54
4.2.2 Le *Gradient Boosting Machine (GBM)* . . . 56
4.3 Optimisation des hyperparamètres *(tuning)* . . . 59
4.3.1 Cas des forêts aléatoires . . . 60
4.3.2 Cas du *Gradient Boosting Machine (GBM)* . . . 63
4.4 Résultats de la méthode de projection individuelle . . . 70
4.4.1 Comparaison des indicateurs de performance sur les bases test . . . 70
4.4.2 Analyse des erreurs de prédictions individuelles . . . 73
4.4.3 Importance des variables dans la construction des modèles . . . 76

5 Backtesting et comparaison avec Chain Ladder **80**
5.1 Présentation du backtesting . . . 80
5.2 Application de la méthode Chain Ladder . . . 83
5.2.1 Validation des hypothèses de Chain Ladder . . . 83
5.2.2 Résultats . . . 85
5.3 Application de la méthode de projection individuelle . . . 87
5.3.1 Adaptation de la méthode de projection individuelle . . . 88
5.3.2 Résultats . . . 89
5.4 Comparaison avec Chain Ladder et discussion sur l'apport des méthodes individuelles . 91

Conclusion **93**

Bibliographie **96**

Annexes **99**
A Recours forfaitaires des conventions inter-assureurs en Responsabilité Civile automobile 99
B Test statistique de Mann - Whitney . . . 100
C Liste des variables explicatives . . . 101
D Valeurs optimales des hyperparamètres pour chaque modèle de projection . . . 102

Introduction

Une estimation précise de la sinistralité en Responsabilité Civile Corporelle (RCC) constitue un enjeu majeur pour les assureurs compte tenu de sa forte volatilité, tant en termes de coût que de délai de règlement. Ajouté à cela, le coût moyen de cette typologie de sinistres ne cesse d'augmenter depuis maintenant plusieurs années et ce, combiné à une règlementation en constante évolution.

Le choix de cette garantie pour cette étude s'explique par la part relativement importante des dommages corporels en responsabilité civile automobile constatée sur le marché des professionnels de l'automobile. La maîtrise de cette sinistralité est donc un enjeu majeur pour la direction dans le pilotage des résultats et donc de la rentabilité de ce marché.

L'objectif de ce mémoire est de développer une méthode d'estimation de la charge ultime des sinistres en RCC automobile basée sur les informations propres aux sinistres étudiés.
L'environnement de travail ainsi que la présentation du contexte et des objectifs de l'étude constituent la première partie de ce mémoire. Quelques généralités sur la Responsabilité Civile Corporelle automobile sont ensuite présentées.

La deuxième partie est consacrée à la récupération d'un maximum d'informations sur les sinistres corporels en responsabilité civile automobile survenus sur la périodes 2009-2017. Compte tenu de la faible volumétrie de cette typologie de sinistres sur le marché des professionnels de l'automobile, la sinistralité issue du marché automobile des particuliers est intégrée pour cette étude. La base sinistres est ensuite enrichie à l'aide d'informations provenant d'une base comptable, d'une base victimes ainsi que d'éléments textuels, à savoir les circonstances des sinistres renseignées par les gestionnaires. Une étude de *Text Mining* est donc réalisée dans le but d'exploiter, d'un point de vue statistique, ces données initialement non structurées.

Après avoir présenté la méthode de projection individuelle développée dans ce mémoire, elle est ensuite implémentée à l'aide de deux méthodes d'apprentissage automatique *(Machine Learning)* : les forêts aléatoires (*Random Forest*) et le *Gradient Boosting Machine (GBM)*. Pour juger de sa qualité prédictive, cette partie fera également l'objet d'une confrontation des résultats issus des deux algorithmes avec les coûts ultimes réellement observés.

Enfin, un benchmark est réalisé en comparant les résultats avec ceux d'une méthode agrégée de place, la méthode Chain-Ladder. Les avantages et inconvénients de l'implémentation d'une méthode individuelle sont ensuite discutés en dernière partie de ce mémoire.

Chapitre 1

Contexte de l'étude

1.1 Le marché des professionnels de l'automobile

L'étude réalisée dans le cadre de ce mémoire porte sur le produit MMA dédié aux professionnels de l'automobile, à savoir « MMA Pros de l'auto ». Ce marché étant relativement moins connu que l'assurance automobile ou bien la Multirisques Habitation, cette partie consiste à présenter brièvement sa composition ainsi que le produit « MMA Pros de l'Auto ».

1.1.1 Présentation du marché

A fin 2015, le potentiel de ce marché était de 160 000 entreprises réparties de la façon suivante [1] :

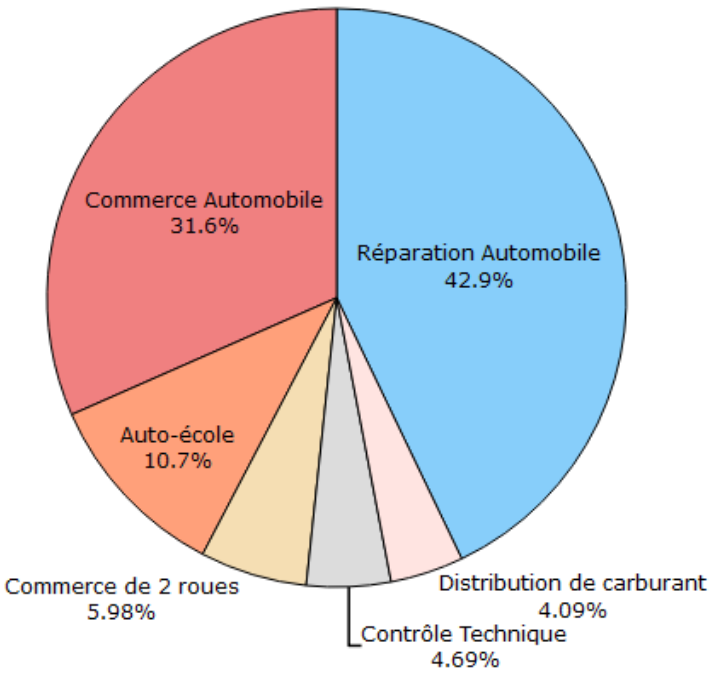


FIGURE 1.1 – Composition du marché des professionnels de l'automobile à fin 2015

1. Source : INSEE

L'activité est séparée globalement en deux segments majeurs : le commerce et la réparation automobile. Ces deux secteurs représentaient en 2015 à eux seuls près de 80% du marché. On retrouve ensuite les auto-écoles (11%), les centres de contrôle technique (5%) et la stations-services (4%).

Pour MMA, ce marché représente en 2018 près de 129 millions d'euros de chiffre d'affaires, soit plus de 26 000 contrats en portefeuille. En partie grâce à sa proximité géographique avec le sport automobile (les 24 heures du Mans), MMA est leader sur ce marché avec près d'un contrat sur six souscrit.

De manière générale, est professionnel de l'automobile toute personne qui vend, contrôle, répare des véhicules terrestres, pour le compte de tiers, ou qui intervient sur ceux-ci, distribue du carburant. A titre indicatif, les mécaniciens réparateurs, les ventes de véhicules automobiles et les carrossiers-tôliers représentent 68% du chiffre d'affaires sur ce marché en 2018.

1.1.2 Le produit « MMA Pros de l'Auto »

Ce produit « 2 en 1 » créé en 2008 permet d'offrir la possibilité aux professionnels de l'automobile de protéger à la fois leur activité mais aussi leurs véhicules. Pour couvrir l'activité de ses assurés, MMA propose de garantir la couverture de la responsabilité civile exploitation et professionnelle, des bâtiments, machines, mobiliers et matériels, des pertes d'exploitation et de la protection juridique.

Concernant l'assurance des véhicules, MMA couvre l'ensemble du parc de véhicules, c'est-à-dire les véhicules confiés par les clients, ceux destinés à la vente et ceux de l'entreprise y compris les véhicules personnels des représentants légaux, de leur conjoint, des associés exerçant une activité permanente dans l'entreprise.

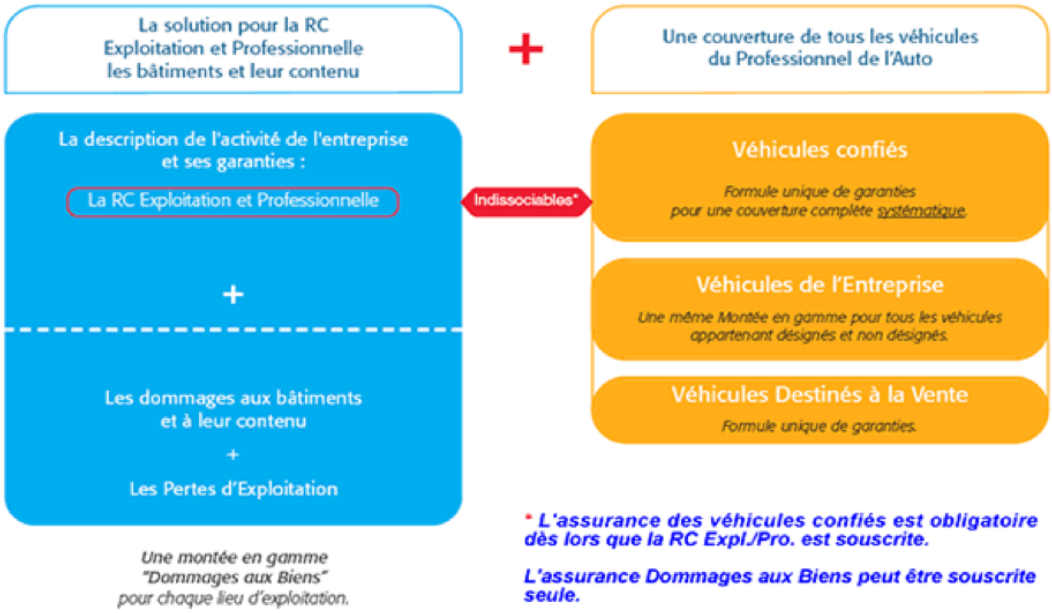


FIGURE 1.2 – Structure du produit MMA Pros de l'Auto

1.1.3 Présentation du contexte et de la problématique

Le marché des professionnels de l'automobile est actuellement au cœur des préoccupations de la Direction Centrale Entreprise de MMA compte tenu des résultats dégradés observés depuis déjà quelques années. Suite à ce constat, un plan d'action est actuellement mené afin d'atteindre le niveau de rentabilité souhaité à moyen terme tout en conservant une position de leader sur ce marché avec un chiffre d'affaires qui ne cesse d'augmenter.

De la conception du produit à la souscription, l'ensemble des processus a été revu afin d'établir des axes d'amélioration. Sur la partie technique, le redressement des résultats passe naturellement par une remise à plat de la structure tarifaire du produit (définition de nouveaux critères techniques, mise à jour des coefficients tarifaires etc..) visant à rééquilibrer le tarif proposé aux professionnels de l'automobile en représentation du risque assuré et ce, de manière structurelle. A cela, il faut également ajouter des actions tarifaires à mener sur les contrats présents en portefeuille qui pourront être appliquées lors de la période de renouvellement.

Pour les contrats « MMA Pros de l'Auto », cette période consiste à définir les évolutions tarifaires pour l'année à venir en fonction des résultats observés sur les exercices passés, essentiellement au travers d'une analyse du ratio sinistres à primes (S/P) des contrats.

En période de renouvellement, une connaissance précise de la sinistralité passée est nécessaire afin d'appliquer la mesure tarifaire adaptée au risque assuré. Cependant, le coût des sinistres survenus n'est pas toujours connue de façon certaine, certains sinistres sont toujours ouverts au moment de l'analyse des résultats techniques.

A titre d'exemple, dans le cadre du renouvellement des contrats « grands comptes », la projection des ratios sinistres à primes à l'ultime est réalisée à l'aide d'une méthode dérivée du Chain-Ladder. Bien que ces méthodes dites agrégées soient très appréciées par les assureurs compte tenu de leur facilité d'implémentation, elles se révèlent particulièrement instables en présence de sinistres présentant des cadences particulières de règlements. De plus, dans certains cas, la faible volumétrie de sinistres ne permet pas d'appliquer rigoureusement de telles méthodes.

L'objectif de ce mémoire est de challenger la méthodologie actuelle en proposant une nouvelle méthode de projection de la charge sinistre à l'ultime. Au-delà du renouvellement des contrats, l'enjeu est de mieux appréhender la sinistralité en responsabilité civile corporelle automobile pour répondre à des problématiques relatives au pilotage des résultats et de la rentabilité du marché des professionnels de l'automobile.

Bien que la construction d'une méthode de projection individuelle soit très souvent abordée dans la littérature actuarielle pour des problématiques de provisionnement, nous nous intéressons dans le cadre ce mémoire uniquement à l'estimation du coût ultime des sinistres. Bien évidemment, la méthode développée peut être utilisée par les équipes en charge du provisionnement.

Le choix d'étudier le développement des sinistres corporels en responsabilité civile automobile est également motivé par son poids relativement important dans la sinistralité du marché des professionnels de l'automobile. En effet, ces sinistres représentent plus de 15% de la charge sinistre globale observée sur la période 2012-2017 mais seulement un peu plus de 2% en nombre.

1.2 Généralités sur la garantie responsabilité civile automobile

La garantie Responsabilité Civile est obligatoirement souscrite lors de la souscription d'un contrat en assurance automobile depuis le 27 février 1958. Elle a pour vocation de protéger les victimes d'accidents matériels et/ou corporels.

1.2.1 Conventions applicables en responsabilité civile automobile

Deux conventions inter-assureurs d'indemnisation existent pour faciliter la gestion des sinistres et ainsi accélérer l'indemnisation des victimes lorsque les montants de préjudices sont relativement faibles. Sauf quelques exceptions, l'ensemble des acteurs opérant dans l'assurance des risques circulation adhérent à ces deux conventions.

Dès lors qu'un préjudice corporel est recensé, le sinistre associé à la garantie responsabilité civile automobile est référencé comme corporel par les gestionnaires sinistres. Néanmoins cela ne signifie pas que seuls des préjudices corporels ont été constatés, il est très fréquent de retrouver des règlements au titre de dommages matériels pour un sinistre référencé en tant que corporel. Rappelons que les sinistres corporels et matériels sont regroupés au sein de la même garantie, la RC automobile. Chez MMA, la distinction se fait à un second niveau en filtrant sur la nature des dommages constatés au moment de la déclaration du sinistre.

En conséquence, la convention IRSA relative à l'indemnisation des dommages matériels peut être applicable pour un sinistre référencé comme corporel en RC automobile.

Convention IRCA

La convention d'Indemnisation et de Recours Corporel Automobile créée en 2002 concerne les sinistres corporels survenus sur le territoire français pour lesquels au moins deux véhicules terrestres à moteur sont impliqués sous réserve que les deux organismes assureurs sont adhérents de cette convention. A cela, il faut également ajouter une contrainte sur le degré de préjudice de la victime : le taux d'AIPP (Atteinte à l'Intégrité Physique et Psychique) ne doit pas dépasser 5%. Dans le cas inverse ou en cas de décès, l'indemnisation du préjudice corporel de la victime est fondée en droit commun, c'est-à-dire en coût réel.

Dans le cas où la convention est applicable, l'assureur de la victime non responsable indemnise en droit commun son assuré tout en présentant un recours à la compagnie d'assurance adverse. Le montant de ce recours est pondéré par le niveau de responsabilité de l'adversaire. A titre d'exemple, dans le cas où la compagnie mandatée indemnise son assuré qui est jugé responsable à hauteur de 50%, alors un « demi » recours sera présenté à la compagnie adverse.

Deux cas sont à distinguer lors de l'application de la convention IRCA :

— Si le taux d'AIPP de la victime est nul alors le recours est forfaitaire, il s'élève à 1 518 euros en 2018 [2].

— Si le taux d'AIPP est compris entre 1% et 5% (inclus) alors le recours est réel sur la base de fourchettes d'indemnités dont les valeurs sont encadrées.

Convention IRSA

La convention d'Indemnisation directe des Assurés (IDA) change de nom en 1974 pour devenir la convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'Assurance automobile (IRSA) [3]. Cette convention est destinée à faciliter l'indemnisation des préjudices matériels en cas d'accidents de la circulation survenus en France et à Monaco en définissant les responsabilités des parties concernées. Elle est applicable lorsqu'au moins deux véhicules terrestres à moteur sont impliqués dans l'accident, sous réserve que les deux organismes assureurs soient adhérents (majoritairement le cas en France). Le barème de cette convention est composé de 13 cas d'accident de la circulation pour lesquels les niveaux de responsabilité des deux parties sont renseignés [4].

Lorsque cette convention est applicable, l'assureur indemnise directement son assuré pour les préjudices subis et évalue lui-même son niveau de responsabilité dans l'accident. En contrepartie, l'assureur se retourne vers l'organisme assureur adverse selon les modalités de recours établis par la convention IRSA.

On distingue deux cas :

— Si le montant des préjudices matériels subis est inférieur à 6 500 euros, alors le recours est forfaitaire [5] et est proportionnel au niveau de responsabilité de l'auteur des dégâts.

— Si l'indemnisation dépasse 6 500 euros alors le recours n'est plus forfaitaire mais réel, c'est-à-dire que le montant correspond à la réparation intégrale des dommages matériels subis.

Cette convention n'est toutefois pas opposable à l'assuré, c'est-à-dire que l'assuré peut très bien choisir d'être indemnisé par application du droit commun.

2. Les montants forfaitaires depuis 2009 sont présentés en annexes A

3. https ://www.index-assurance.fr/pratique/sinistre/convention-irsa

4. Le barème de la convention IRSA est consultable au lien suivant : https ://fredericlassureur.fr/wp-content/uploads/2015/01/Barême-IRSA.pdf

5. Les montants forfaitaires depuis 2009 sont présentés en annexes A

1.2.2 Projet de réforme de la garantie responsabilité civile automobile

Dans le cadre de la modernisation de la justice commencée en 2016, le ministère de la justice a pour intention de réformer la responsabilité civile automobile. Parmi les mesures proposées dans le projet de réforme présenté en mars 2017 [6], une mesure visant à améliorer l'indemnisation des dommages corporels inquiète tout particulièrement les assureurs.

L'article 1287 de ce projet de réforme prévoit que « *En cas de dommage corporel, la faute de la victime est sans incidence sur son droit à réparation* » et mentionne également que « *Toutefois, la faute inexcusable prive la victime de tout droit à réparation si elle a été la cause exclusive de l'accident* ».

A ce jour, la responsabilité civile automobile de l'assuré est engagée lorsqu'il est lui-même responsable des dommages causés à un tiers. Les dommages corporels qu'il subi dans le cas où celui-ci est fautif ne sont pas indemnisés au titre de cette garantie. La garantie du conducteur qui est quant à elle optionnelle est prévue à cet effet.

Afin d'améliorer l'indemnisation des dommages corporels, ce projet de réforme suggère que la responsabilité civile automobile de la victime non responsable peut alors être engagée afin d'indemniser les préjudices corporels du conducteur fautif (sauf faute inexcusable).

L'intégration de l'indemnisation des conducteurs fautifs dans la responsabilité civile automobile conduirait logiquement à une hausse des cotisations en assurance automobile (estimée à 25% [7]), et ce en raison d'une augmentation de la sinistralité afférente à cette garantie. Dans le cas où ce projet de loi serait adopté, on pourrait également s'interroger sur la définition des risques couverts par la garantie du conducteur.

A la date à laquelle ce mémoire est rédigé (janvier 2019), aucune décision sur l'adoption de ce projet de loi n'a encore été prononcée. Un espace participatif a été ouvert entre mars et avril 2018 pour consulter l'avis du public sur certaines interrogations que peut susciter ce projet de réforme.

6. « *Projet de réforme de la responsabilité civile », Ministère de la justice (Mars 2017)*
7. Argus de l'assurance « RC auto : la réforme qui fait peur aux assureurs » A.Abadie le 02/07/2018.

Chapitre 2

Construction de la base d'étude

Le choix de développer une méthode basée sur des données individuelles est principalement motivé par la volonté d'utiliser un maximum d'informations sur les sinistres pour améliorer la prédiction de leur coût ultime.

Cette partie consiste donc à présenter l'ensemble des données utilisées lors de la réalisation de cette étude ainsi que les différents retraitements effectués.

2.1 La base sinistres

La première étape fut de récupérer l'ensemble des sinistres corporels en responsabilité civile automobile survenus sur une période suffisamment large pour mener à bien cette étude.

Dans cette base, les informations suivantes ont été extraites pour les sinistres survenus entre le 1er janvier 2009 et le 31 décembre 2017 :

- Le marché (Professionnel/Particulier)
- Le numéro de sinistre
- Le numéro de contrat
- Le code de la garantie
- La date de survenance du sinistre
- La date de connaissance société
- La date d'ouverture de la garantie
- La vision au 31/12/2017 de l'état du sinistre (classé/en-cours)
- La vision en fin d'année de l'état du sinistre sur la période 2009-2017
- La vision en fin d'année de la charge dossier/dossier, des montants de règlements, des recours encaissés et évalués sur la période 2009-2017 ainsi qu'à l'ouverture du sinistre
- La vision en fin d'année du service en charge de la gestion du sinistre sur la période 2009-2017, ainsi qu'à l'ouverture du sinistre
- Le taux de responsabilité de l'assuré actualisé
- Le taux de responsabilité de l'assuré dans le cadre de la convention IDA
- Application de la convention IDA
- La nature judiciaire
- Le département où le sinistre est survenu

— Les codes causes sinistre 1 et 2

Chez MMA, la gestion des sinistres corporels est répartie principalement en trois services selon la gravité des préjudices :

— Corporels légers : Les sinistres d'un montant inférieur à 50 000 euros et pour lesquels le taux d'AIPP des victimes ne dépasse pas 10% sera confié au service des sinistres corporels légers.
— Corporels médians : Dès lors qu'une victime est décédée ou que le taux d'AIPP est compris entre 10 et 50% le sinistre est géré par le service des corporels médians.
— Corporels importants : Les sinistres dont l'évaluation est susceptible de dépasser les 1 500 000 euros ou ceux pour lesquels le taux d'AIPP d'une victime est supérieur à 50% sont gérés au service des coporels importants.

Parmi les valeurs prises par les codes causes sinistres 1 et 2, on retrouve notamment les 13 cas d'accident de la circulation répertoriés dans le barème de la convention IRSA.

Toutefois, et ce malgré une période d'observation de neuf années (2009-2017), seulement **6 352 dossiers sinistres corporels** ont pu être récupérés.

2.2 Enrichissement de la base sinistres

Suite à ce constat de faible volumétrie de données, l'idée est d'intégrer les sinistres corporels en responsabilité civile automobile issus du marché des particuliers.

2.2.1 Intégration de la sinistralité issue du marché des particuliers

Cet ajout nous amène toutefois à supposer que les sinistres du périmètre des particuliers et ceux issus du marché des professionnels de l'automobile sont homogènes. En termes de couverture d'assurance, les deux produits ne proposent évidemment pas les mêmes garanties. Néanmoins, les risques liés à la circulation présent dans ces deux couvertures d'assurance restent comparables en termes d'indemnisation en cas de sinistre. En effet, que l'on soit un professionnel de l'automobile ou un particulier, les accidents de la circulation sont les mêmes pour tous.

La couverture de la responsabilité civile automobile du professionnel se résume principalement au cas où il circule avec ses propres véhicules ou bien ceux de l'entreprise (objet de risque « Véhicules de l'entreprise »), lorsqu'il déplace/teste des véhicules qui lui sont confiés dans le cadre de son activité (objet de risque « Véhicules confiés ») et dans le cas où des véhicules en vente sont essayés par de potentiels clients (objet de risque « Véhicules à la vente »). A cela, il faut également ajouter la couverture du risque circulation dans le cas d'un prêt de véhicules de courtoisie.

L'activité du professionnel n'a intrinsèquement pas d'impact sur la gravité d'un sinistre corporel en RC automobile, seule la nature des véhicules assurés peut potentiellement moduler la gravité du sinistre et donc le montant d'indemnisation versé aux victimes. Les véhicules confiés gérés par le professionnel ainsi que les véhicules de courtoisie sont fondamentalement de même nature que les véhicules assurés sur le marché automobile grand public.

Ces différents arguments permettent de justifier d'un point de vue qualitatif l'hypothèse d'homogénéité des montants de sinistres issus de ces deux marchés. Cette hypothèse doit toutefois être vérifiée statistiquement.

2.2.2 Vérification de l'homogénéité du coût des sinistres

Afin de vérifier statistiquement l'homogénéité des montants de sinistres issus du marché des particuliers et des professionnels, un test statistique est réalisé. Afin de s'affranchir des hypothèses sur la distribution des données, les tests non paramétriques (*distribution free*) sont privilégiés.

Parmi eux, le test de Mann - Whitney est vraisemblablement le plus populaire et permet de comparer deux échantillons en utilisant la notion de rangs. L'idée sous-jacente est de ne plus s'intéresser aux valeurs en elles-mêmes mais plutôt à leur numéro d'ordre dans l'échantillon global. Ce test revient dans le cas présent à tester si le coût des sinistres provenant du marché des professionnels de l'automobile ne sont pas majoritairement plus faibles ou plus élevés que ceux du marché des particuliers (hypothèse nulle du test).

Pour ne pas alourdir le corps de ce mémoire, le test de Mann - Whitney est présenté succinctement en annexe B de ce mémoire.

Avant de procéder au test de Mann - Whitney, la densité du coût des sinistres par marché est présentée sur la Figure 2.1.

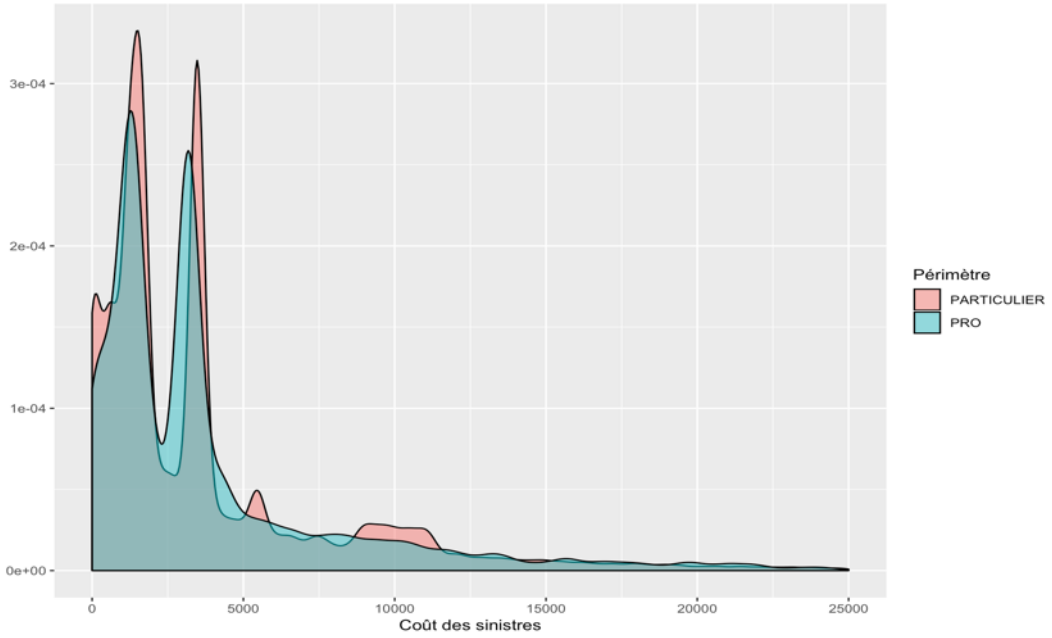


FIGURE 2.1 – Densité du coût des sinistres bornés entre 0 et le 95ème centile en fonction du périmètre

Les deux densités sont globalement très proches hormis deux légers décalages visibles autour de 3 000 et 10 000 euros. Ces derniers doivent être mis au regard de la différence de volumétrie de sinistres RCC sur ces deux marchés : **96 397** sinistres pour les particuliers contre **6 352** sinistres pour les professionnels de l'automobile.

Afin de confirmer cette analyse graphique, le test bilatéral de Mann - Whitney est réalisé à l'aide du logiciel R à partir de la fonction « *wilcox.test* » dont les sorties sont présentées en Figure 2.2. Le seuil critique (noté α) est ici fixé à 5%.

```
        Wilcoxon rank sum test with continuity correction

data:  data[marche == "PRO", ]$cout and data[marche == "PARTICULIER", ]$cout
W = 188850000, p-value = 0.34
alternative hypothesis: true location shift is not equal to 0
```

FIGURE 2.2 – Résultats du test de Mann - Withney

La p-valeur du test de Mann - Withney obtenue est de **0,34**, elle est supérieure au seuil critique choisi (α = 5%). En conséquence, l'hypothèse nulle selon laquelle les deux sous-échantillons proviennent d'une même distribution (i.e que les montants de sinistres sont homogènes) n'est pas rejetée pour un seuil α de 5 %. Ces résultats viennent conforter les arguments précédemment énoncés.

Les sinistres issus du marché des particuliers sont donc intégrés à notre base sinistres initiale. Celle-ci recense dorénavant **102 749** sinistres corporels en responsabilité civile automobile survenus sur la période 2009-2017. L'apport de la sinistralité issue du marché des particuliers est très important puisque ces derniers représentent plus de 93% de la base d'étude.

2.3 La base comptable

Les extractions réalisées jusqu'à présent ont permis de récupérer la vision en fin d'année des sinistres sur la période 2009-2017. La base comptable recense quant à elle l'ensemble des règlements et recours effectués par les gestionnaires sinistres. En plus des montants et dates de règlements ou de recours, cette base donne également des informations sur les motifs et la nature des règlements.

Les informations disponibles par règlement sont ensuite agrégées afin de conserver une base de données au niveau sinistre, c'est-à-dire qu'une ligne de cette base correspond à un unique sinistre.

Les données suivantes ont été extraites afin de compléter les informations déjà récupérées dans la base sinistres :

- Date du premier règlement ou recours
- Application de la convention IRCA
- Règlements matériels uniquement
- Présence d'au moins un règlement matériel
- Nombre total de règlements versés et de recours encaissés

Grâce à cette base, il est possible de récupérer l'information sur l'application de la convention IRCA lors du règlement du sinistre (cette information n'étant pas présente dans la base sinistres).

Etant donné que cette information est disponible au niveau règlement (ou recours), on considère qu'un sinistre est concerné par la convention IRCA dès lors qu'au moins un de ses règlements a été versé au titre de cette convention.

La date du premier règlement nous permet d'analyser la liquidation du sinistre après ouverture de la garantie dans le système de gestion.

Enfin, bien qu'ils soient référencés comme corporels, il semble pertinent de distinguer les sinistres pour lesquels seuls des règlements matériels ont été versés. En effet, par précaution, dès lors qu'un assuré mentionne des dommages corporels lors de la déclaration d'un sinistre, il sera référencé comme corporel. Après réception d'informations complémentaires, il est tout à fait possible que seuls des préjudices matériels ont été réellement constatés et dans ce cas, le sinistre reste en gestion auprès des services corporels.

Par ailleurs, la vision au niveau règlement sera utilisée lors de la mise en « *as-if* » des coûts de sinistres en utilisant la date de versement des règlements et des recours.

2.4 La base victimes

La troisième et dernière base de données utilisée recense des informations au niveau victime (une ligne de cette base correspond à une victime) pour un sinistre donné. Elle ne possède cependant pas d'historique, seules les dernières informations connues concernant les victimes y sont recensées, c'est-à-dire vues au 31/12/2017 dans notre cas.

La base victimes est tenue par les gestionnaires sinistres et recense de nombreuses informations sur les victimes pour les sinistres survenus à partir de 2012. Pour les sinistres survenus avant 2012 (soit 33% des sinistres présents dans la base), très peu d'informations sur les victimes ont pu être récupérées.

Pour près de la moitié des sinistres survenus après 2012, aucune information victimes n'a pu être récupérée. Plusieurs raisons peuvent justifier la présence de valeurs manquantes pour ces sinistres. Tout d'abord, il est possible qu'aucune victime n'ait subi de dommages corporels et dans ce cas, aucune information ne sera renseignée dans la base victimes. Il est également possible que les victimes ne soient pas encore connues par l'organisme assureur. Ce dernier cas concerne principalement les sinistres relativement récents.

Malgré la grande proportion de valeurs manquantes, il a été décidé dans un premier temps de conserver ces informations en raison du pouvoir discriminant potentiellement important qu'elles peuvent apporter pour expliquer les différents comportements de liquidation de sinistres.

Pour chacune des victimes, les informations suivantes ont pu, le cas échéant, être récupérées :

- La nature du préjudice (corporel ou mortel)
- La qualité de la victime (Assuré ou Tiers)
- Le taux d'AIPP de la victime
- L'âge de la victime au moment de la survenance du sinistre

— Le nombre d'ayant-droit
— Le rôle de la victime (piéton/cycliste/conducteur ou passager)
— Le type du véhicule (deux roues ou quatre roues)

Les informations recensées dans cette base peuvent concerner d'une part les préjudices subis par des tiers mais également ceux subis par les assurés. La garantie engagée n'étant pas renseignée, les victimes concernées par la responsabilité civile automobile des assurés ne peuvent pas être récupérés par ce biais.

Néanmoins, les informations issues de la variable « Qualité de la victime » peuvent nous aider à faire cette distinction. Nous conservons alors seulement les informations relatives aux tiers.

Cependant, pour certains dossiers sinistres présents dans notre base d'étude, seules les préjudices subis par les assurés sont renseignés dans la base victimes. Pour ce cas particulier, il a été décidé de tout de même conserver ces informations puisqu'elles peuvent nous apporter des indications sur la gravité des sinistres.

Pour un même sinistre, il est naturellement possible d'observer plusieurs victimes. Comme pour la base comptable, les informations disponibles pour chacune des victimes sont agrégées afin de conserver une base de données au niveau sinistre (i.e une ligne pour un sinistre).

Les informations associées aux victimes décédées ont été privilégiées à celles concernant les préjudices corporels uniquement, et ce en raison d'un coût moyen observé plus élevé (122 711 vs 21 803 euros).

En suivant le même principe pour les autres informations, les aménagements suivants ont été réalisés :
— Le taux d'AIPP retenu correspond au maximum des taux d'AIPP de toutes les victimes d'un même sinistre.
— Par simplification, l'âge retenu correspond à la moyenne arithmétique des âges des victimes au moment de la survenance du sinistre.
— L'ordre de priorité suivant a été retenu pour l'alimentation de la qualité de la victime : Piéton -> Cycliste -> Conducteur ou passager.
— Si un deux roues est impliqué dans l'accident, ce type de véhicule est retenu au niveau sinistre. Dans le cas inverse, ce champ est alimenté à quatre roues.

A ce stade, la base d'étude recense **102 749 sinistres** en responsabilité civile corporelle automobile. L'utilisation de la base victimes a permis d'enrichir la base d'étude pour un peu plus de **40%** des sinistres représentant un apport non négligeable d'informations.

2.5 Retraitement de la base d'étude

Le retraitement des données est une étape indispensable dans un travail de modélisation. En effet, la qualité des données mises en entrée des modèles conditionne naturellement la fiabilité des résultats obtenus. Le plus grand soin y a donc été apporté afin d'utiliser une base de données la plus propre possible.

Le premier retraitement opéré sur la base concerne les **24 743 sinistres classés sans suite** (soit **24%** de la base). Au sein de la Direction Centrale Entreprise, un sinistre considéré comme sans suite est un sinistre classé dont la dernière évaluation du gestionnaire sinistre ainsi que les derniers montants de règlements et de recours sont nuls. La définition de la Fédération Française de l'Assurance (FFA) inclue également les sinistres pour lesquels seuls des règlements accessoires ont été versés (majoritairement des frais d'expertise). Cette nuance n'a que très peu d'impact ici puisque seuls 166 sinistres sont concernés par ce cas particulier.

L'analyse des sinistres classés ne présente pas d'intérêt particulier dans le cadre de cette étude, ils sont donc supprimés de la base. A ce stade, notre base recense **78 006** sinistres en RC corporelle automobile.

Certains tests de cohérence des données provenant de la base sinistres ont ensuite été mis en place afin de détecter les éventuelles anomalies. Après avoir vérifié que pour chaque sinistre classé, une date de classement est bien renseignée et que celle-ci n'est pas antérieure à la date de survenance du sinistre, une analyse sur les montants cumulés de règlements et de recours a été effectuée. En partant du principe que ces montants ne peuvent pas être négatifs, **145 sinistres** ont été supprimés de notre base.

Une analyse de cohérence des montants de sinistres est réalisée entre la base sinistres et la base comptable. Les coûts des sinistres observés dans ces deux bases doivent être en théorie identiques.

Rappelons que le coût d'un sinistre vu au 31/12/2017 est obtenu de la façon suivante :

$$Cout\ sinistre_{31/12/2017} = Règlements\ totaux_{31/12/2017} - Recours\ encaissés_{31/12/2017}$$

Les règlements totaux sont obtenus en sommant les règlements principaux avec les règlements accessoires.

Toutefois, certains traitements comme les procédures de clôture rapide des comptes en fin d'exercice (« *fast close* ») peuvent entraîner des visions différentes du coût des sinistres entre les deux bases. De plus, certains sinistres ouverts au mois de décembre 2017 ne sont pas présents dans la base comptable.

En prenant la base sinistres comme référence, **383 sinistres** se retrouvent en anomalie, les coûts de ces derniers doivent donc être corrigés.

Pour ces sinistres, l'écart de coût observé entre la base sinistres et la base comptable est corrigé par la création d'un flux comptable (un règlement ou un recours selon le signe de l'écart de coût observé) d'un montant correspondant à la valeur de cet écart qui est versé à la date de clôture du sinistre. Pour les sinistres en-cours, on considèrera que ce flux est versé à la date de vision des données, c'est-à-dire au 31/12/2017.

Au-delà de la cohérence des bases utilisées pour notre étude, cette correction est nécessaire pour effectuer la revalorisation des sinistres au niveau règlement dont la méthodologie sera décrite infra.
Après cette phase de retraitement des données, notre base d'étude recense dorénavant **77 861 sinistres** corporels en RC automobile.

2.6 *Text Mining* sur les circonstances sinistres

Parmi les différents éléments renseignés par les gestionnaires sinistres à l'ouverture d'un sinistre, ils ont la possibilité de résumer succinctement les circonstances du sinistre. Ils sont libres quant à la rédaction de celles-ci, le champ n'est pas formaté comme peuvent l'être par exemple les codes causes sinistres 1 et 2 récupérés dans la base sinistres. Il n'est pas possible d'exploiter ces circonstances sinistres ou de les analyser en l'état pour en tirer de la connaissance.

L'objectif ici est donc d'extraire de l'information statistique à partir de ces circonstances sinistres, c'est-à-dire d'éléments textuels qui sont par essence non structurés.

A noter que les circonstances d'un sinistre sont uniquement renseignées à l'ouverture du sinistre, elles ne sont pas mises à jour par le gestionnaire dès lors qu'il a connaissance de nouvelles informations.

Une étude de *Text Mining* (fouille de texte en français) est réalisée sur les circonstances de l'ensemble des sinistres présents dans la base d'étude pour en extraire un maximum d'informations pouvant s'avérer utile pour estimer leur charge ultime. Cette technique permet de mettre au jour de manière automatisée l'information utile initialement cachée dans des masses importantes de données. L'intérêt ici n'est pas de détailler avec précision la démarche entreprise pour réaliser une étude de ce type mais d'en présenter les principales étapes ainsi que les résultats obtenus. Le lecteur intéressé par la réalisation d'une étude de *Text Mining* à l'aide du logiciel R peut s'appuyer sur le tutoriel proposé par EducTechWiki.fr *« Tutoriel tm text mining package »*.

Bien que les éléments textuels récupérés par les assureurs contiennent bon nombre d'informations intéressantes à analyser, il faut tout de même avoir à l'esprit qu'une étude de *Text Mining* s'avère très couteuse en temps. En effet, elle nécessite encore à ce jour une intervention humaine importante que ce soit pour le nettoyage des données afin de leur donner un sens métier ou bien dans l'analyse et l'interprétation des résultats.

L'ensemble de ce travail est réalisé à l'aide d'une formation dispensée à MMA en utilisant principalement le package « *tm* » [1] du logiciel R.

2.6.1 Prétaitements des données et normalisation du texte

La première étape consiste à récupérer les circonstances de l'ensemble des sinistres présents dans la base d'étude. Cependant, les circonstances sinistres ne sont pas recensées dans une base de données, elles sont uniquement consultables sur un outil dédié à la gestion des sinistres. Une demande a donc été formulée auprès du département informatique pour récupérer les circonstances sinistres en format texte.

Pour près de **40%** des sinistres présents dans notre base d'étude, les circonstances sinistres ne sont pas renseignées par le gestionnaire. Aucune information supplémentaire ne pourra donc être extraite pour ces sinistres à partir de leurs circonstances sinistres.

1. https ://cran.r-project.org/web/packages/tm/tm.pdf

Après import de ces données sous le logiciel R, les circonstances sinistres sont ensuite découpées en mots pour effectuer une première analyse de leur occurrence.

On observe alors le phénomène de « *long tail* » (longue traîne en anglais) représenté sur la Figure 2.3 :

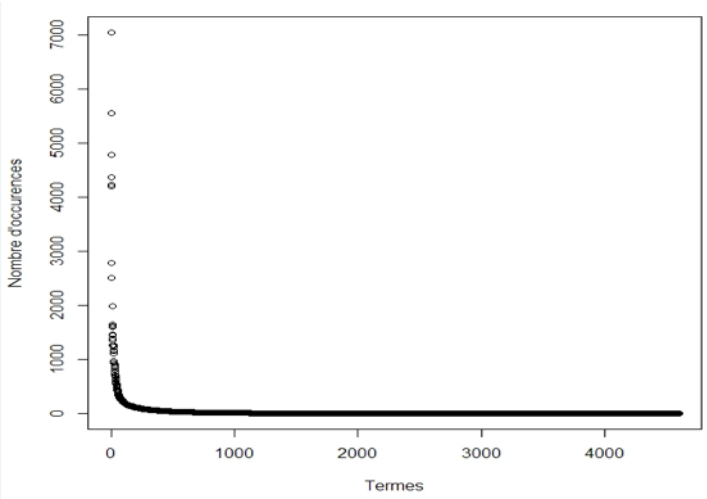


FIGURE 2.3 – Représentation du phénomène de long tail

Ce phénomène correspond au cas où une minorité des termes (i.e les plus courants de la langue française) représentent une part importante du nombre total d'occurrences alors qu'un très grand nombre de termes n'apparaissent que très rarement.

L'étape de prétraitements des données est nécessaire pour d'une part réduire le nombre total de mots analysés mais aussi pour augmenter les fréquences associées aux mots de la même famille grammaticale mais déclinés sous différentes formes (exemple : assurés, assuré, assurée ...). Les techniques de normalisation des données vont permettre de rassembler différents mots ayant la même signification pour ne pas les compter comme deux formes différentes.

Tout d'abord, les majuscules, les accents, la ponctuation, les chiffres sont supprimés des formes à analyser. De plus, certains mots n'apportent pas d'informations utiles dans l'analyse du contenu des circonstances sinistres (*stepwords*, ou littéralement en français « mots vides »). On retrouve par exemple les pronoms, les prépositions, les conjonctions mais également les mots non pertinents vis-à-vis de l'étude réalisée (exemple : le nom des compagnies d'assurance).

Une autre problématique importante dans une étude de *Text Mining* concerne les mots fléchis. Ces mots correspondent aux formes conjuguées des verbes et différents accords, aux adjectifs et aux noms. La présence de ce type de mots provoque l'apparition de termes sous différentes formes mais qui possèdent néanmoins un sens proche, ils peuvent donc être regroupés au sein d'une unique forme graphique commune.
Deux techniques sont à ce jour utilisées pour effectuer cette tâche : la **lemmatisation** et la **racinisation**.

La lemmatisation consiste à transformer les formes fléchies en leur lemme (exemple : ramener à l'infinitif

un verbe conjugué) alors que la racinisation a pour objectif de les transformer en leur racine, c'est-à-dire la partie du mot restante lorsque le suffixe a été supprimé. Compte tenu de la spécificité de langue française, les algorithmes de lemmatisation se révèlent être plus adaptés. Néanmoins, l'application de ce type d'algorithme nécessite que les termes présents au sein du texte à analyser soient correctement orthographiés. Etant donné que les gestionnaires sinistres sont limités en nombre de caractères dans la rédaction des circonstances sinistres, le texte à analyser contient beaucoup d'abréviations qu'il faut au préalable identifier et traiter.

Au-delà de leur ressemblance d'un point de vue grammaticale, certains termes peuvent également être proches en termes de sens sans pourtant partager le même lemme. Il faut là aussi procéder à quelques regroupements de termes pour en simplifier l'analyse. A titre d'exemple, le mot « piéton » regroupe d'une part l'ensemble des termes partageant le même lemme mais également les termes suivants : « passage piéton », « passage protégé », « passage clouté ».

Après ces étapes de prétraitements des éléments textuels, les formes peuvent être représentées au travers d'une matrice d'occurrences, couramment appelée « matrices termes-documents » ou « matrices documents-termes » selon le sens de représentation. Les documents représentent ici les 77 861 circonstances sinistres à analyser.

Naturellement, un même terme peut apparaître plusieurs fois au sein d'une même circonstance sinistre, le nombre d'occurrences d'un mot donné ne correspond donc pas forcément au nombre de documents dans lesquels le mot est renseigné mais bien au nombre total de fois où ce terme apparaît dans l'ensemble des circonstances sinistres.

A ce stade, la matrice termes-documents construite possède 13 666 termes pour 77 861 circonstances sinistres.

Les termes très peu fréquents peuvent être supprimés en utilisant le paramètre de **sparsité** de la matrice documents-termes pour en simplifier l'analyse. Le terme est conservé s'il apparaît dans au moins $100 \times (1 - sparse)\%$ des circonstances sinistres. Dans notre cas, ce terme est fixé à 99,987% de sorte à ce que les termes présents dans moins de 100 circonstances sinistres soient supprimés.

La taille de la matrice documents-termes est ainsi réduite, l'analyse se restreint dorénavant à **228 termes**.

2.6.2 Analyse des résultats du *Text Mining*

Les résultats du Text Mining peuvent être représentés à l'aide d'un nuage de mot (« *wordcloud* ») pour effectuer une première analyse des mots les plus renseignés au sein des circonstances sinistres. Pour faciliter la lecture graphique du nuage de mots, seuls les 30 mots les plus fréquemment cités sont représentés. La taille des mots est proportionnelle à son nombre d'occurrence au sein des circonstances sinistres, leur position sur le graphique est quant à elle aléatoire dans le cas présent.

FIGURE 2.4 – Nuage de mots

Les gestionnaires sinistres renseignent très fréquemment les mots « heurter », « percuter » (regroupés en un unique terme « heurter_percuter ») ainsi que « responsable » avec respectivement 15 001 et 11 067 occurrences dans les circonstances sinistres. Le mot « responsable » est à analyser avec précaution puisqu'il peut très bien concerner l'assuré ou bien l'adversaire. Ces mots n'ont en soi pas d'intérêt pour expliquer le comportement de liquidation des sinistres, il faut donc repérer les mots moins occurrents pouvant apporter de l'information. Les termes « pieton », « cycliste_moto », « alcool_drogue » et « sens_inverse » apparaissent respectivement 4 782, 4193 et 1 261 fois et peuvent potentiellement être intéressant pour notre étude, reste à savoir, s'ils peuvent être utiles pour expliquer le coût des sinistres.

Pour ce faire, il est possible de créer un nuage de mots comparatif sous R à l'aide du package « *wordcloud* » [2] et de la fonction *comparison.cloud* qui permet de représenter sous forme graphique la relation entre l'occurrence des termes et une variable illustrative catégorielle.

Dans notre cas, il peut être intéressant d'utiliser cette fonction afin de constater une éventuelle relation entre les termes présents dans les circonstances sinistres et le coût de ces derniers. Afin de ne pas sous-estimer le coût des sinistres en-cours en retenant uniquement les montants réglés, les sinistres présents dans notre base sont répartis en fonction de leur charge dossier/dossier.

Les cinq modalités de notre nouvelle variable catégorielle sont les suivantes :

— Charge dossier/dossier inférieure à 2 000 euros
— Charge dossier/dossier comprise entre 2 000 et 10 000 euros
— Charge dossier/dossier comprise entre 10 000 et 50 000 euros
— Charge dossier/dossier comprise entre 50 000 et 150 000 euros
— Charge dossier/dossier supérieure à 150 000 euros

2. https ://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

Le montant de 150 000 euros correspond au seuil de graves utilisé par la Direction Centrale Entreprise de MMA. Les autres montants représentent différents niveaux de gravité des sinistres qui ont été choisis arbitrairement.

Sur la Figure 2.5, chaque modalité est associée à une couleur distincte. Par exemple, les termes associés à la modalité « Plus de 150 000 » sont représentés en vert.
La taille d'un mot est proportionnelle à l'écart entre son taux d'occurrence sur la modalité et son taux d'occurrence moyen sur l'ensemble des modalités. La taille des mots sera par conséquent plus grande dès lors que le taux d'occurrence sur une modalité donnée est bien plus important que sur les autres modalités.

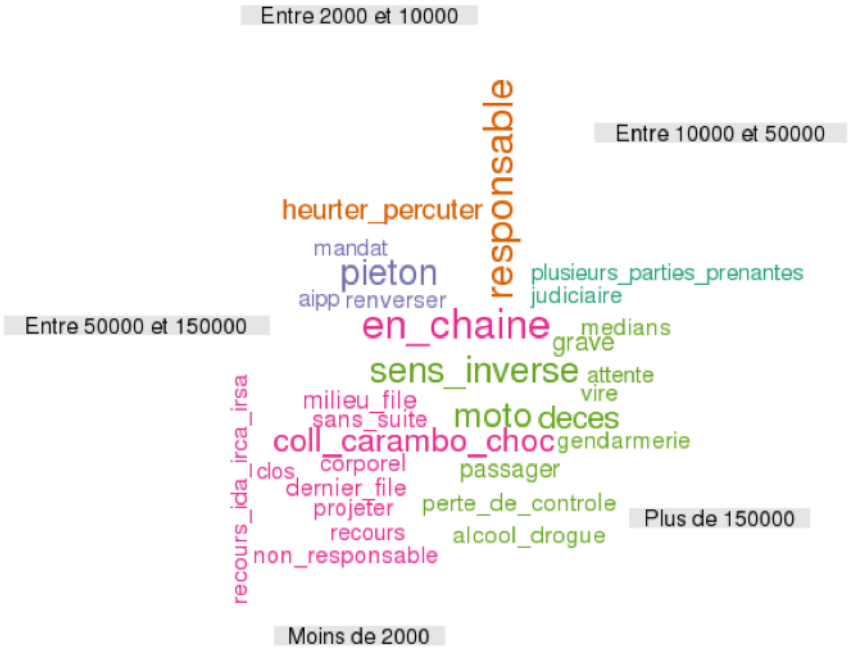


FIGURE 2.5 – Nuage de mots comparatif

Les termes « sens_inverse », « moto », « décès », « alcool_drogue », « perte_de_contrôle » et « grave » sont plus fréquemment associés aux sinistres évalués à plus de 150 000 euros qu'aux autres. Le terme « grave » regroupe, en autre, les mots « violent », « traumatisme crânien », « polytraumatisme » et « hôpital ». Ce constat peut toutefois être nuancé en raison d'une part du faible nombre de sinistre présent dans cette modalité (1 026 sinistres) mais aussi de la relative faible occurrence dans les circonstances sinistres de certains termes comme « décès » ou « grave » qui n'apparaissent respectivement que 522 et 264 fois.

En ce qui concerne les sinistres dont la charge est comprise entre 50 000 et 150 000 euros, les mots « pieton » et « renverser » sont plus souvent associés à cette modalité qu'aux autres. Les termes « plusieurs_parties_prenantes » et « judiciaire » sont fréquemment associés aux sinistres d'un montant compris entre 10 000 et 50 000 euros. L'enjeu concernant les sinistres dont le montant évalué est inférieur à 10 000 est moindre, les mots renseignés dans les circonstances sinistres de ces derniers ne seront donc récupérés. Même si le terme « cycliste » n'apparaît pas, il sera également retenu en vue de compléter les valeurs non renseignées de la variable « Qualité de la victime » provenant de la base victimes.

A ce stade, neuf termes ont été retenus pour d'une part compléter (si possible) les valeurs manquantes des variables présentes dans la base victimes mais aussi pour ajouter de nouvelles informations dans notre base d'étude.

Ces termes ainsi que le nombre de circonstances dans lesquelles ils apparaissent sont présentés dans le Tableau 2.1.

Termes	décès	piéton	cycliste	moto	sens inverse	alcool drogue	perte de contrôle	grave	judiciaire
Nombre d'apparitions	522	3 756	2 535	1 326	3 361	1 157	1 973	264	630

TABLE 2.1 – Termes extraits des circonstances sinistres et leurs occurences

Reste à savoir maintenant sous quelle forme ces informations peuvent être intégrées dans la base d'étude.

2.6.3 Intégration des informations issues du *Text Mining* dans la base d'étude

Les neuf termes extraits des circonstances sinistres sont directement intégrés dans la base d'étude en créant neuf variables binaires (1 si le terme est présent dans les circonstances du sinistre, 0 sinon).

D'autre part, certaines de ces informations peuvent être utilisées à des fins de remplacements de valeurs manquantes concernant la nature du préjudice (variable « décès ») et le rôle de la victime (« cycliste » et « piéton »).

Avant de compléter les informations issues de la base victimes, il convient tout d'abord de vérifier si les termes récupérés à partir de l'analyse des circonstances sinistres apportent une information de qualité qui pourra potentiellement être utilisée lors de l'étape de modélisation. Ces informations sont donc confrontées avec les éléments connus issus de la base victimes, lorsque ces dernières sont connues.

Cependant, près de 40% des circonstances des sinistres présents dans notre base ne sont pas renseignées par les gestionnaires, les informations manquantes ne pourront donc pas être compléter en totalité à partir des résultats du *Text Mining*.

Pour au moins **20 171** sinistres (soit **26%** de la base), aucune information n'est connue au niveau victime, que ce soit par le biais de l'analyse des circonstances sinistres ou bien par la base victimes.

Toutefois, la base comptable peut être utilisée à des fins de remplacements de valeurs manquantes concernant la nature du préjudice subi.

Nature du préjudice subi

En partant du principe que dans notre cas, si l'information concernant le décès d'une victime est absente dans les circonstances sinistres alors le préjudice est en théorie corporel. Cependant, pour certains sinistres, seuls des règlements au titre de dommages matériels ont été versés bien qu'ils soient référencés comme corporels.

Etant donné que la nature des règlements est renseignée au niveau garantie, elle est par conséquent plus précise que celle concernant la nature préjudice (information au niveau sinistre). De ce fait, il semble opportun de qualifier le préjudice de matériel lorsque seuls des règlements matériels ont été versés. De plus, contrairement aux circonstances sinistres, la nature des règlements versés est une information connue pour tous les sinistres présents dans la base d'étude. Il convient néanmoins de s'assurer que ce remplacement est cohérent d'un point de vue statistique en comparant les montants de sinistres observés.

Toutefois, l'information sur la nature des règlements d'un sinistre donné n'est complète que lorsque ce dernier est clôturé. En effet, pour un sinistre en-cours, cette information est dite censurée à droite dans le sens où la nature des règlements futurs est inconnue au moment où ce sinistre est analysé. Il est tout à fait possible que pour un sinistre en-cours donné, seuls des règlements au titre de dommages matériels ont été constatés au 31/12/2017 (date de vision des données) mais que l'indemnisation des préjudices corporels intervienne seulement l'année d'après.

Dans ce cas, il est jugé préférable de conserver l'information sur la nature du préjudice issue de la base victimes lorsque celle-ci est renseignée plutôt que de la remplacer par une information incomplète. De ce fait, l'état du sinistre est à prendre en compte pour alimenter la nature du préjudice, qu'elle soit renseignée ou non.

Remplacement des valeurs manquantes

Tout d'abord, la nature du préjudice n'est pas renseignée pour **47 018** sinistres. Dans le but de remplacer les natures de préjudice manquantes, l'information relative à la nature des règlements (base comptable) est croisée avec les résultats issus du *Text Mining* (variable « décès ») et l'état du sinistre.

SINISTRES CLOS			
Nature des règlements	***Décès – Text mining***	***Nombre de sinistres***	***Coût moyen***
Aucun règlement	Non renseignée	16	(727)
	Non	19	(1 167)
	Oui	1	(700)
Règlements matériels uniquement	Non renseignée	9 407	1 887
	Non	10 512	2 029
	Oui	30	5 067
Indemnisation corporelle	Non renseignée	9 560	5 425
	Non	14 657	8 509
	Oui	141	57 627
Total		**44 343**	**5 049**

TABLE 2.2 – Coût des sinistres clos pour lesquels la nature du préjudice n'est pas renseignée en fonction de la nature des règlements et de l'information de décès issue des circonstances sinistres

SINISTRES EN-COURS			
Nature des règlements	*Décès - Text mining*	*Nombre de sinistres*	*Coût moyen évalué*
Aucun règlement	Non renseignée	785	5 416
	Non	893	6 948
	Oui	26	111 227
Règlements matériels uniquement	Non renseignée	340	4 051
	Non	451	4 720
	Oui	-	-
Indemnisation corporelle	Non renseignée	63	23 700
	Non	111	517 518
	Oui	6	56 930
Total		2 675	28 461

TABLE 2.3 – Charge dossier/dossier des sinistres en-cours pour lesquels la nature du préjudice n'est pas renseignée en fonction de la nature des règlements et de l'information de décès issue des circonstances sinistres

Au vu des montants présentés sur le Tableau 2.2 et 2.3, la nature du préjudice est alimentée à l'aide de la nature des règlements lorsqu'aucune information de décès n'est extraite du *Text Mining* et ce quel que soit l'état des sinistres.

Lorsque des informations relatives à un décès sont récupérées à partir des circonstances sinistres, le préjudice subi est jugé comme mortel, sous réserve que ce remplacement n'engendre pas de biais important.

Afin de vérifier cela, nous confrontons sur le Tableau 2.4 la nature du préjudice avec l'information de décès (*Text Mining*) lorsque cette dernière est connue, c'est-à-dire pour 47 400 sinistres.

		Décès - Text Mining		
		Non	*Oui*	*Total en ligne*
Nature du préjudice (base victimes)	**Corporel**	20 073	143	20 216
	Décès	162	175	337
	Non renseignée	26 643	204	26 847
	Total en colonne	46 878	522	47 400

TABLE 2.4 – Nature du préjudice VS Décès - *Text Mining*

Dans le cas où l'information de décès est présente dans la base victimes, elle l'est également dans 175 circonstances sinistres (soit 52% des cas). Ceci s'explique principalement par le fait que les circonstances sinistres sont renseignées à l'ouverture du sinistre par le gestionnaire. Il est tout à fait probable qu'aucune information concernant le décès d'une victime ne soit connue à cette date, expliquant son absence dans les circonstances du sinistre.

Enfin, afin de juger de la cohérence de ces remplacements, il convient d'étudier le taux d'erreur lorsqu'un préjudice mortel est renseigné dans les circonstances du sinistre.

Tout d'abord, dans 45% des cas (soit 143 circonstances sinistres), l'information de décès est présente dans les circonstances sinistres mais absente dans la base victimes.

Bien que les différents indicateurs calculés ne confortent pas totalement l'idée d'effectuer le remplacement des valeurs manquantes à partir des résultats du *Text Mining*, il est tout de même très largement justifié par l'étude des charges dossier/dossier moyennes. Etant donné que notre objectif est d'expliquer la liquidation des sinistres à l'aide d'informations propres aux sinistres, il a été décidé de qualifier la nature du préjudice comme mortel lorsque l'information de décès est présente dans les circonstances du sinistre.

In fine, la nature du préjudice est dorénavant renseignée pour **45 340 des 47 018 sinistres** pour lesquels cette information n'était initialement pas connue, soit en utilisant les données issues de la base comptable ou des circonstances sinistres. Pour les 1 678 sinistres en-cours pour lesquels aucun règlement n'a été versé, cette information reste manquante dans notre base d'étude.

Reste à savoir maintenant s'il paraît opportun d'améliorer l'information statistique apportée par la nature du préjudice en utilisant la nature des règlements.

Modification de l'alimentation de la nature du préjudice

Après s'être intéressé au cas où la nature du préjudice n'était pas renseignée, nous pouvons affiner l'alimentation de celle-ci en la croisant avec la nature des règlements versés.

Pour les sinistres clôturés, la nature des règlements versés apportent naturellement une information plus précise que celle provenant de la base victimes. En conséquence, il semble pertinent de modifier la nature du préjudice lorsque ce dernier est renseigné comme corporel et que seuls des règlements matériels ont été versés.
Cette intuition doit tout de même être confirmée d'un point de vue statistique, les coûts moyens évalués des sinistres clos en fonction de la nature des règlements et de la nature du préjudice lorsqu'elle est renseignée sont présentés dans le Tableau 2.5.

SINISTRES CLOS			
Nature des règlements	***Nature du préjudice***	***Nombre de sinistres***	***Coût moyen évalué***
Aucun règlement	Corporel	5	(627)
	Mortel	-	-
Règlements matériels uniquement	Corporel	5 146	2 102
	Mortel	17	11 074
Indemnisation corporelle	Corporel	15 831	11 915
	Mortel	154	165 198
Total		**21 153**	**10 640**

TABLE 2.5 – Coût des sinistres clos en fonction de la nature des règlements et de la nature du préjudice (renseignée)

Le coût moyen des sinistres clos dont le préjudice est corporel et pour lesquels seuls des règlements matériels ont été effectués est très largement inférieur à celui observé pour les sinistres corporels (2 102 contre 11 915 euros). Dans le cas où le préjudice est mortel, le coût moyen évalué est nettement supérieur, s'élevant à 11 074 euros.

De ce fait, lorsqu'un décès est renseigné dans la base victimes, la nature du préjudice n'est pas modifiée, elle reste alimentée à « mortel ». Au vu de ces éléments, la nature du préjudice des sinistres clos est référencée comme matérielle lorsque seuls des règlements matériels ont été versés.

Pour les sinistres en-cours au 31/12/2017, lorsque la nature du préjudice n'est pas renseignée, elle est alimentée à l'aide de la nature des règlements (le cas échéant), même si cette information est censurée.

La Figure 2.6 résume l'ensemble des retraitements effectués sur l'alimentation de la nature du préjudice.

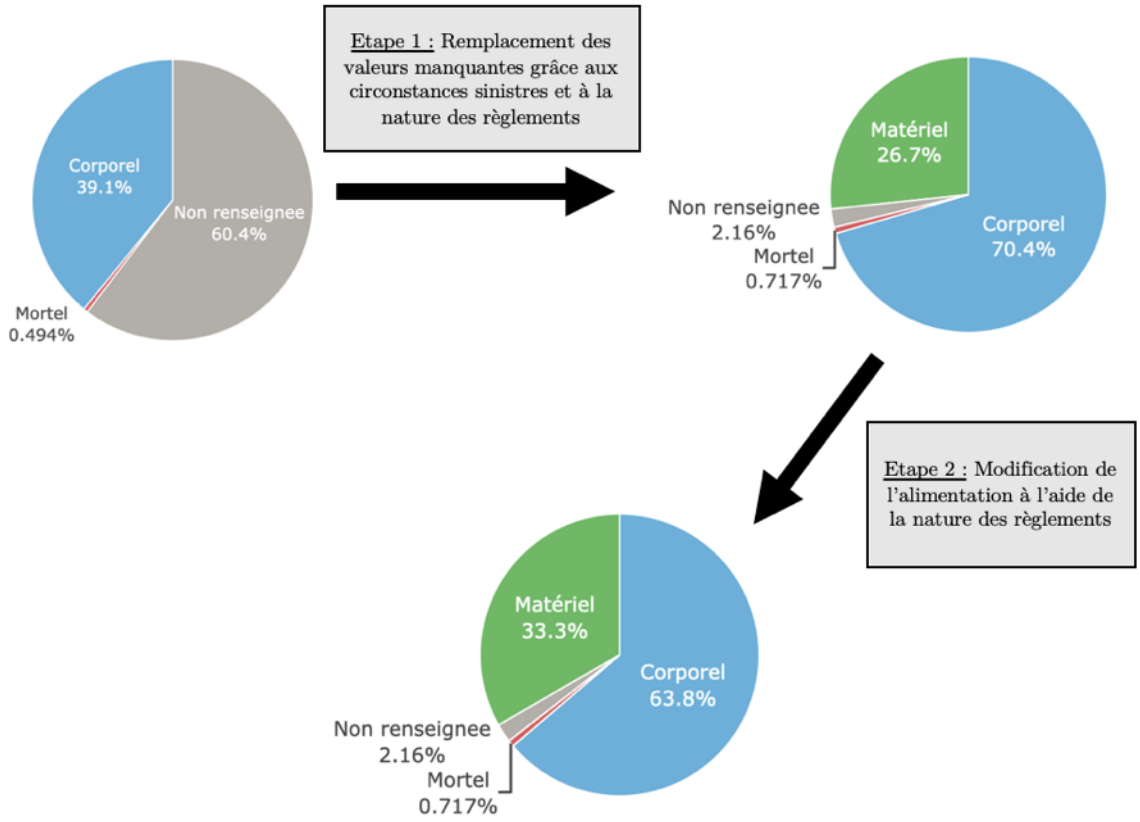


FIGURE 2.6 – Synthèse des différents retraitements réalisés sur la nature du préjudice

A noter qu'après ce retraitement, **un tiers** des sinistres référencés comme corporels ne présentent que des règlements au titre de préjudices matériels.

Le rôle de la victime

Les informations extraites des circonstances sinistres permettent également de renseigner le rôle de la victime lorsque celui-ci est manquant.

		Piéton - Text Mining		
		Non	Oui	Total en ligne
Rôle de la victime (base victimes)	Autres cas	344	141	485
	Cycliste	1258	34	1 292
	Piéton	370	1872	2 242
	Passager ou Conducteur	16 481	53	16 534
	Non renseigné	25 191	1 656	26 847
	Total en colonne	43 644	3 756	47 400

		Cycliste - Text Mining		
		Non	Oui	Total en ligne
Rôle de la victime (base victimes)	Autres cas	333	152	485
	Cycliste	168	1 124	1 292
	Piéton	2 233	9	2 242
	Passager ou Conducteur	16 504	30	16 534
	Non renseigné	25 627	1 220	26 847
	Total en colonne	44 865	2 535	47 400

TABLE 2.6 – Rôle de la victime VS Piéton - Text Mining (à gauche) | Cycliste - Text Mining (à droite)

Contrairement à ce qui a pu être observé concernant la nature du préjudice, les données issues du *Text Mining* semblent cohérentes avec celles présentes dans la base victimes. En effet, lorsque l'information concernant la présence d'un piéton est présente dans la base victimes, elle est renseignée dans 1 872 des circonstances sinistres (soit 83% des cas). Pour les cyclistes, ce même taux s'élève à 87% (soit 1 124 circonstances sinistres) .

Afin de juger de la pertinence du remplacement des valeurs manquantes par les informations contenues dans les circonstances du sinistre, il convient, comme pour l'information relative au décès d'une victime, de vérifier que l'erreur commise n'est pas trop importante.

Lorsque l'information est renseignée dans les circonstances du sinistre, elle l'est également dans plus de 89% des cas pour les piétons et 85% des cas pour les cyclistes.

Au vu de ces indicateurs, les valeurs manquantes quant au rôle de la victime sont donc remplacées par les informations récupérées à partir des circonstances sinistres. Même après ces remplacements, cette information n'est pas disponible pour **20 176** sinistres.

2.7 Revalorisation des sinistres

La base mise à disposition pour cette étude recense l'ensemble des sinistres corporels en responsabilité civile automobile survenus sur la période 2009-2017. Le choix d'une période d'observation aussi large est principalement motivé par le souhait de récupérer une plus grande volumétrie de sinistres afin de développer une méthode d'estimation de la charge ultime suffisamment robuste.

En contrepartie, les montants des sinistres survenus sur différents exercices ne sont néanmoins pas comparables, et ce en raison de l'inflation du coût des préjudices corporels. Une étape de revalorisation des sinistres est nécessaire afin de rendre l'évolution des coûts sur la période 2009-2017 exogène à notre étude. Néanmoins, les processus de gestion des sinistres corporels ont pu être modifiés au cours de ces neuf années d'observation venant possiblement modifier la liquidation des sinistres. De ce fait, l'exercice de survenance des sinistres peut potentiellement être source d'hétérogénéité dans le développement des sinistres même après retraitement de l'inflation du coût des dommages corporels.

Afin de pallier à la problématique d'inflation des préjudices corporels, les montants des sinistres sont revalorisés en « *as-if* » afin de les rendre comparables pour une même date de vision des sinistres, c'est-à-dire au 31/12/2017 dans le cas présent.

Pour ce faire, il nous faut récupérer un indice représentatif de l'évolution du coût des sinistres présents dans notre base sur la période 2009-2017. Pour les dommages corporels, cette tâche s'avère particulièrement complexe au vu du grand nombre de postes de préjudices existants dans l'indemnisation des victimes d'accidents corporels [3].

Le coût d'un sinistre corporel est étroitement lié à la gravité du préjudice subi par les victimes, il n'existe donc à priori pas d'indice de référence reflétant l'évolution globale du coût des dommages corporels tous degrés de préjudices confondus.

Néanmoins, il est possible récupérer des informations sur l'évolution de l'indemnisation des corporels pour un degré de gravité constant, comme les différents indices publiés en avril 2013 par la FFSA (Fédération Française des Sociétés d'Assurance) et GEMA (Groupement des Entreprises Mutuelles d'Assurance) [4] représentés en Figure 2.7.

3. La nomenclature Dinthilac publiée en juillet 2005 établit une liste non exhaustive des principaux postes de préjudices corporels.

4. Ces deux organismes ont maintenant fusionné pour former la Fédération Française de l'Assurance (FFA).

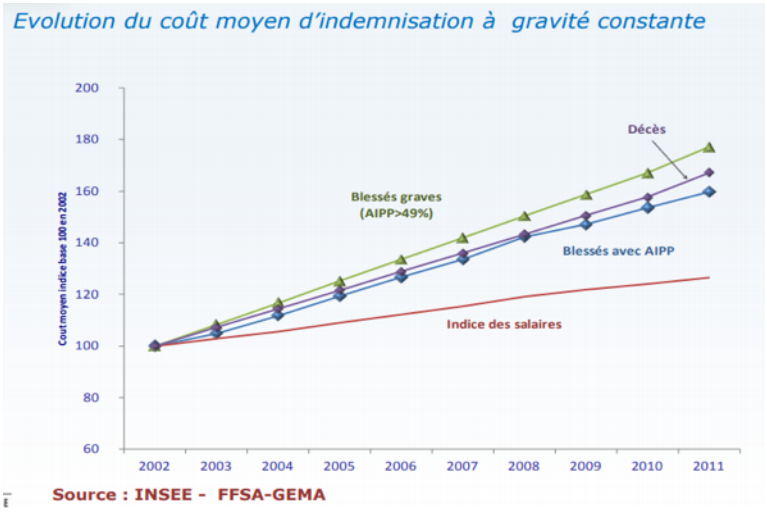


FIGURE 2.7 – Evolution du coût moyen d'indemnisation des sinistres corporels à gravité constante sur la période 2002-2011

Quel que soit le degré de gravité, l'indemnisation moyenne des dommages corporels a évolué bien plus rapidement que la revalorisation des salaires sur la période 2002-2011. De ce fait, il ne parait pas pertinent d'utiliser ce dernier indice ou bien l'indice des prix à la consommation pour réindexer les sinistres corporels. En effet, ce choix entraînerait une sous-estimation du coût des sinistres passés.

Sur ces 10 années, l'inflation moyenne était de 6% par an pour les cas de blessures graves, de décès et pour les blessés avec un taux d'AIPP positif.
En absence d'AIPP de la victime, soit 71% des accidents corporels en 2013 [5], l'indemnisation versée est forfaitaire sous réserve que l'assureur adverse soit adhérent à la convention IRCA.

Rappelons que dès lors que la victime d'un accident de la route déclare un préjudice corporel au près des gestionnaires sinistres MMA, le sinistre est ouvert comme étant corporel même si in fine, aucun versement au titre de dommages corporels n'a eu lieu, seuls les dommages matériels sont indemnisés le cas échéant.
Pour les sinistres présents dans notre base, les règlements matériels représentent en moyenne près de 22% de la charge sinistre. A partir de ce constat, il peut être pertinent de s'intéresser à l'évolution du coût total des réparations automobiles.

A partir des données publiées par SRA [6], le coût des réparations automobiles a augmenté en moyenne d'un peu moins de 3% par an sur la période 2009-2017. Un autre indicateur reflétant l'inflation des coûts matériels est l'évolution moyenne des montants des forfaits de la convention IRSA depuis 2009 qui se révèle être légèrement inférieur à 2% [7].

Après réflexion et en prenant en compte l'évolution des coûts de ces deux types de préjudices, il a été convenu d'utiliser un taux de revalorisation annuel constant fixé à **4%** correspondant à une pondération de ces évolutions annuelles moyennes en fonction du poids du préjudice associé (matériel/corporel) dans la charge globale des sinistres présents dans notre base.

5. Source : AFA-ACP
6. Disponibles sur le site du SRA : http ://www.sra.asso.fr
7. Les montants forfaitaires sont présentés en annexe A

Bien que ce choix engendre une simplification de la revalorisation des sinistres, il ne détériore en rien la qualité de la méthodologie développée dans ce mémoire.

2.7.1 Construction des statistiques *as-if*

Sachant qu'une base comptable est à notre disposition pour cette étude, il est possible de récupérer pour chaque sinistre un certain nombre d'informations associées aux règlements déjà effectués et en particulier leur date de versement.

Afin d'affiner la revalorisation des sinistres présents dans notre base, il a été préféré d'adopter un pas mensuel en utilisant cette date de versement pour calculer nos statistiques « *as-if* ».

Soit :

- n l'année pendant laquelle le règlement (resp. le recours) a été effectué
- m le mois de l'année n pendant lequel le règlement (resp. le recours) a été effectué
- $Regt_m^n$ le montant réglé au mois m de l'année n
- $Regt\ as\ if_m^n$ le montant réglé au mois m de l'année n vu à fin décembre 2017
- Rec_m^n le montant de recours encaissé au mois m de l'année n
- $Rec\ as\ if_m^n$ le montant de recours encaissé au mois m de l'année n vu à fin décembre 2017

La statistique « *as-if* », c'est-à-dire les montants observés au mois m de l'année n réindexés au 31/12/2017, est obtenue de la façon suivante :

Avec :

$Ecart\ annuel_n = 2017 - n$
$Ecart\ mensuel_m = 12 - m$

$$Regt\ as\ if_m^n = Regt_m^n \times (1 + 4\%)^{(Ecart\ annuel_n + \frac{Ecart\ mensuel_m}{12})}$$

En suivant le même principe, les montants de recours encaissés « *as-if* » sont obtenus de la façon suivante :

$$Rec\ as\ if_m^n = Rec_m^n \times (1 + 4\%)^{(Ecart\ annuel_n + \frac{Ecart\ mensuel_m}{12})}$$

Le coût « *as-if* » (noté $Cout\ as\ if_m^n$) de chaque sinistre observé au mois m de l'année n est obtenu par différence entre la somme des règlements et des recours « *as-if* ».

Après s'être intéressé aux montants réglés, il faut également, par soucis de cohérence, retraiter les charges dossier/dossier. Cependant, la méthodologie à appliquer n'est pas la même puisque ces montants ne correspondent pas à des règlements ou à des encaissements.

Par simplicité, la charge dossier/dossier « *as-if* » (notée $Charge\ d/d\ as\ if_m^n$) observée au mois m de l'année n est obtenue en ajoutant au coût « *as-if* » la provision dossier/dossier à cette même date, c'est-à-dire la différence entre la charge dossier/dossier et le coût du sinistre non réindexés :

$$Charge\ d/d\ as\ if_m^n = Cout\ as\ if_m^n + (Charge\ d/d_m^n - Cout_m^n)$$

Cette procédure nous permet de réindexer au 31/12/2017 les coûts et les charges dossier/dossier observés au mois m de l'année n pour les rendre comparables.

Sans ce niveau de granularité des données, les statistiques « *as-if* » auraient pu être construites à partir d'une vision de fin d'année des montants de sinistres en utilisant un pas annuel de revalorisation. Cette alternative, bien que plus simple, aurait été toutefois moins précise puisque les montants réglés en début d'année civile auraient été réindexés de la même façon que ceux versés en fin d'année.

2.8 Analyse descriptive de la base d'étude

Avant de procéder à l'étape de modélisation, il convient tout d'abord d'avoir un aperçu des différentes données présentes dans la base d'étude. Par soucis de cohérence, les statistiques descriptives sont réalisées à partir des montants de sinistres « *as-if* » afin de les rendre comparables.

2.8.1 Présentation générale

Pour rappel, la base d'étude est composée de **77 861** sinistres corporels en responsabilité civile automobile survenus sur la période 2009-2017. Parmi ces sinistres, **12 365** sont toujours en cours au 31/12/2017, représentant plus de **20%** des sinistres présents dans la base.

Analyse de la variable cible

Le Tableau 2.7 présente les statistiques descriptives du coût des sinistres en euros :

Min.	*1er quartile*	*Médiane*	*Moyenne*	*3ème quartile*	*Max.*
-16 604	1 397	2 525	9 598	4 734	8 978 211

TABLE 2.7 – Statistiques descriptives du coût des sinistres au 31/12/2017

L'étendue du coût est relativement grande puisque certains sinistres présentent des montants négatifs alors pour d'autres, les coûts peuvent s'élever à plusieurs millions d'euros. La présence de montants négatifs s'explique dans le cas où les recours encaissés (notamment forfaitaires) sont supérieurs aux règlements versés. D'autre part, la base d'étude recense quelques sinistres majeurs entraînant des préjudices corporels importants. Malgré cette grande étendue, la moitié des sinistres présentent des coûts compris entre 1 397 et 4 734 euros.
Pour compléter cette analyse, la Figure 2.8 illustre la distribution du coût des sinistres.

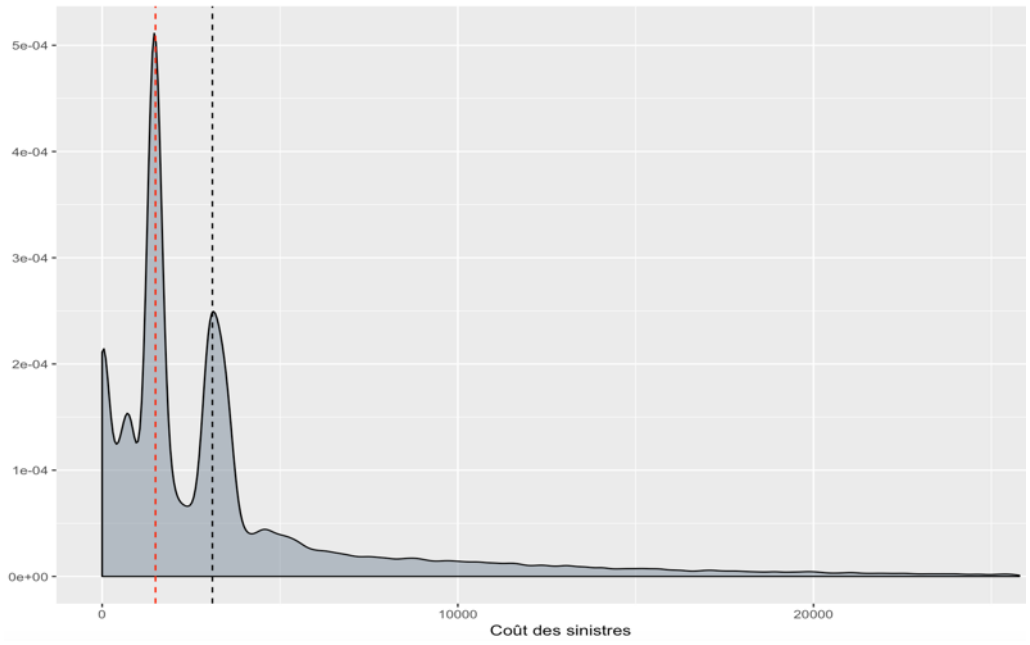


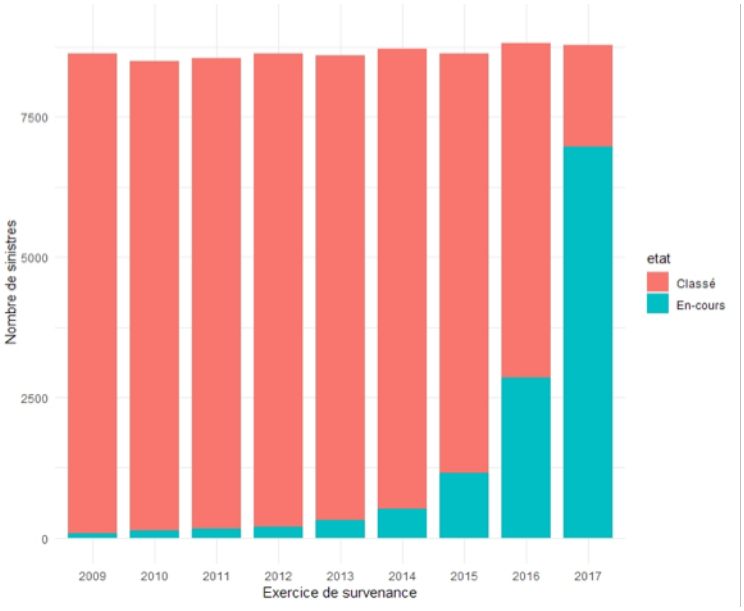
TABLE 2.8 – Distribution du coût des sinistres au 31/12/2017 bornée entre 0 et le 95ème centile

La distribution du coût des sinistres présente deux pics pour des valeurs proches de 1 500 euros (pointillés rouges) et 3 000 euros (pointillés noirs) induits par l'application des deux conventions d'indemnisation, l'IRCA et l'IRSA, dont MMA est signataire.

Le premier pic de distribution autour de 1 500 euros correspond majoritairement aux recours forfaitaires réglés dans le cadre de l'une des deux conventions [8]. Le second correspond quant à lui aux sinistres réglés par l'application simultanée des deux conventions d'indemnisation.

Etat des sinistres

La Figure 2.8 présente la répartition de ces sinistres par exercice de survenance en fonction de leur état.



Exercice de survenance	Nombre de sinistres		Proportions sinistres en-cours
	En-cours	Clos	
2009	83	8 547	1%
2010	134	8 369	2%
2011	158	8 397	2%
2012	198	8 442	2%
2013	310	8 292	4%
2014	514	8 197	6%
2015	1 149	7 476	13%
2016	2 856	5 954	32%
2017	6 963	1 822	79%

FIGURE 2.8 – Répartition des sinistres par exercice de survenance selon leur état

Parmi les sinistres en-cours présents dans la base d'étude, plus de **56%** sont survenus en 2017, soit **6 963** sinistres. De plus, on observe logiquement que la proportion de sinistres en-cours par exercice est d'autant plus importante que l'exercice de survenance est récent, elle est même de 79% en 2017.

Bien que les sinistres en-cours ne représentent en nombre qu'une part marginale sur les exercices passés (moins de 6% sur les exercices allant de 2009 à 2014), ce constat est sensiblement différent en raisonnant en charge sinistre, comme l'illustre la Figure 2.9.

8. Les montants forfaitaires des conventions IRCA et IRSA depuis 2009 sont présentés en annexe.

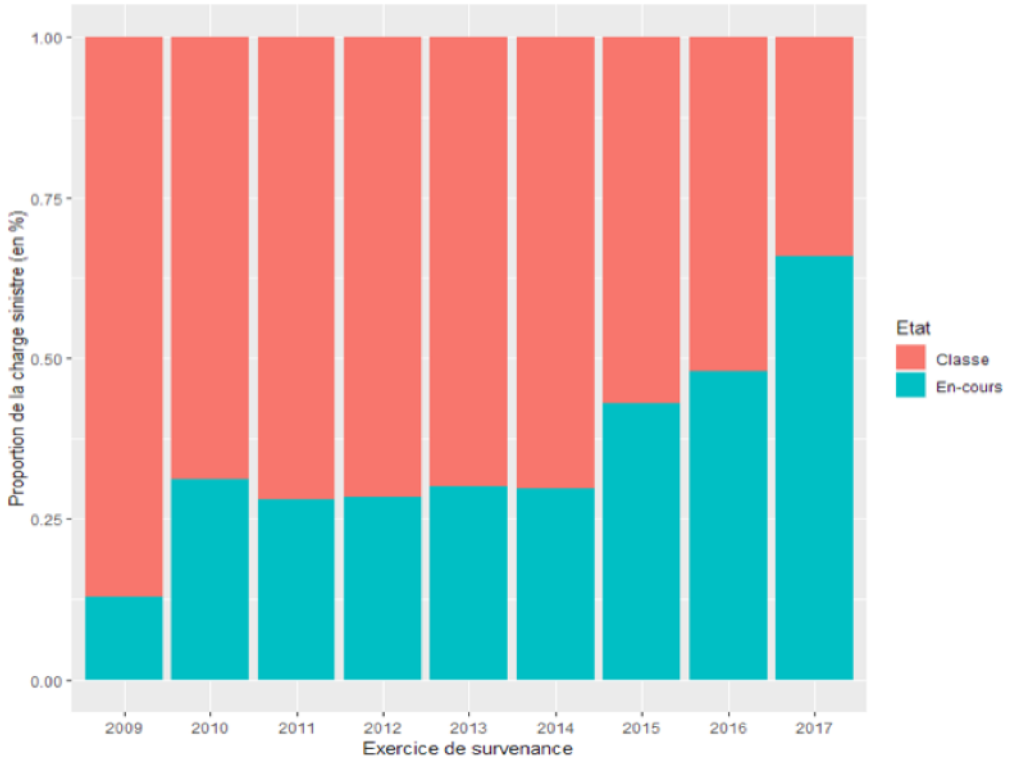


FIGURE 2.9 – Répartition des sinistres par exercice de survenance selon leur état

En prenant le cas de l'exercice 2010, les sinistres en-cours ne représentent que 2% des sinistres survenus sur cet exercice (soit 158 sinistres) mais plus de 30% de la charge sinistre globale observée au 31/12/2017 sur ce même exercice.

Durée de règlement

En s'intéressant maintenant à la durée de règlements, les sinistres présents dans notre base se liquident en un an et demi en moyenne. La moitié d'entre eux sont réglés en moins d'un an. Ce constat est à priori étonnant pour une branche longue comme la responsabilité civile corporelle automobile.

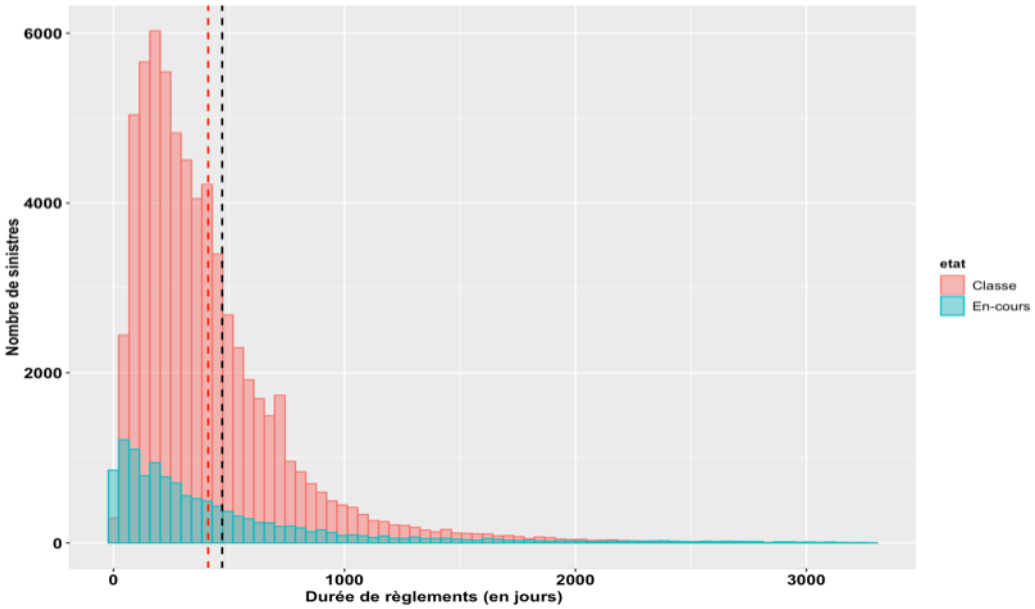


FIGURE 2.10 – Distribution de la durée de règlement des sinistres en fonction de leur état

Nous observons sur la Figure 2.10 que la durée de règlement des sinistres en-cours calculée au 31/12/2017 est légèrement plus longue que celle des sinistres classés (pointillés noire VS pointillés rouge), d'autant plus que, par définition, cette information est censurée à droite.
Notons également la forte concentration des sinistres en-cours sur l'extrémité droite de l'histogramme présenté en Figure 2.10. Ces derniers représentent plus de 60% des sinistres ayant vécus au moins sept ans.

A ce stade, il semble pertinent de s'intéresser à la relation entre le coût moyen d'un sinistre et son délai de règlement.

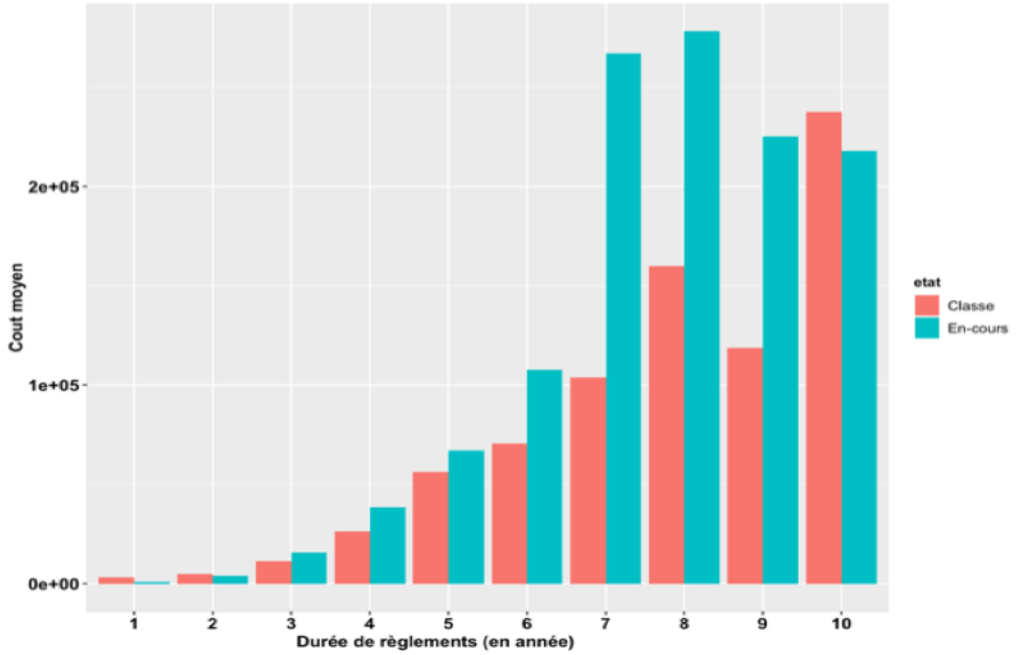


FIGURE 2.11 – Coût moyen des sinistres en fonction de la durée de règlement et selon leur état

Nous pouvons observer très nettement sur la Figure 2.11 une relation croissante entre le coût moyen des sinistres et leur durée de règlement (exprimée en années). A noter que dans ce graphique, un délai de règlement annuel égal à 1 pour un sinistre clos signifie qu'il s'est clôture en moins d'un an. Pour un sinistre en-cours, cela signifie qu'à la date d'extraction des données (au 31/12/2017) ce sinistre est ouvert depuis moins d'un an.

D'autre part, le coût moyen des sinistres en-cours ouverts en gestion depuis plus de 5 ans dépasse les 200 000 euros et est supérieur à celui des sinistres classés. Outre le fait que la durée de règlement des sinistres figure parmi les variables explicatives de notre méthode individuelle, elle sera utilisée lors de la construction des modèles de projection [9].

9. La construction de la méthode individuelle est détaillée dans la partie 3.3.1

2.8.2 Analyse des variables explicatives

La base d'étude recense de nombreuses informations sur les sinistres, la liste exhaustive est présentée en annexe C. Après une analyse préalable, seules les variables les plus pertinentes pour expliquer la liquidation des sinistres sont présentées dans cette section.

Le taux d'AIPP

Même si le taux d'AIPP n'est pas toujours renseigné (cf partie 2.4), il reflète le niveau de gravité du sinistre. Cette variable a été discrétisée en quatre modalités selon différents niveaux de gravité définis de façon à conserver des coûts de sinistres homogènes dans chaque classe.

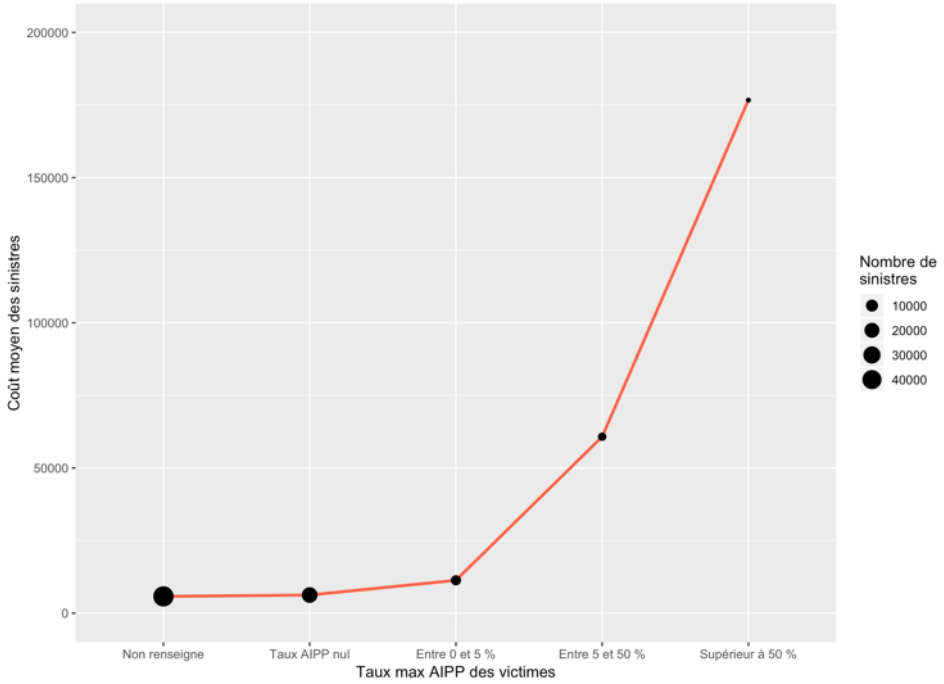


FIGURE 2.12 – Coût moyen des sinistres en fonction du taux d'AIPP max des victimes

La Figure 2.12 illustre la relation croissante entre le coût des sinistres et le taux d'AIPP des victimes en notant tout de même que peu de sinistres dans la base d'étude présentent des taux d'AIPP de plus de 5% (2 475 sinistres). Intuitivement, plus le niveau de gravité du préjudice subi est élevé, plus l'indemnisation versée sera importante.

La nature du préjudice subi par les victimes est également une information intéressante à exploiter. Elle permet de distinguer les préjudices matériels, corporels et mortels. L'analyse de cette variable a été réalisée dans la partie 2.6.3.

L'évaluation à l'origine

L'évaluation à l'ouverture d'un sinistre peut être réalisée en utilisant un montant forfaitaire ou bien en l'estimant à son coût réel. En Responsabilité Civile Corporelle automobile, le montant forfaitaire est différent selon la section en charge de la gestion du dossier reflétant la gravité du sinistre (cf partie 2.1).

Ces montants correspondent bien souvent au coût moyen observé des sinistres gérés par la section en question.

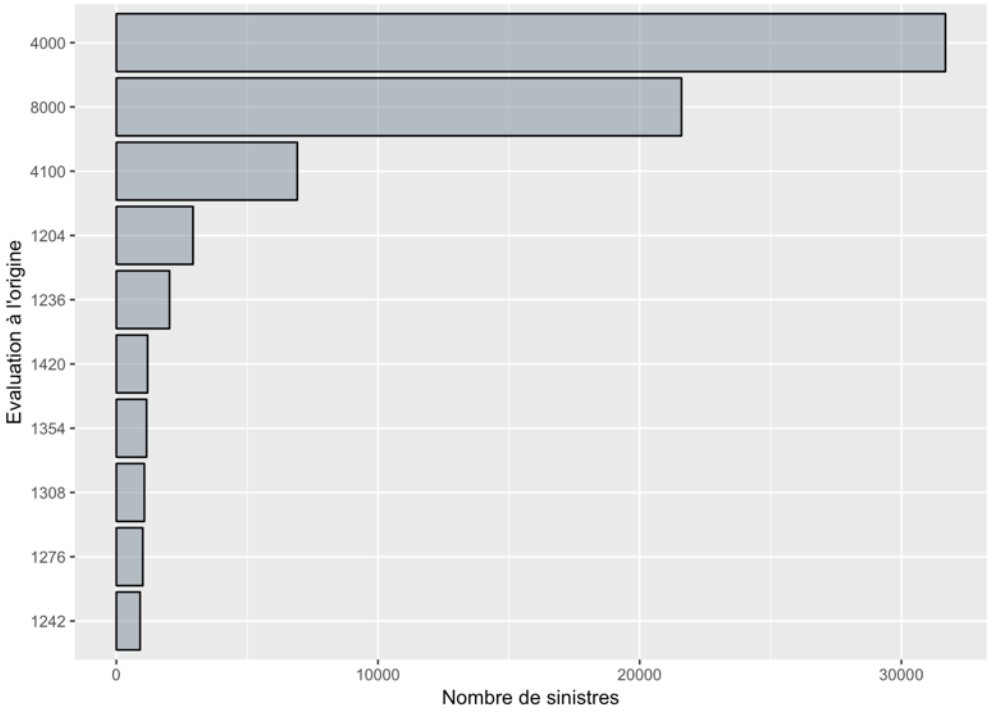


FIGURE 2.13 – Les 10 évaluations à l'origine les plus fréquentes dans la base d'étude

Plus de 90% des sinistres présents dans la base d'étude sont évalués à l'ouverture à l'aide de forfaits dont une partie sont proches des montants de recours fixés par la convention IRSA (cf. annexe A). Pour ces sinistres, seuls des préjudices matériels ont vraisemblablement été déclarés au moment de leur ouverture dans le système de gestion. Les évaluations fixées à 4 000 et 8 000 euros (soit 77% des sinistres) concernent l'indemnisation de préjudices corporels.

Au final, très peu de sinistres sont estimés à leur coût réel à l'ouverture (moins de 10%).

La section de gestion du sinistre

Comme présenté dans la partie 2.1, la section en charge de la gestion du sinistre est définie selon sa gravité (mesurée par le taux d'AIPP) ou bien en fonction de son coût. Pour rappel, la gestion des sinistres corporels est réalisée principalement par trois sections : corporels **légers**, **médians** et **importants**. D'autres sections annexes sont sollicitées pour la gestion de sinistres matériels ou ceux survenus à l'étranger (regroupées au sein de la modalité "Autres").

Les sinistres gérés par les sections coporels médians et importants se distinguent des autres avec un coût moyen nettement supérieur (Figure 2.14). Ces derniers sont toutefois sous-représentés dans la base d'étude, on en décompte seulement 3 947.

Les sinistres ont été répartis en fonction de la section de gestion sinistres observée au 31/12/2017. Cette information est tout à fait susceptible de changer au cours de la liquidation d'un sinistre.

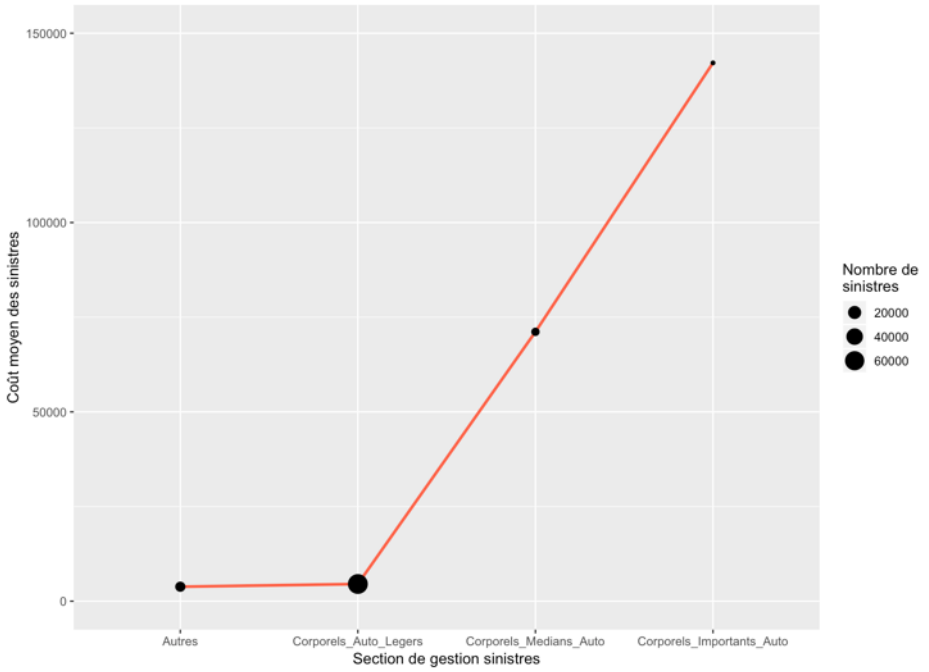


FIGURE 2.14 – Coût moyen des sinistres en fonction de la section de gestion sinistres au 31/12/2017

Nature judiciaire

Certains sinistres peuvent faire l'objet de procédures judiciaires notamment pour fixer les montant d'indemnisation à verser aux victimes, leurs modalités de règlements etc.. Elles interviennent généralement lors de sinistres majeurs ou bien lorsqu'une victime est décédée.

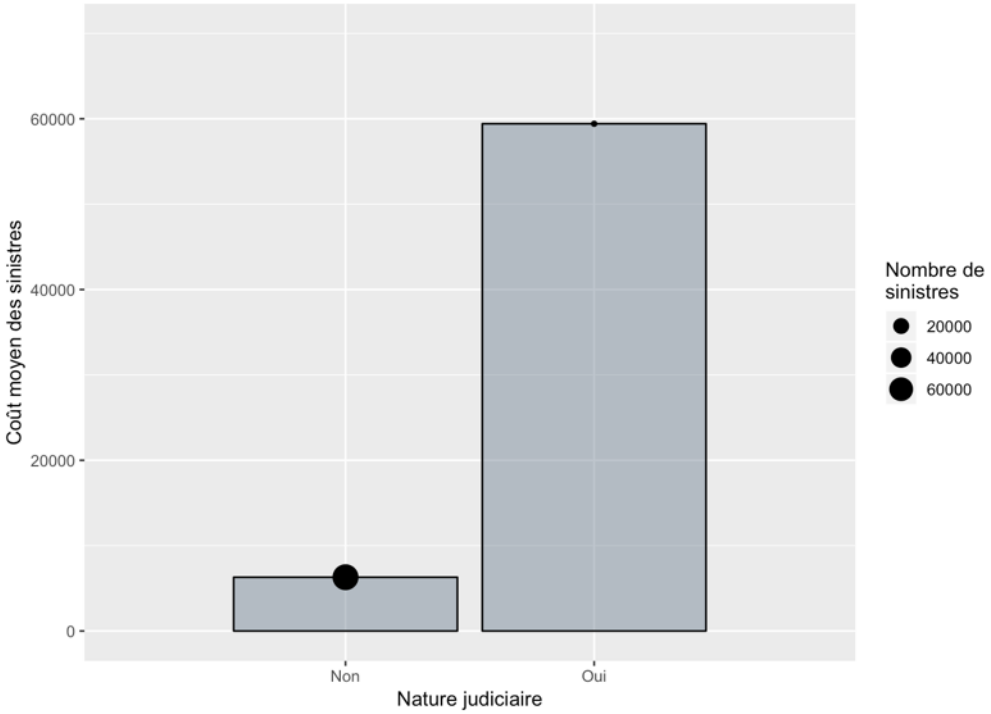


FIGURE 2.15 – Coût moyen des sinistres en fonction de la nature judiciaire (Oui/Non)

Les sinistres réglés suite à une procédure judiciaire sont nettement plus coûteux (59 423 euros contre 6 301 euros) mais ne représente seulement qu'un peu plus de 3% des sinistres présents dans la base d'étude.

Chapitre 3

Méthode d'estimation de la charge ultime basée sur des données individuelles

Depuis maintenant plusieurs années, l'intérêt pour les méthodes individuelles n'a cessé de s'accroître, notamment pour répondre à des problématiques de provisionnement.

Parmi les références en la matière, Larsen [15], Antonio et Plat [1], Pigeon et al. [19] et bien d'autres, proposent de modéliser l'occurrence et le développement des sinistres par une approche probabiliste. Les modèles développés dans ces travaux reposent néanmoins sur des structures paramétriques peu flexibles pouvant se révéler difficiles à implémenter en pratique.

Plus récemment, les articles de M.V.Wüthrich [23], M.Baudry et C.Y.Robert [3] et M.V.Wüthrich [24] mettent en avant l'inconvénient majeur des méthodes agrégées, à savoir l'absence d'utilisation des informations propres aux sinistres dans l'estimation de leur coût ultime. Ces auteurs proposent d'utiliser des concepts issus du *Machine Learning* (resp. arbres de décision CART, *Random Extra Trees*, réseaux de neurones) pour implémenter leur approche ligne à ligne.

La méthode développée dans ce mémoire s'inspire principalement de ces trois deniers articles. La problématique du provisionnement ne sera pas en tant que telle abordée, nous nous intéressons ici uniquement à la projection des sinistres à l'ultime.

3.1 Motivations du développement d'une méthode individuelle

Une méthodologie basée sur des données individuelles vise à répondre aux différentes lacunes induites par l'utilisation de méthodes agrégées pour projeter les sinistres à l'ultime.

Tout d'abord, les méthodes individuelles fournissent des estimations du coût ultime par sinistre contrairement aux méthodes agrégées. Généralement, ces dernières permettent uniquement d'obtenir des prédictions par exercice de survenance ou bien par année de déclaration.

La perte des informations individuelles induite par l'agrégation des données constitue la principale lacune des méthodes agrégées. Les données récoltées sur les sinistres peuvent pourtant se révéler précieuses pour améliorer la prédiction de leur coût ultime.

De plus, la granularité des données en entrée des triangles de liquidation est une question qui se pose lors de l'utilisation de méthodes agrégées. Bien qu'elles se révèlent particulièrement efficaces en présence d'une grande volumétrie de sinistres homogènes, il n'est souvent pas possible de les utiliser sur des périmètres plus réduits.

Afin de faire le lien avec la problématique de ce mémoire, le nombre de sinistres par contrat ou par garantie ne permet généralement pas de construire des triangles de liquidation suffisamment stables pour accorder un niveau de confiance convenable aux coûts ultimes estimés et ce, en raison de l'hétérogénéité des sinistres individuels dont l'impact est démultiplié en présence d'une faible volumétrie de données.

D'autre part, la présence de sinistres graves se révèle bien souvent problématique. Souvent source d'instabilité, un traitement spécifique sur les sinistres graves est réalisé d'un point de vue opérationnel afin d'éviter les effets multiplicatifs des méthodes agrégées dans les triangles de liquidation. Le développement d'un sinistre majeur vient augmenter drastiquement la charge sinistre associée à son exercice de survenance ayant pour conséquence une modification significative des facteurs de développement et donc de l'ultime estimé.

Là aussi, la méthodologie développée dans ce mémoire tentera de répondre à cette problématique tout en sachant que ces sinistres sont peu nombreux et très hétérogènes.

Un autre point relevant plutôt d'une problématique technique concerne la validité des hypothèses sous-jacentes aux méthodes agrégées de type Chain Ladder. L'idée ici est de construire une méthode ne reposant pas sur des hypothèses trop contraignantes pour qu'elle soit relativement flexible et ainsi faciliter son utilisation.

Enfin, la projection individuelle des sinistres permet d'appliquer des traités de réassurance de type excédent de sinistres ce qui est difficilement réalisable en utilisant des méthodes agrégées.

Pour toutes ces raisons, l'utilisation d'une approche ligne à ligne apparait comme une alternative intéressante.

Avant de présenter la démarche entreprise dans cette étude, il convient tout d'abord d'effectuer un bref rappel sur la vie d'un sinistre en assurance IARD.

3.2 Généralités sur la vie d'un sinistre en assurance IARD

Un sinistre survient à une date notée T_1 appelée date de l'accident ou date d'occurrence du sinistre pendant la période d'assurance du contrat. Ce sinistre est ensuite déclaré auprès de l'assureur à une

date notée T_2 (avec $T_2 \geq T_1$) appelée date de connaissance du sinistre.

Cette date est toutefois à distinguer de la date d'ouverture de la garantie (notée T_3) dans le système de gestion de l'assureur. En effet, il est possible que dans certains cas, le gestionnaire sinistre soit dans l'attente de nouvelles informations sur le sinistre pour pouvoir ouvrir les garanties concernées. Cependant, sauf exceptions, le délai entre ces deux dates est généralement très court et est même nul pour près de 80% des sinistres présents dans notre base.

Après ouverture des garanties concernées par le sinistre survenu, plusieurs règlements vont avoir lieu à différentes dates. Une fois que toutes les indemnités dues par l'assureur sont versées, le sinistre peut être clôturé à une date notée T_4 (avec $T_4 \geq T_3$). En toute généralité, une fois clos, un sinistre le reste indéfiniment. Cependant dans certains cas, un sinistre peut être réouvert si de nouvelles informations sur les circonstances du sinistre sont dorénavant connues de l'assureur ou bien lorsque l'état des victimes s'est aggravé venant modifier le montant des indemnités à verser. Ce dernier cas arrive ponctuellement pour les sinistres corporels en responsabilité civile automobile, près de 9% des sinistres présents dans la base d'étude ont été ré-ouverts.

Certains sinistres présents dans notre base ne sont pas encore classés à la date de vision où ces derniers ont été extraits, la date de clôture ainsi que les montants d'indemnités non encore réglés par l'assureur sont alors inconnus, on parlera alors de sinistres RBNS (*Reported But Not Settled*).

Soit T_v la date de vision des données, la vision d'un sinistre RNBS ($T_3 < T_v < T_4$) à cette date est représentée sur la Figure 3.1.

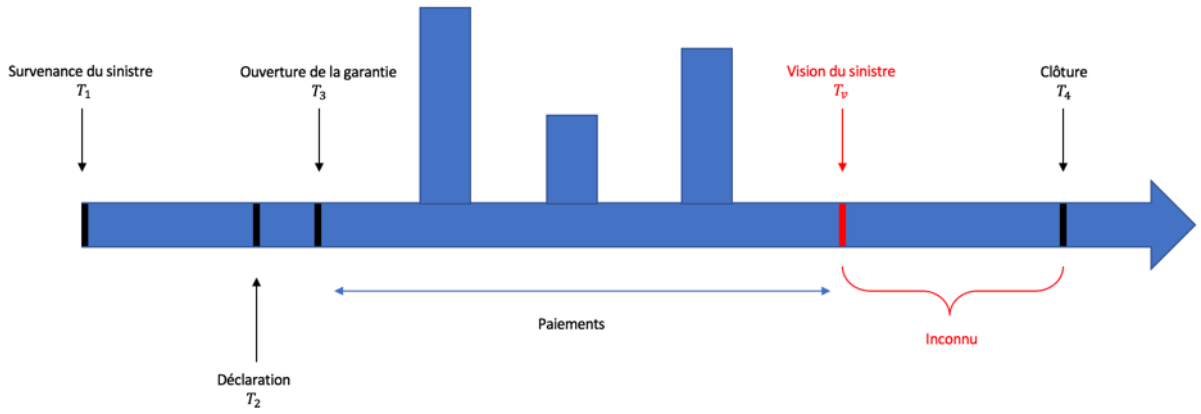


FIGURE 3.1 – Représentation d'un sinistre RBNS

Les sinistres survenus pendant la période d'assurance mais non encore déclarés à la compagnie d'assurance à la date de vision des données (cas où $T_1 < T_v < T_2$) sont appelés INBR (*Incurred But Not Reported*). Ces sinistres sont alors inconnus de l'assureur mais doivent en théorie être pris en compte dans la projection du coût des sinistres d'un exercice quelconque.

La méthodologie développée dans ce mémoire se consacre uniquement à la projection de la charge des sinistres RNBS à l'ultime, les sinistres IBNR ne seront pas étudiés. Ce choix aura son importance lors de la comparaison des résultats obtenus avec la méthode Chain Ladder.

3.3 Présentation de la méthode de projection individuelle

Dans cette partie, la méthode développée dans ce mémoire est, dans un premier temps, présentée d'un point de vue théorique. Ensuite, après avoir redéfini notre périmètre d'étude, les étapes de construction des bases en entrée des modèles sont détaillées.

La méthodologie développée est implémentée d'une part à l'aide de *forêts aléatoires (Random Forest)*. Les résultats seront ensuite challengés en utilisant l'algorithme du *Gradient Boosting Machine (GBM)*.

3.3.1 Formalisation

L'objectif de ce mémoire est de proposer une méthode d'estimation du coût ultime des sinistres RBNS à partir des informations connues jusqu'à la date T_v, la date de vision des données.

En reprenant les notations de la partie 3.2 et en introduisant I le nombre de sinistres en-cours à la date T_v, $(\mathcal{F}_t)_{t\geq 0}$ l'information disponible à la date t et C_t^i le coût du sinistre i à la date t, l'estimation du coût ultime d'un sinistre i à la date t (notée $\hat{C}_t^i$) peut alors s'exprimer de la façon suivante :

$$\hat{C}_{T_v}^i = \mathbb{E}[C_{T_{4,i}}^i | \mathcal{F}_{T_v}] \ \forall \ i \ \in \ \{1, ..., I\}$$

Une première approche serait de considérer uniquement les dernières informations connues à la date de vision des sinistres pour en estimer leur coût ultime. Cependant, l'évolution du sinistre depuis son ouverture dans le système de gestion jusqu'à la vision de ce dernier à la date T_v ne serait par conséquent pas prise en compte.

La démarche entreprise dans ce mémoire consiste à utiliser un maximum d'informations sur les sinistres, cette première hypothèse de travail est alors écartée.

L'idée ici est d'adopter une vision par période de développement, c'est-à-dire que les sinistres sont observés à différentes étapes de leur liquidation sur des intervalles de temps discrets. Ces visions sont ensuite utilisées pour estimer le coût ultime de chaque sinistre.

On considère alors une grille de taille $P \times \delta$ constituée de P périodes d'observation des sinistres, chacune de durée δ. A la fin de chacune des périodes considérées, les informations relatives aux sinistres sont récupérées.

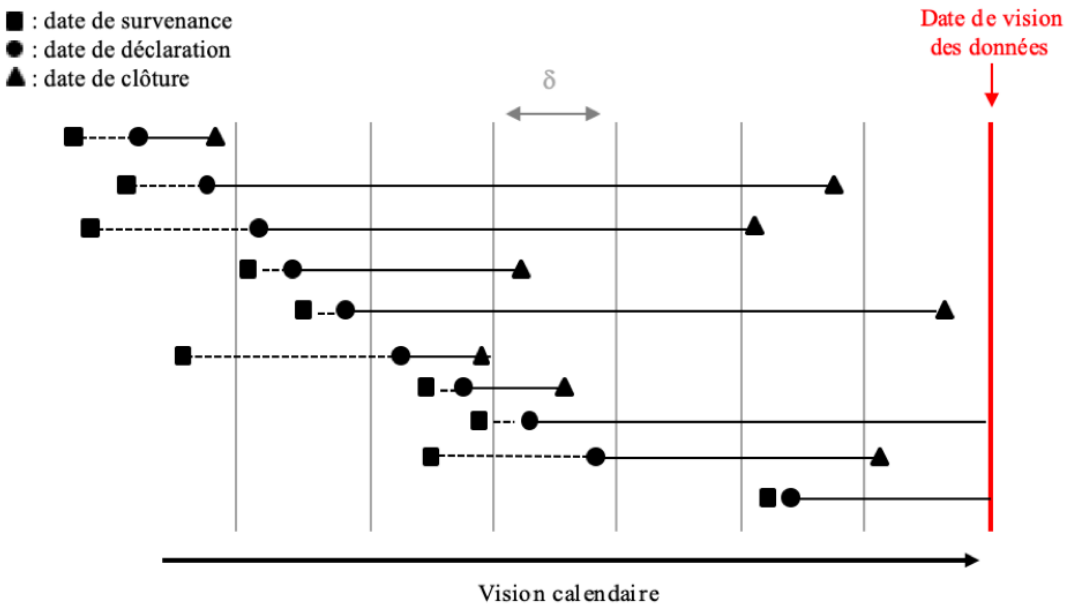


FIGURE 3.2 – Exemple de grille de taille $P \times \delta$ pour $P = 7$

A la date de vision des données, l'ensemble des sinistres en-cours présents dans la base d'étude ne sont évidemment pas tous rendus au même stade dans leur liquidation. En effet, la profondeur d'historique est plus ou moins importante en fonction de la date à laquelle les sinistres ont été ouverts dans le système de gestion.

Les sinistres en-cours peuvent alors être regroupés en (P-1) groupes de telle sorte à ce que le groupe p soit constitué des sinistres en-cours pour lesquels les informations connues à la fin de la période p correspondent à la dernière vision connue ($\forall\ p\ \in\ \{1, ..., P\}$).

Pour ces sinistres, les informations à la fin des périodes p+1 à P sont alors inconnues. Si pour un sinistre en-cours donné, toutes les informations le concernant sont connues jusqu'à la fin de la période P, alors ce sinistre est considéré comme clos (*).

Soit J_p le nombre de sinistres en-cours appartenant au groupe p et $(\mathcal{G}_t^j)_{t\geq 0}$ l'information disponible sur le sinistre j à la date t le coût ultime estimé peut s'exprimer de la façon suivante :

$$\hat{C}_{T_v}^j = \mathbb{E}[C_{T_{4,j}}^j | \mathcal{G}_{T_v}^j]\ \forall\ j\ \in\ \{1, ..., J_p\}$$

Avec :

$$\mathcal{G}_{T_v}^j = \{T_{3,j} < T_v < T_{4,j},\ (\mathcal{H}_u^j)_{0<u\leq p}\}$$

Où $(\mathcal{H}_u^j)_{0<u\leq p}$ représente l'ensemble des informations connues sur le sinistre j à la date de vision des données, observées à la fin de chacune des périodes d'observation.

Pour chacun des groupes p, un modèle sera ainsi construit afin d'estimer le coût ultime des sinistres affectés au groupe en question. Contrairement à la méthodologie développée par M.Baudry et C.Y.Robert [3] ou bien par D.Fabre Rudelle [11], les sinistres en-cours sont ici directement projetés à l'ultime sans estimation du coût de ces derniers à la fin de chaque période de développement future.

Cette méthodologie nécessite donc de calibrer P-1 modèles afin de projeter chacun des sinistres présents dans la base d'étude en fonction de la profondeur d'historique dont nous disposons sur ces derniers.

En prenant par exemple le groupe 2 ($p = 2$), notre méthode de projection des sinistres RBNS à l'ultime peut se représenter graphiquement de la façon suivante :

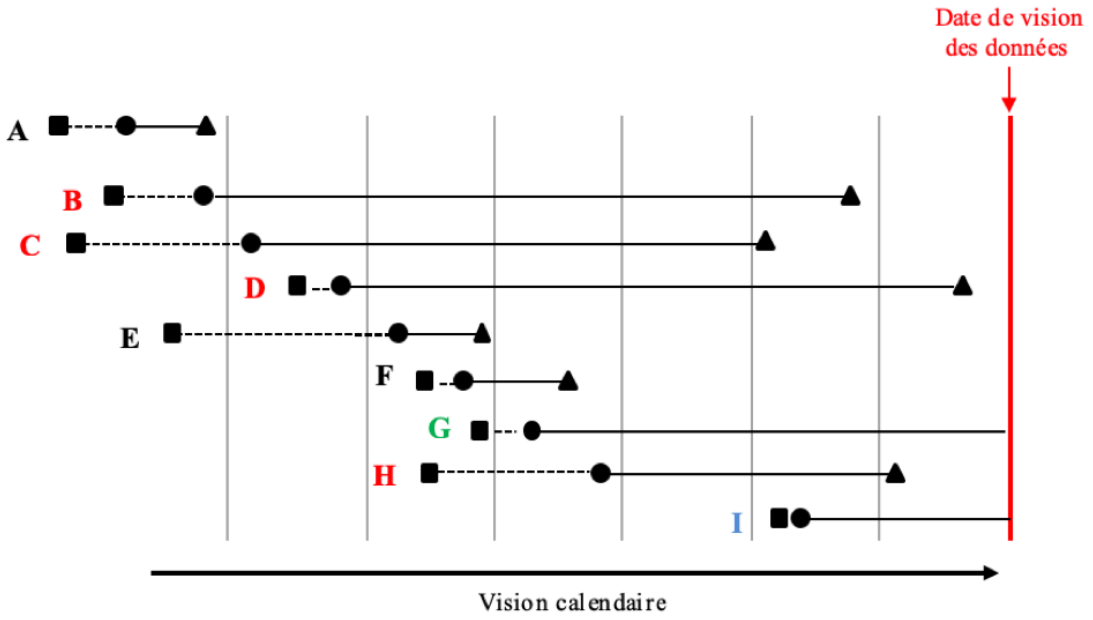


FIGURE 3.3 – Présentation de la méthode pour $p = 2$

Sur la Figure 3.3, le sinistre en-cours I appartient au 2ème groupe puisqu'il a été observé à la fin de deux périodes de durée δ. Pour estimer son coût ultime, seules les informations des sinistres B, C, D et H sont utilisées. Le modèle est construit à partir de ces sinistres de façon à apprendre le passage depuis la fin de la deuxième période d'observation jusqu'à l'ultime. Les données observées à l'ouverture ainsi qu'à la fin de la première période d'observation sont également utilisées pour cette estimation.

La mise en place d'une telle méthode suggère donc que les sinistres présents dans la base d'apprentissage soient obligatoirement clos puisque la connaissance du coût ultime est nécessaire. Le sinistre G n'est alors pas utilisé dans la phase d'entraînement puisque son coût ultime n'est pas connu. De plus, dans le cadre de notre exemple ($p = 2$), les informations observées sur les périodes suivantes et connues à la date T_v (périodes 3, 4 et 5 pour le sinistre B par exemple) ne sont pas utilisées dans la projection de la charge ultime. En effet, elles ne sont pas connues pour les sinistres en-cours observés sur deux périodes uniquement, il ne semble donc pas pertinent de les intégrer dans la base d'apprentissage.

Dans le cadre de ce mémoire, nous retiendrons des périodes d'observations annuelles ($\delta = 1$ an). Compte tenu de l'historique disponible, le développement des sinistres survenus est observé sur neuf exercices (P=9). De ce fait, les sinistres déclarés en 2009 sont considérés comme clos (*).

3.3.2 Suppression des sinistres « millionnaires »

Pour rappel, la base d'étude est composée de 77 861 sinistres observés sur neuf années (2009-2017). L'analyse descriptive des données présentée dans la partie 2.8.1 nous permet de constater des montants de sinistres très importants, pouvant s'élever à plusieurs millions d'euros. Pour pallier ce problème

d'homogénéité des montants de sinistres, une séparation entre sinistres graves et attritionnels est fréquemment réalisée en assurance non-vie afin d'opérer un traitement différent à ces deux typologies de sinistres.

Néanmoins, l'enrichissement de la base sinistres en intégrant diverses informations vise à détecter les différentes typologies de comportement de liquidation des sinistres. Ainsi, nous supposons que le développement des sinistres présentant une charge importante peut potentiellement s'expliquer au travers d'informations propres aux sinistres dont nous disposons dans la base d'étude. C'est pourquoi, une séparation entre sinistres graves et attritionnels n'est pas retenue dans le cadre de ce mémoire.

Toutefois, les sinistres présentant une charge dossier/dossier supérieure à un million d'euros bénéficient d'un traitement spécifique de la part des gestionnaires. En effet, ces sinistres entraînent bien souvent des procédures judiciaires et nécessitent donc un suivi particulier.

Compte tenu de leur comportement de liquidation atypique, les sinistres dont la charge dossier/dossier au 31/12/2017 est supérieure à un million d'euros sont écartés du périmètre de l'étude.

Après cette suppression, la base d'étude recense dorénavant **77 709** sinistres corporels en RC automobile.

3.3.3 Construction des bases pour l'implémentation

Le principe de la méthode individuelle développée dans ce mémoire consiste à projeter individuellement chaque sinistre en fonction de la quantité d'informations disponibles les concernant.
En conséquence, une base d'étude distincte est construite pour chacun des huit modèles considérés.

D'autre part, l'apprentissage des différents modèles nécessite la connaissance de la charge ultime des sinistres, c'est pourquoi seuls les sinistres clos au 31/12/2017 sont utilisés à des fins de calibrage, soit **65 537 sinistres**.

En pratique, pour projeter les sinistres en-cours appartenant au groupe p, la base en entrée du modèle est constituée de tous les sinistres clos au 31/12/2017 qui était toujours ouverts à la fin de la période p (cf Figure 3.3 pour p=2).

Le Tableau 3.1 dénombre les sinistres utilisés dans la construction des huit modèles. La volumétrie des sinistres en-cours appartenant à chaque groupe est également présentée.

Groupe / Modèle	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Base en entrée du modèle p	48 142	16 508	5 410	2 097	975	448	195	97
Sinistres en-cours	7 255	2 691	1 055	476	278	170	139	108

TABLE 3.1 – Composition des bases en entrée des différents modèles de projection

Une problématique de volumétrie de sinistres se pose pour projeter les sinistres ouverts en gestion depuis au moins 5 ans. En effet, à partir du modèle 6, les bases en entrée sont constituées de peu de

sinistres (moins de 500). Le coût ultime de ces 417 sinistres en-cours devra être estimé par une méthode alternative.

Après avoir réalisé différents tests, le coût ultime de ces sinistres sera estimé en retenant la dernière charge dossier/dossier connu. En présence d'une très faible volumétrie de sinistres, ce montant constitue à priori la meilleure estimation dont nous disposons au niveau individuel lorsque le sinistre est ouvert en gestion depuis au moins 5 ans.

Après observation des volumes disponibles pour chaque groupe de sinistre, 5 modèles de projection distincts seront construits. Notons tout de même que la méthode de projection individuelle développée dans ce mémoire permet de projeter les **11 755** sinistres ouverts en gestion depuis moins de 6 ans, soit plus de **96%** des sinistres en-cours au 31/12/2017.

En termes de notation, chaque groupe p est associé au modèle de projection p. A titre d'exemple, la projection des sinistres en-cours à la fin de la deuxième année de développement est réalisée à l'aide du deuxième modèle.

Découpage Apprentissage / Test des bases en entrée des modèles

Traditionnellement, un découpage base d'apprentissage / base de validation / base test est réalisé pour d'une part calibrer le modèle construit à partir de la base d'apprentissage (à l'aide de la base de validation) puis pour confronter les prédictions aux observations sur un échantillon indépendant, la base test.

Cependant, compte tenu de la faible volumétrie de sinistres dont nous disposons pour chaque modèle, chacune des bases est découpée en deux échantillons : une base d'apprentissage et une base test. Le calibrage des modèles est réalisé sur l'échantillon *Out-Of-Bag (OOB)* pour les forêts aléatoires, par validation croisée pour le *Gradient Boosting Machine (GBM)*[1].

Ces deux alternatives permettent d'utiliser l'ensemble des données dans la phase d'apprentissage et de ne pas se restreindre à un sous-échantillon.

La Figure 3.4 illustre le découpage réalisé pour chacune des bases en entrée des cinq modèles en retenant une répartition 80% - 20%.

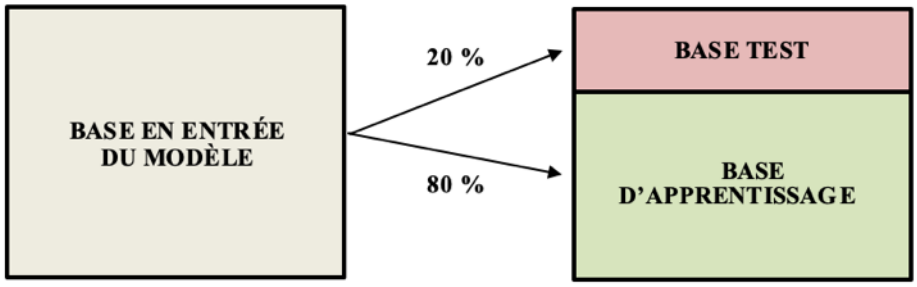


FIGURE 3.4 – Découpage de la base en entrée des modèles

1. Ce point sera davantage détaillé dans la partie 4.3

Le risque entraîné par ce découpage est d'observer des répartitions par typologie de sinistres significativement différentes dans les deux échantillons, principalement en termes de coûts. Une faible proportion de sinistres majeurs dans la base d'apprentissage conduirait à d'éventuelles mauvaises prédictions sur la base test.

Afin d'obtenir deux bases homogènes en termes de montants de sinistres, nous réalisons un échantillonnage stratifié sur le coût ultime. Ce procédé a pour objectif de conserver dans le nouvel échantillon créé les mêmes proportions de sous-population (des strates) que celles observées dans la base initiale.

Nous définissions alors, pour chaque base en entrée des modèles, 9 tranches (sous-population) correspondant à des quantiles de différents niveaux du coût ultime présentés dans le Tableau 3.2.

Tranches	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T8*	*T9*
Niveaux de quantiles	12,5%	25%	37,5%	50%	62,5%	75%	87,5%	99%	100%

TABLE 3.2 – Stratification des bases en entrée des modèles

Le choix de créer une neuvième tranche rassemblant les 1% des sinistres les plus coûteux permet de s'assurer que l'apprentissage des modèles sera bien réalisé en intégrant ces sinistres majeurs dont le coût ultime s'avère plus délicat à prédire.

Données historisées VS non historisées

Nous distinguons deux grandes catégories d'informations présentes dans la base d'étude.

Tout d'abord, certaines informations sont **non historisées** c'est-à-dire qu'elles restent identiques pendant toute la durée de vie du sinistre. Bien souvent, elles sont connues dès son ouverture dans le système de gestion. Parmi ces données, nous retrouvons par exemple la date de survenance du sinistre, la date de connaissance société, l'évaluation à l'origine du gestionnaire sinistre ou bien les causes sinistres.

Il y a également des informations pouvant potentiellement évoluées pendant la durée de vie du sinistre mais qui ne sont malheureusement pas historisées. On peut par exemple citer les données victimes qui sont seulement observées à la date d'extraction des données, c'est-à-dire au 31/12/2017. Pour un même sinistre, ces données sont par conséquent identiques sur toutes les périodes d'observation.

D'autre part, notre base de données recense également des informations pouvant évoluer pendant toute la durée de vie des sinistres et dont l'historique a pu être récupéré (données **historisées**). On retrouve bien évidemment le coût du sinistre, la charge dossier/dossier, son état mais également des informations comme la section de gestion sinistre en charge du dossier qui peut changer d'une période d'observation à une autre. C'est tout particulièrement ce type d'informations qui peut expliquer le changement de comportement de liquidation d'un sinistre.

La liste exhaustive des variables explicatives utilisées dans les modèles de projection est présentée en annexe C, en précisant si chacune des variables est historisée ou non historisée.

Chapitre 4

Implémentation

Avant de procéder à l'étape d'implémentation de la méthode de projection individuelle, il convient tout d'abord d'effectuer un bref rappel sur les deux algorithmes considérés dans ce mémoire.

4.1 Introduction aux méthodes ensemblistes

Les forêts aléatoires (*Random Forest*) et le *Gradient Boosting* sont deux algorithmes d'apprentissage automatique (*Machine Learning*) dont le principe consiste à agréger plusieurs modèles dans le but d'obtenir un meilleur pouvoir prédictif, ce sont des méthodes ensemblistes. Dans le cadre de ce mémoire, les modèles de base sont des arbres de décisions de type *CART (Classification And Regression Tree)*.

La construction de ce type d'arbres de décision binaires ne sera pas présentée ici, il est toutefois fondamental d'en comprendre le fonctionnement afin d'appréhender correctement les forêts aléatoires et le *Gradient Boosting*. Pour davantage de précisions sur l'algorithme CART, le lecteur intéressé peut se référer à l'article de L.Brieman « *Classification and regression trees* » [8].

Bien que les arbres CART ne soient pas présentés dans ce mémoire, il convient néanmoins d'en rappeler les principales lacunes afin de promouvoir l'utilisation des méthodes ensemblistes.

En tout premier lieu, le risque de surapprentissage reste tout de même très présent en utilisant ce type d'arbres de décision binaires, même s'il peut être réduit lors de la procédure d'élagage visant à améliorer la capacité de généralisation de l'arbre sur de nouvelles données. Ce risque fait référence au dilemme biais-variance bien connu en statistiques [1].

Outre cette première limite, le défaut majeur des arbres de décision CART concerne leur robustesse. En effet, les prédictions issues de ces modèles se révèlent être particulièrement instables, elles peuvent varier sensiblement lorsque la base d'apprentissage est légèrement modifiée.

Les méthodes ensemblistes visent à répondre à cette lacune en générant un ensemble d'arbres de décision CART dont les prédictions résultantes sont ensuite agrégées. Cependant, l'agrégation de modèles n'a d'intérêt que si les différents modèles ne génèrent pas des prédictions individuelles relativement proches. Dans le cas inverse, un seul modèle suffirait.

1. Le dilemme biais-variance est présenté dans le mémoire de M.Barbaste [2]

Dans le cadre de ce mémoire, les deux algorithmes sélectionnés pour l'implémentation de notre méthode de projection individuelle utilisent deux stratégies d'agrégation relativement différentes. Les forêts aléatoires reposent sur une stratégie de *bagging* (pour *Bootstrap Aggregating*) alors que le *Gradient Boosting* repose quant à lui, comme son nom l'indique, sur une stratégie de *boosting*.

4.2 Présentation des méthodes d'apprentissage automatique

L'objectif de cette partie est de présenter succinctement les forêts aléatoires et le *Gradient Boosting* pour en comprendre les principaux mécanismes. Pour davantage de précisions sur les aspects techniques, le lecteur intéressé pourra se référencer aux articles [7] et [13].

Le lecteur familier sur le fonctionnement de ces deux algorithmes peut directement se rendre à la partie 4.3 relative à l'optimisation des hyperparamètres.

4.2.1 Les forêts aléatoires (*Random Forest*)

Les forêts aléatoires ont été introduites en 2001 par L.Brieman [7] et constituent, sans nul doute, la méthode d'apprentissage automatique la plus populaire. En effet, les forêts aléatoires sont très fréquemment choisies parmi les méthodes d'apprentissage automatique en raison de la qualité de ses prédictions et constituent bien souvent un modèle de référence.

L'utilisation de cet algorithme dans le secteur assurantiel est de plus en plus répandue, tout particulièrement sur le périmètre IARD. Par exemple, il permet notamment de challenger les résultats issus des modèles linéaires généralisés (MLG), dans le cadre d'une problématique de tarification.

Dans le cadre de la problématique de ce mémoire, cette première méthode permet d'obtenir une charge ultime individuelle en utilisant les informations propres aux sinistres au travers d'un grand nombre de variables explicatives (*features*).

Le principe des forêts aléatoires consiste à agréger des arbres de décision CART supposés indépendants afin d'obtenir un estimateur « moyenné » permettant d'améliorer la robustesse de l'estimation de la variable cible.

Les forêts aléatoires introduisent deux formes d'aléas : le ré-échantillonnage aléatoire (*Bootstrap*) et la sélection d'un sous ensemble aléatoire de variables explicatives (*random features selection*).

Bootstrap Aggregating (Bagging)

Le *bagging* consiste à réaliser un tirage aléatoire avec remise des individus (*bootstrap*) présent dans l'échantillon initial afin de construire un nouvel échantillon (échantillon *In-Bag*) sur lequel sera construit un arbre de décision CART. Ce procédé est répété pour chaque arbre constituant la forêt aléatoire. Les arbres CART sont ensuite agrégés afin de former un nouvel estimateur.

Généralement, l'échantillon construit par *bootstrap* est obtenu à l'aide d'un tirage uniforme sur la base de données initiale mais il est également envisageable d'intégrer des probabilités de sélection différentes selon les individus. De plus, il est également possible de construire les arbres de décision à partir d'un échantillon de taille inférieure à la base de données initiale, ce qui peut s'avérer utile afin de limiter les temps de calculs. Ces deux options sont notamment proposées dans le package R « *ranger* » [2] qui sera utilisé dans ce mémoire pour la construction des forêts aléatoires.

Le fait de construire chaque arbre de la forêt aléatoire à partir d'un échantillon *bootstrap* permet de stabiliser la prédiction obtenue et donc d'augmenter la robustesse du modèle. A noter qu'une stratégie de *bagging* permet uniquement de diminuer la variance de l'estimateur, le biais reste quant à lui inchangé. Dans le support de L.Rouviere [22], le calcul de la variance de l'estimateur baggé est détaillé.

Sélection d'un sous-ensemble aléatoire *(random features selection)*

La deuxième source d'aléa introduite par les forêts aléatoires provient de la sélection d'un sous-ensemble aléatoire venant modifier la construction des arbres CART. Parmi l'ensemble des variables explicatives considérées, seul un sous-ensemble sélectionné par tirage aléatoire est retenu pour réaliser la séparation binaire optimale pour chaque nœud, et ce pour chaque arbre de décision CART constituant la forêt aléatoire. Cette forme d'aléa a pour objectif de construire des arbres de décisions décorrélés assurant la diminution la variance de l'estimateur obtenu.

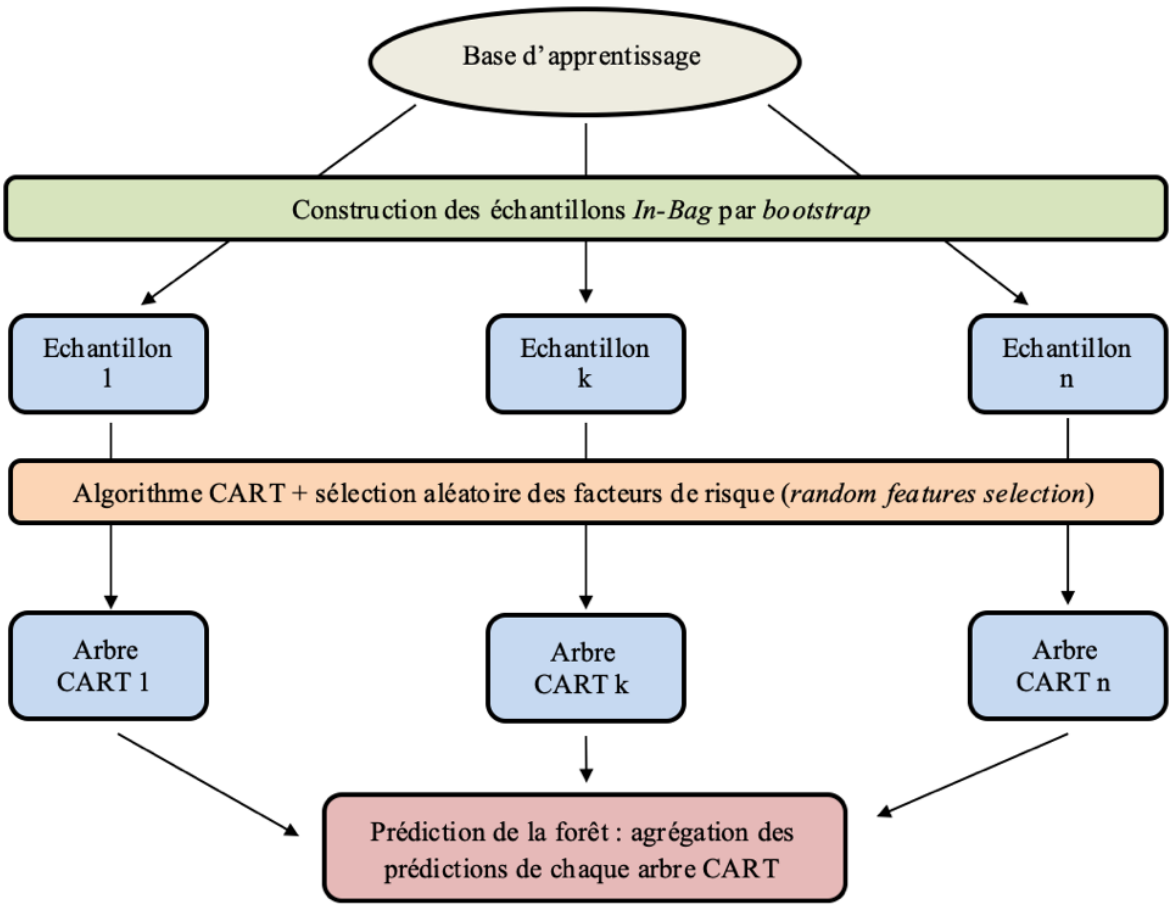


FIGURE 4.1 – Etapes de construction d'une forêt aléatoire

2. https ://cran.r-project.org/web/packages/ranger/ranger.pdf

Le nombre de variables tirées aléatoirement parmi les variables candidates (c'est-à-dire la taille du sous-ensemble aléatoire) est un paramètre de *tuning* de l'algorithme qu'il faudra calibrer au préalable. Le graphique en Figure 4.1 présente les différentes étapes de construction d'une forêt aléatoire.

4.2.2 Le *Gradient Boosting Machine (GBM)*

Le *boosting* a été introduit en 1996 par Y.Freund et R.Schapire [12]. Comme les forêts aléatoires, le *boosting* est également une méthode ensembliste qui consiste à agréger des arbres de décision binaires. La stratégie sous-jacente est toutefois bien différente du *bagging* puisqu'elle repose sur une construction adaptative séquentielle d'estimateurs.

L'idée de l'algorithme consiste à introduire un nouvel arbre de décision venant corriger les défauts du précédent et ce, de façon récursive jusqu'à obtenir un estimateur optimal. Les méthodes de *boosting* sont très largement plébiscitées dans le milieu de la datascience compte tenu de leur forte capacité prédictive. Par ailleurs, elles sont fréquemment utilisées par les gagnants des concours Kaggle et se démocratisent de plus en plus pour la résolution de problématiques assurantielles.

Il existe à ce jour plusieurs algorithmes de *boosting*, tous reposant sur le principe d'amélioration des performances d'un « *weak learner* » dans le but de construire un modèle agrégé robuste. La première version d'algorithmes de *boosting Adaboost* a pour vocation à être utilisé pour résoudre des problèmes de classification.

Etant donné que la projection de la charge à l'ultime est un problème de régression, l'algorithme *Adaboost* ne sera pas présenté. Toutefois, le lecteur intéressé pourra se référer à l'article publié par Y.Freund et R.Schapire [12].

Dans le cadre de ce mémoire, on s'intéressera davantage au *Gradient Boosting*.

Le *Gradient Boosting*

La méthode du Gradient Boosting a été introduite en 2001 par J.Friedman [13] faisant référence à la stratégie de *boosting* mais également à la descente de gradient, une méthode fréquemment utilisée pour approcher les solutions de problèmes d'optimisation.

L'idée d'introduire la notion de gradient est de pouvoir généraliser la stratégie de *boosting* à différents types de problèmes en considérant d'autres fonctions de perte. Afin de faire le lien avec *Adaboost*, C.M Bishop [5] montre que la construction de cet algorithme revient à minimiser une fonction de perte particulière, qui dans ce cas précis est de forme exponentielle.

Grace cette généralisation, la méthode du *Gradient Boosting* peut alors être utilisée dans le cadre d'une problématique de régression, en considérant éventuellement plusieurs fonctions de perte.

Bien que l'idée du *boosting* soit relativement intuitive, l'algorithme du *Gradient Boosting* n'est pas facile à appréhender. L'objectif de cette partie n'est pas de détailler précisément cet algorithme mais d'adopter un formalisme mathématique simplifié afin d'en comprendre le fonctionnement. Cette pré-

sentation s'inspire du support de R.Rakotomalala [21] où les notations seront reprises. Toutefois, le lecteur intéressé par les aspects techniques pourra se référer à l'article de J.Friedman [13].

En considérant un échantillon de taille N et en notant Y la variable cible de notre problème de régression (de taille N) associée à une matrice de variables explicatives notée X de taille $N \times P$ où P est le nombre de prédicteurs. On considère également une séquence de B modèles (« *week learner* ») dénotés par M_b ainsi qu'une fonction de perte L différentiable.

L'algorithme générique du *Gradient Boosting* se présente de la façon suivante :

Initialisation du modèle : $F_0(x) = \underset{a}{\text{argmin}} \sum_{i=1}^{N} L(y_i, a)$

Pour un nombre d'itérations allant de 1 à B :

- Calcul de l'opposé du gradient de la fonction de perte L : $- L(y_i, F_{b-1}(x_i))$
- Construire un arbre de régression M_b sur $- L(y_i, F_{b-1}(x_i))$
- $F_b(x) = F_{b-1}(x) + \eta * M_b(x)$

Fin

FIGURE 4.2 – Présentation de l'algortihme du *Gradient Boosting*

Après initialisation du modèle, l'algorithme du *Gradient Boosting* consiste à approcher de façon itérative le gradient de la fonction de perte L par la construction d'un modèle M_b qui, dans le cadre de ce mémoire, est un arbre de régression CART. A noter qu'il est tout à fait envisageable de considérer d'autres modèles. L'estimation de la variable cible obtenue est ensuite ajoutée à celles issues des modèles précédents en appliquant un coefficient multiplicatif noté η.

Cet hyperparamètre, le *learning rate*, permet de contrôler la vitesse de convergence de l'algorithme vers la solution optimale en pénalisant l'ajout d'un nouveau modèle dans l'estimation finale. L'introduction de ce paramètre est essentielle afin de se prémunir du surapprentissage du modèle et donc d'améliorer son pouvoir prédictif sur un échantillon indépendant.

En plus d'une réduction de variance obtenue par agrégation des arbres de décision, cette stratégie itérative a pour principal avantage de réduire également le biais d'estimation, contrairement aux forêts aléatoires où le biais reste inchangé.

La dernière ligne de cet algorithme présenté en Figure 4.2 fait directement référence à la méthode de descente de gradient. En effet, le modèle M_b est construit à partir de la direction opposée au gradient de la fonction de perte L. La constante d'apprentissage η joue le rôle du pas dans la descente de gradient.

Le choix de la fonction de perte

Avant de passer à l'optimisation du modèle de *Gradient Boosting*, il convient tout d'abord de s'intéresser à la fonction de perte introduite dans l'algorithme. Comme présenté précédemment, chaque arbre de décision CART est construit de façon à approcher le gradient de la fonction de perte. En conséquence, la forme de la fonction de perte considérée conditionne naturellement les prédictions obtenues.

Pour des problématiques de régression, l'erreur quadratique (norme L_2) est souvent privilégiée en raison de sa sensibilité aux valeurs extrêmes.

Cette fonction s'exprime de la façon suivante :

$$l(y, F(x)) = \frac{1}{2}(y - F(x))^2$$

En considérant cette métrique, la fonction de perte considérée classiquement dans un algorithme de *Gradient Boosting* est la suivante :

$$L(y, F) = \sum_{i=1}^{N} l(y_i, F(x_i)) = \frac{1}{2}\sum_{i=1}^{N}(y_i - F(x_i))^2$$

Le gradient de L par rapport à $F(x_i)$ s'écrit :

$$\frac{\partial L}{\partial F(x_i)} = \frac{\partial \sum_{i=1}^{N} l(y_i, F(x_i))}{\partial F(x_i)} = F(x_i) - y_i$$

Dans ce cas précis, le gradient de la fonction de perte correspond à l'opposé du résidu du modèle. En revenant sur l'algorithme du *Gradient Boosting*, chaque arbre de décision est ainsi construit de façon à modéliser les résidus obtenus à l'itération précédente. L'estimation de chaque arbre correspond dans ce cas à la moyenne des résidus des individus présents dans chaque feuille. Cette dernière est ensuite ajoutée aux prédictions issues des modèles précédents, pondérée par le taux d'apprentissage η.

Nous retrouvons ainsi un processus itératif où le nouvel arbre généré vise à répondre aux insuffisances prédictives des arbres précédents. L'introduction d'une fonction de perte quadratique permet également d'appréhender plus facilement le fonctionnement de l'algorithme du *Gradient Boosting*. Une présentation très détaillée est disponible dans le support de R.Rakotomalala [21].

A noter que d'autres fonctions de perte peuvent également être utilisées, comme par exemple l'erreur en valeur absolue (norme L_1) :

$$m(y, F(x)) = |y - F(x)|$$

D'un point de vue pratique, l'algortihme du *Gradient Boosting* est implémentée pour ce mémoire à l'aide du package « *caret* » [3] du logiciel R.

3. https ://cran.r-project.org/web/packages/caret/caret.pdf

4.3 Optimisation des hyperparamètres *(tuning)*

Les forêts aléatoires et le *Gradient Boosting* présentent des hyperparamètres qu'il faut au préalable calibrer afin d'optimiser leurs performances. Précisions tout de même que la qualité prédictive d'un modèle est mesurée par sa capacité à se généraliser sur un jeu de données indépendant, c'est à dire qui n'a pas été utilisé lors de sa construction. On cherche donc à calibrer les différents modèles de telle sorte à minimiser les erreurs de prédictions sur la base test.

Rappelons que les hyperparamètres sont des paramètres fixés par l'utilisateur en amont de l'apprentissage du modèle. A l'inverse, les valeurs des paramètres inhérents au modèle sont déterminées lors de sa construction, généralement par optimisation.

Les méthodes d'apprentissage automatique sont à ce jour fréquemment utilisées dans le cadre de problématiques assurantielles. En revanche, le calibrage des hyperparamètres est une étape qui n'est pas toujours abordée dans la littérature actuarielle, les valeurs par défaut sont souvent utilisées.

Ces valeurs par défaut sont définies de telle sorte à obtenir en général des résultats relativement satisfaisants. Néanmoins, les valeurs optimales des hyperparamètres dépendent du problème étudié et de la base de données, ils sont donc à calibrer au cas par cas.

Compte tenu du fait qu'il n'existe pas à ce jour de résultats théoriques sur le calibrage optimal de ces hyperparamètres, l'objectif de cette partie est donc de proposer une méthode qui permettrait d'optimiser le pouvoir de généralisation des modèles.

Les deux stratégies présentées ici reposent sur une recherche par quadrillage (*grid search*). Cette méthode consiste à définir un ensemble d'hyperparamètres puis à évaluer sur un échantillon indépendant la qualité prédictive du modèle pour chacune des combinaisons considérées. Une variante de cette méthode (*random search*) consiste à ne sélectionner qu'une partie des combinaisons d'hyperparamètres par tirage aléatoire.

La recherche par quadrillage est fréquemment utilisée pour le calibrage des hyperparamètres des modèles d'apprentissage automatique et ce, malgré qu'elle soit relativement coûteuse en temps de calculs. D'autres méthodes comme l'optimisation Bayésienne permet, en autre, de pallier à ce problème en sélectionnant les combinaisons d'hyperparamètres ayant le plus grand potentiel de diminution de la fonction de perte.

L'avantage d'opérer une recherche par quadrillage est également de pouvoir visualiser l'impact marginal de chacun des hyperparamètres sur la qualité des résultats obtenus, facilitant par conséquent la compréhension des algorithmes de *Machine Learning* considérés.

4.3.1 Cas des forêts aléatoires

Pour les forêts aléatoires, nous dénotons trois principaux hyperparamètres présents dans la fonction « *ranger* » du package du même nom :

— *num.trees* : correspond au nombre d'arbres de décision construit dans la forêt aléatoire. Une valeur relativement élevée de ce paramètre permet d'obtenir de meilleures performances ainsi qu'une plus grande stabilité des prédictions. Toutefois, un nombre d'arbres de décision important entraîne des temps de calcul relativement longs. Ce paramètre ne possède pas de valeur par défaut, il devra alors être fixé afin d'optimiser les performances du modèle tout en veillant à ce que les temps de calculs restent convenables.

— *mtry* : correspond au nombre de facteurs de risque candidats à la construction de chaque nœud des arbres de décision. Comme évoqué lors de la présentation des forêts aléatoires, ce paramètre permet de contrôler la corrélation des arbres construits. Un nombre important de variables candidates entraînera potentiellement une instabilité des prédictions de la forêt. A l'inverse, le pouvoir prédictif de la forêt construite pourrait être réduit.
La valeur par défaut proposée par la fonction « *ranger* » correspond à la racine carrée du nombre de facteurs de risque.

— *min.node.size* : correspond au nombre minimum d'individus nécessaires dans un nœud pour réaliser une séparation supplémentaire. Ce paramètre permet de contrôler la profondeur des arbres de régression construits.
Par défaut, ce paramètre est fixé à cinq pour un problème de régression [4].

Comme évoqué au sein des précédentes parties, la stratégie de *bagging* permet de diminuer la variance de l'estimateur construit, le biais reste quant à lui inchangé.
Pour cette raison, le nombre minimum d'individus à la création d'une séparation supplémentaire sera fixé à 1 afin de construire des arbres maximaux permettant de minimiser le biais de l'estimateur obtenu.

En conséquence, seuls les paramètres *num.trees* et *mtry* seront calibrés afin d'optimiser le pouvoir prédictif des forêts aléatoires.

4. Pour un problème de classification, ce paramètre est fixé par défaut à 1.

Méthode d'optimisation des hyperparamètres

Les deux hyperparamètres *num.trees* et *mtry* doivent être calibrés sur un échantillon indépendant afin d'optimiser le pouvoir de généralisation du modèle.

Dans le cas des forêts aléatoires, chaque arbre de décision est construit à partir d'un échantillon *In-Bag* obtenu à l'aide d'un tirage aléatoire avec remise sur la base d'apprentissage. L'échantillon *Out-Of-Bag (OOB)* n'a par conséquent pas servi à la construction de l'arbre de décision, il peut donc servir à la calibration des hyperparamètres du modèle.

La stratégie d'optimisation retenue pour les forêts aléatoires consiste naturellement à sélectionner le couple d'hyperparamètres *(num.trees, mtry)* minimisant l'erreur de prédiction sur l'ensemble des échantillons *Out-Of-Bag* de la forêt aléatoire.

L'erreur de prédiction est ici mesurée à l'aide du RMSE (*Root Mean Square Error*) :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

Ce processus est répété pour chacun des 5 groupes introduits dans la partie 3.3.3, les prédictions sont ensuite confrontées aux observations des bases test respectives.

Application au premier modèle de projection

Nous proposons d'illustrer le processus d'optimisation des hyperparamètres des forêts aléatoires en considérant le modèle servant à projeter les sinistres du 1er groupe, c'est-à-dire les sinistres déclarés en 2017 toujours ouverts en gestion au 31 décembre de la même année.

En appliquant une recherche par quadrillage (*grid search*), nous définissons une séquence de valeurs pour chaque hyperparamètre sur laquelle l'algorithme des forêts aléatoires est évalué : de 5 à 15 pour le nombre de variables candidates à chaque séparation des arbres CART *(mtry)*, de 500 à 1 700 par pas de 50 pour le nombre d'arbres de la forêt *(num.trees)*.

Pour rappel, ce modèle est construit à partir de 48 182 sinistres décrits par 32 variables explicatives. La valeur proposée par défaut pour l'hyperparamètre *mtry* est donc de 6 ($\sqrt{p}$ où p est le nombre de variables explicatives).

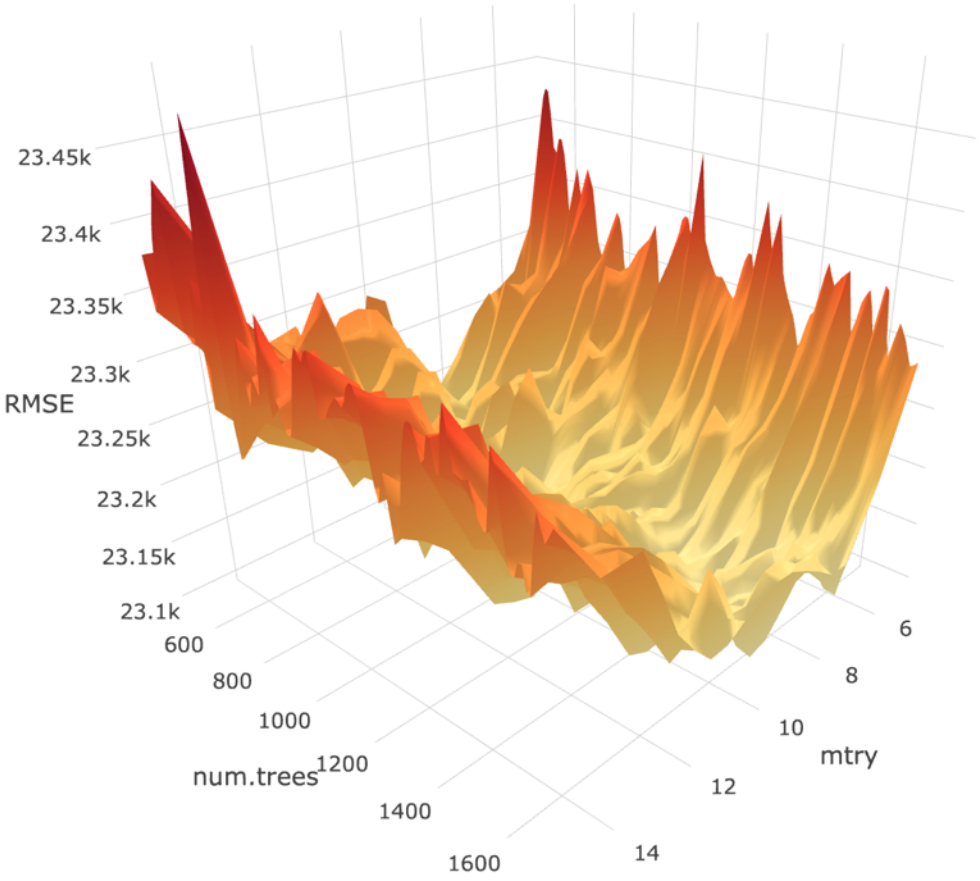


FIGURE 4.3 – RMSE sur l'échantillon *Out-Of-Bag* en fonction du nombre d'arbres *(num.trees)* et du nombre de variables candidates *(mtry)*

La Figure 4.3 illustre la valeur du RMSE calculé sur l'échantillon *Out-Of-Bag* pour chaque combinaison du couple *(num.trees, mtry)*. L'hyperparamètre *mtry* impacte significativement le RMSE obtenu, sa valeur doit donc être choisie judicieusement. Comme l'illustre la Figure 4.3, le RMSE est minimale lorsque *mtry* est fixé à **8**.
Un nombre d'arbres élevé permet de se focaliser uniquement sur des forêts aléatoires suffisamment robustes, obtenues par minimisation de la variance (propriété du *bagging*). C'est ce choix qui a été retenu pour cette optimisation.

Une coupe de la surface représentée en Figure 4.3 est réalisée afin de déterminer le nombre d'arbres optimal lorsque l'hyperparamètre *mtry* est fixé à 8.

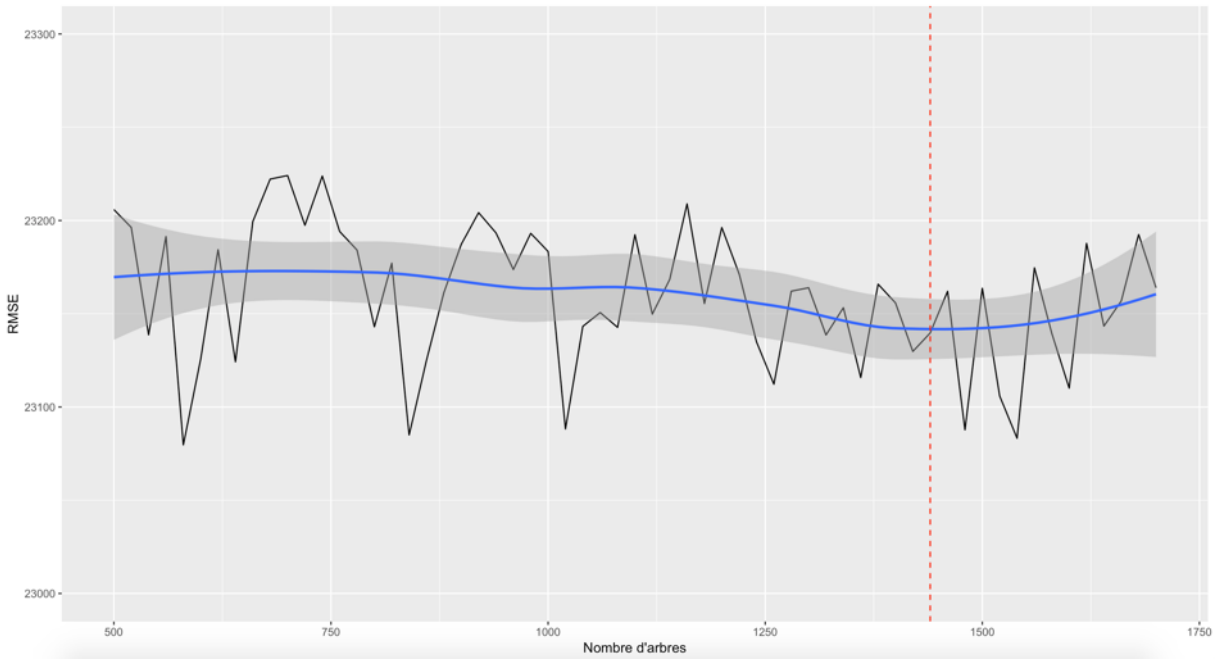


FIGURE 4.4 – RMSE en fonction du nombre d'arbres pour $mtry = 8$

Malgré que le RMSE soit relativement volatile, sa valeur minimale semble être obtenue pour un nombre d'arbres fixé à **1 440** (ligne en pointillé rouge), c'est cette valeur que nous retenons pour le paramètre *num.trees*. Au delà, le RMSE tend à augmenter, l'algorithme rentre dans une situation de surapprentissage (*overfitting*). Il ne semble donc pas pertinent de considérer des forêts aléatoires composée de plus de 1 750 arbres.

Au final, le couple d'hyperparamètres ***(num.trees, mtry)* = (1 440, 8)** est retenu pour projeter les sinistres déclarés en 2017 à l'aide de l'algorithme des forêts aléatoires.

4.3.2 Cas du *Gradient Boosting Machine (GBM)*

L'algorithme du *Gradient Boosting Machine* possède davantage d'hyperparamètres que les forêts aléatoires. L'optimisation de ces derniers est par conséquent plus complexe et nécessite plusieurs étapes.

Tout d'abord, parmi les nombreux hyperparamètres présents dans le *Gradient Boosting*, seulement 4 seront optimisés :

— *shrinkage* : le taux d'apprentissage permet de contrôler la vitesse de convergence de l'algorithme vers la solution optimale en pénalisant l'ajout d'un nouveau modèle dans l'estimation finale.

— *n.trees* : correspond au nombre d'itérations considérées (i.e le nombre d'arbres CART construits). Pour un taux d'apprentissage donné, un grand nombre d'itérations améliore la robustesse du modèle en réduisant la variance de l'estimateur mais peut également conduire au surapprentissage.
Le couple d'hyperparamètres *(shrinkage,n.trees)* doit donc être calibré conjointement afin d'optimiser le pouvoir de généralisation du modèle mesuré sur la base test.

— *interaction.depth* : correspond au nombre de séparations (split) réalisées dans chaque arbre de décision. Le nombre de feuilles (*leaves*) obtenu est égal à *interaction.depth* + 1. Cet hyperparamètre permet de contrôler la profondeur des arbres construits, une valeur élevée peut conduire au surapprentissage du modèle.

— *n.minobsinnode* : correspond au nombre minimum d'individus présents dans chaque feuille de l'arbre. Ce dernier hyperparamètre permet également de contrôler le surapprentissage du modèle.

Les hyperparamètres en bleu interviennent dans la construction de l'algorithme de *boosting* alors que les hyperparamètres en vert définissent les caractéristiques des arbres CART construits.

L'hyperparamètre $bag.fraction$ permet de sélectionner un sous-échantillon de la base d'apprentissage de taille $bag.fraction \times N$ [5] utilisé pour construire chaque de décision. Il permet d'introduire une source d'aléa dans l'algorithme, on parlera alors de *Stochastic Gradient Boosting* [14]. Une valeur inférieure à 1 améliore la robustesse du modèle, cet hyperparamètre sera fixé à 0.8 [6].

Le risque de surapprentissage est omniprésent lors de la construction d'un modèle de *boosting* et doit être réduit au maximum afin d'augmenter son pouvoir de généralisation sur un nouveau jeu de données. En effet, contrairement aux forêts aléatoires, le *boosting* permet de réduire le biais de l'estimateur final au risque d'apprendre des interactions spécifiques à la base d'apprentissage.

Comme pour les forêts aléatoires, les valeurs optimales des hyperparamètres sont obtenues par minimisation du RMSE.

Méthode d'optimisation des hyperparamètres

Pour le *Gradient Boosting*, la méthode d'optimisation considérée dans ce mémoire s'inspire très largement des approches utilisées par les participants aux concours Kaggle.

Contrairement aux forêts aléatoires, les hyperparamètres sont ici optimisés par validation croisée, à savoir la *K-fold Cross Validation*.

Cette méthode d'échantillonnage consiste à diviser la base d'apprentissage en K échantillons de même taille. Parmi ces échantillons, $K-1$ sont sélectionnés pour construire le modèle, l'échantillon restant est utilisé en tant que base de validation. L'erreur de prédiction est estimée en moyennant les K erreurs quadratiques moyennes.

La Figure 4.5 illustre la *K-fold Cross Validation* pour $K = 5$, la valeur choisie pour calibrer les hyperparamètres du modèle.

5. N étant la taille de la base d'apprentissage.
6. Cette valeur permet en général d'obtenir de bons résultats

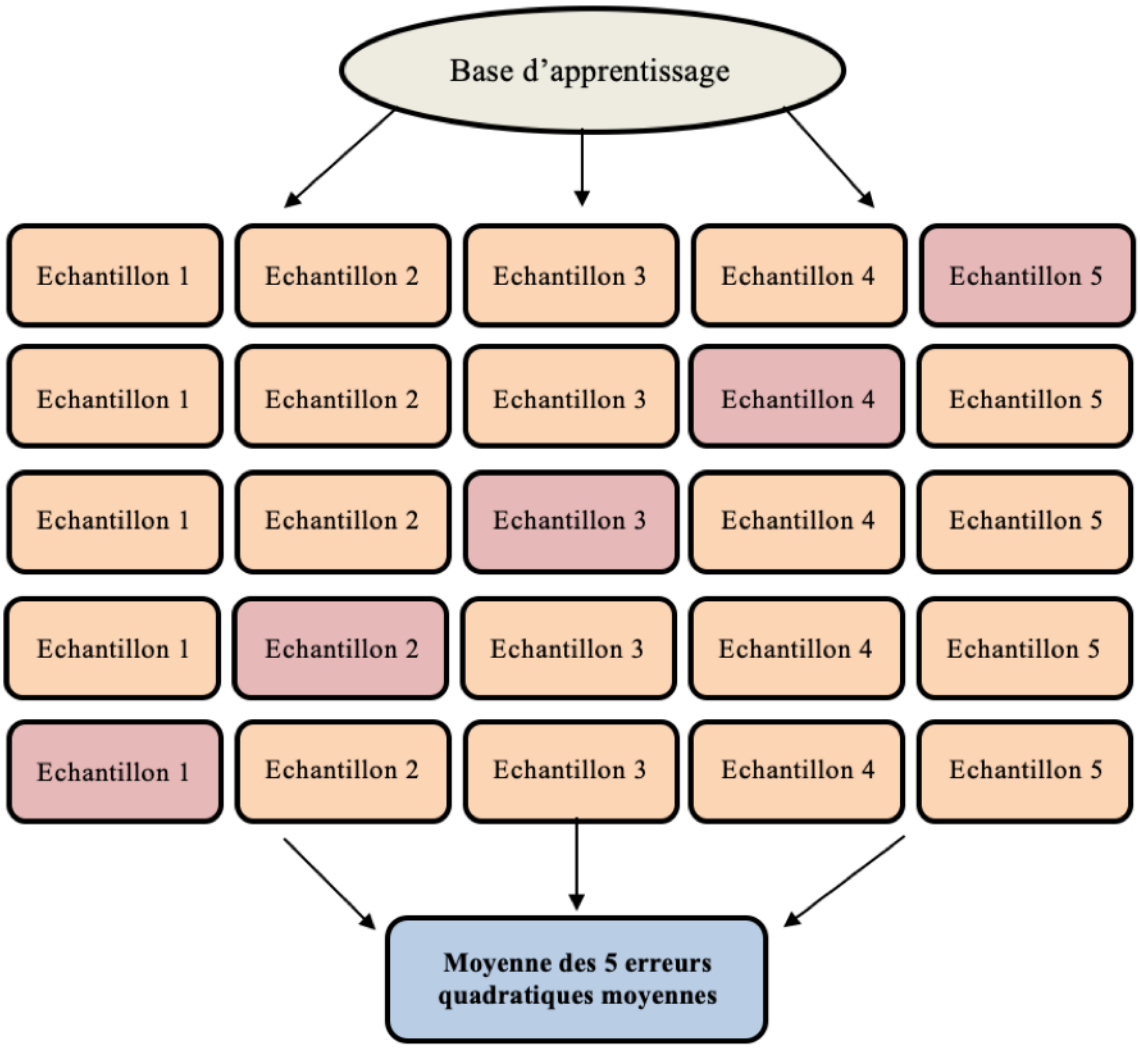


FIGURE 4.5 – Illustration de la *K-fold Cross Validation* pour $K = 5$

Concernant l'algorithme de *boosting*, les hyperparamètres sont calibrés en suivant les étapes suivantes :

1. Initialisation du nombre d'itérations *(n.trees)* :
   On fixe un taux d'apprentissage *(shrinkage)* suffisamment grand de telle sorte à ce que l'erreur de validation minimale soit obtenu pour un nombre d'itérations relativement faible (entre 50 et 100 itérations). Les valeurs initiales du couple *(shrinkage, n.trees)* doit permettre de tester rapidement plusieurs valeurs des hyperparamètres relatifs aux arbres de décision construits *(interaction.depth, n.minobsinnode)*.
   Pour le calibrage de l'ensemble des modèles de *boosting*, le *shrinkage* est fixé initialement à 0.1.

2. Calibrage du couple *(interaction.depth, n.minobsinnode)* :
   Pour le couple *(shrinkage, n.trees)* fixé à l'étape 1, les hyperparamètres *(interaction.depth, n.minobsinnode)* sont calibrés en minimisant l'erreur de validation croisée. L'hypothèse sous-jacente consiste à supposer que le couple optimal de ces deux hyperparamètres sera le même quel que soit le taux d'apprentissage et le nombre d'itérations considérés. Cette hypothèse a été vérifiée empiriquement.

3. Calibrage du couple *(shrinkage, n.trees)* :
   Une fois les hyperparamètres *interaction.depth* et *n.minobsinnode* calibrés, on diminue le taux d'apprentissage tout en augmentant le nombre d'itérations afin d'augmenter la robustesse du modèle construit.

Comme pour les forêts aléatoires, ce processus d'optimisation des hyperparamètres est répété pour chacun des 5 groupes introduits dans la partie 3.3.3, les prédictions sont ensuite confrontées aux observations des bases test respectives.

Application au premier modèle de projection

Comme pour les forêts aléatoires, la méthode d'optimisation des hyperparamètres du modèle de *boosting* est appliquée au modèle servant à projeter les sinistres en-cours au 31/12/2017 déclarés au cours de la même année.

Etape 1 : Initialisation du nombre d'itérations :

La première étape consiste à déterminer un nombre d'itérations permettant d'évaluer en un temps de calculs raisonnable différentes valeurs des autres hyperparamètres. Cette valeur est obtenue de telle sorte à ce que la valeur minimale du RMSE soit minimale pour un nombre d'itérations compris entre 50 et 100. Le taux d'apprentissage permet de contrôler la convergence de l'algorithme du *Gradient Boosting*. Celui-ci est fixé à 0,1 de façon à approcher rapidement la solution optimale (pour la combinaison d'hyperparamètres définie initialement).

Nous testons un nombre d'itérations variant de 10 à 150 par pas de 5, les résultats obtenus sont présentés en Figure 4.6.

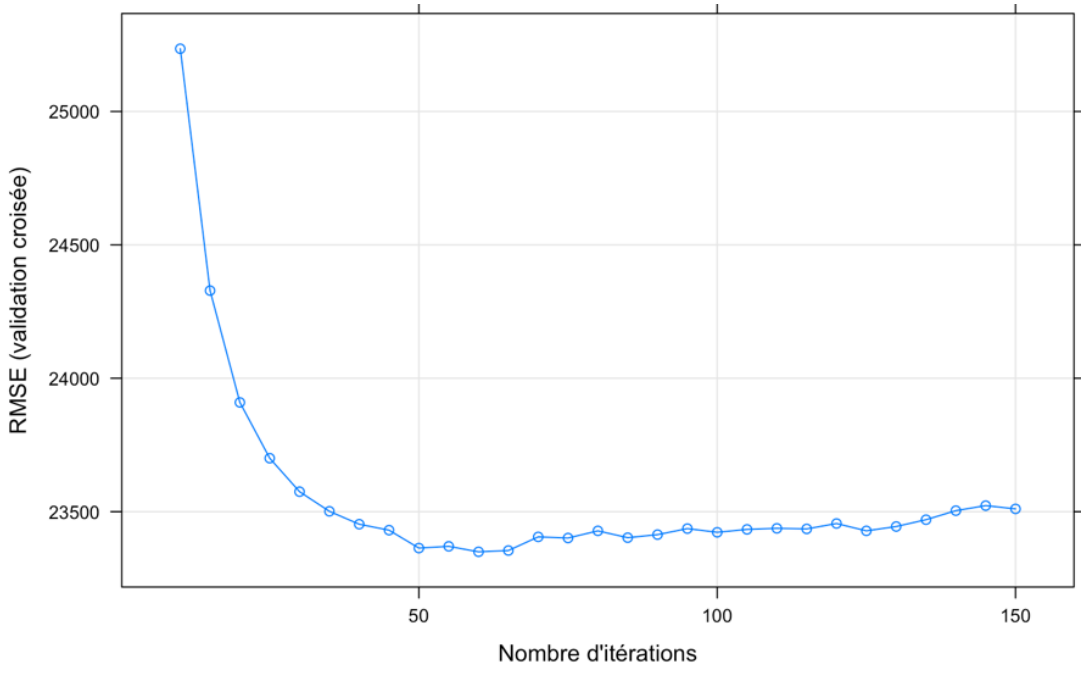


FIGURE 4.6 – Initialisation du nombre d'itérations

Dans un premier temps, l'erreur de validation croisée diminue lorsque le nombre d'itérations augmente puis se stabilise lorsque cet hyperparamètre est fixé à 60. Au delà, la qualité prédictive du modèle se dégrade. La Figure 4.6 illustre le phénomène de surapprentissage de l'algorithme.

Pour le calibrage les hyperparamètres relatifs à la construction des arbres CART *(interaction.depth, n.minobsinnode)*, le nombre d'itérations est fixé à **60**.

Etape 2 : Calibrage du couple *(interaction.depth, n.minobsinnode)* :

Après avoir déterminé le nombre initial d'itérations, la prochaine étape consiste à calibrer les hyperparamètres *(interaction.depth, n.minobsinnode)*. Ils déterminent la complexité des arbres CART construits.

Contrairement aux forêts aléatoires, les arbres construits sont bien souvent moins profonds pour éviter le surapprentissage de l'algorithme. Nous faisons varier le nombre maximal de séparation de chaque arbre *(interaction.depth)* entre 3 et 16 et le nombre d'observations minimum dans chaque feuille entre 0 et 150 par pas de 25. Les résultats sont présentés sur la Figure 4.7.

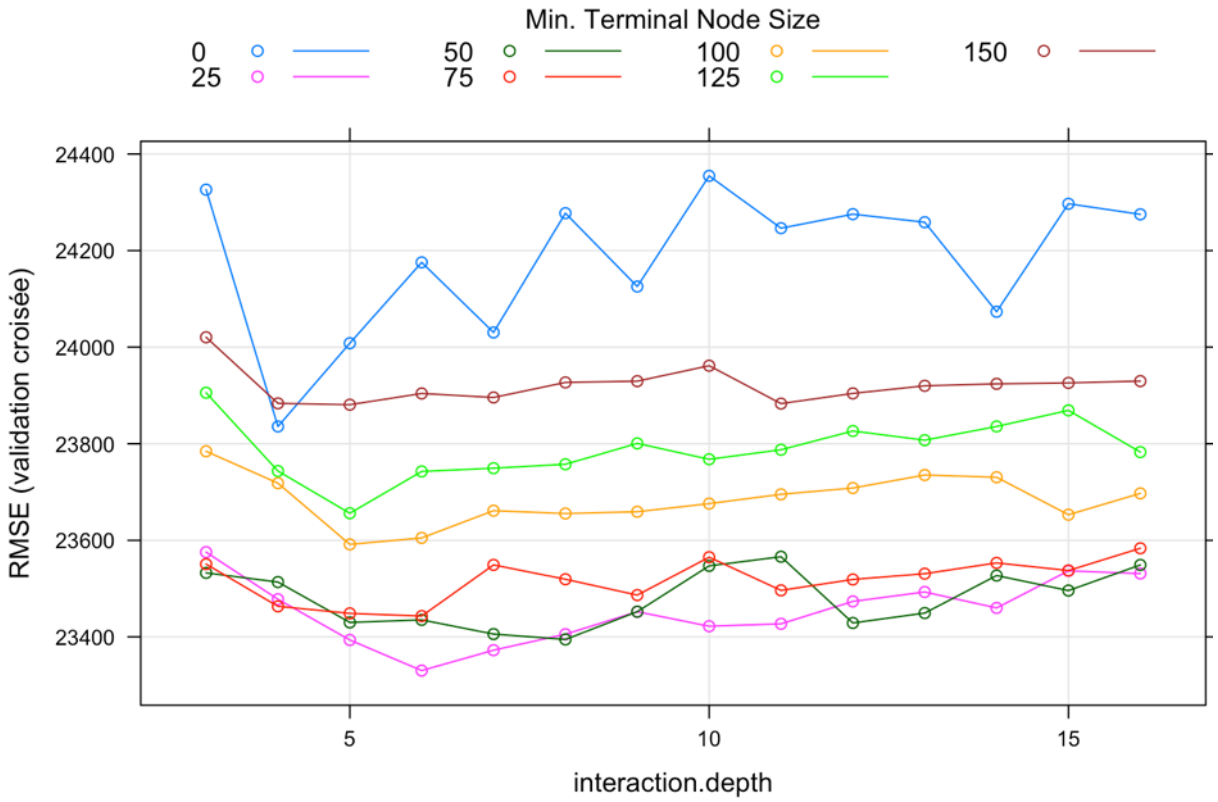


FIGURE 4.7 – Calibrage du couple *(interaction.depth,n.minobsinnode)*

Un nombre minimum d'observations dans chaque feuille fixé à 0 entraîne une valeur du RMSE relativement élevée variant sensiblement en fonction de la profondeur des arbres construits. Dans ce cas précis, la qualité prédictive du modèle dépend fortement de la constitution des cinq blocs utilisés pour opérer la validation croisée. Cet effet n'est pas souhaitable pour déterminer les valeurs optimales des deux hyperparamètres.

Le RMSE semble minimal lorsque *n.minobsinnode* vaut **25** et lorsque l'hyperparamètre *interaction.depth* est fixé à **6**. Ce sont ces deux valeurs que nous retenons pour la suite de l'optimisation.

Etape 3 : Calibrage du couple *(shrinkage, n.trees)* :

La dernière étape consiste à calibrer le taux d'apprentissage et le nombre d'itérations. Contrairement aux hyperparamètres relatifs à la construction des arbres CART, le couple *(shrinkage, n.trees)* définissent la structure du modèle de *Gradient Boosting* et sont à calibrer simultanément puisque la valeur optimal de l'un est fortement dépendante de la valeur de l'autre.

A l'étape 1, le taux d'apprentissage a été fixé volontairement à une valeur relativement élevée afin d'obtenir un RMSE minimal pour un faible nombre d'itérations. Pour augmenter la robustesse du modèle construit, on cherche à diminuer le taux d'apprentissage de façon à minimiser le RMSE pour un nombre élevé d'itérations. Par ce biais, nous minimisons la variance de l'estimateur construit par agrégation des arbres CART.

Après avoir réalisé des tests au préalable, deux séquences sont construites : le taux d'apprentissage varie de 0,004 à 0,014 par pas de 0,02 tandis que le nombre d'itérations varie quant à lui de 300 à 1500 par pas de 50.

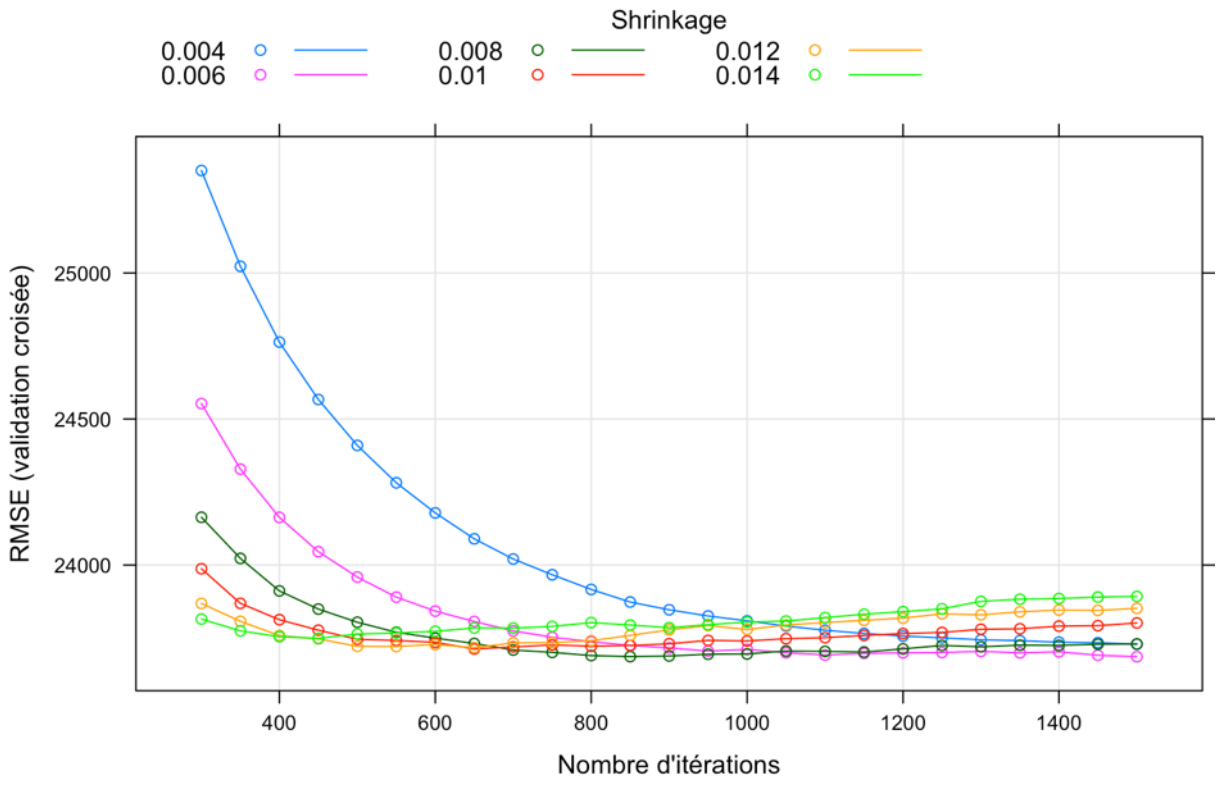


FIGURE 4.8 – Calibrage du couple *(shrinkage,n.trees)*

Pour un taux d'apprentissage fixé à 0,014, le modèle rentre en surapprentissage à partir de 450 itérations. Pour une valeur plus élevée, le RMSE est strictement croissant avec le nombre d'itérations. A partir de ce seuil, le modèle apprend des itérations spécifiques à la base d'apprentissage venant potentiellement diminuer sa qualité prédictive sur un nouveau jeu de données. Au contraire, pour un

taux d'apprentissage fixé à 0,004, le RMSE minimal ne semble toujours pas être atteint pour 1 600 itérations (sous-apprentissage du modèle).

Pour une valeur fixée à **0,006**, le RMSE se stabilise à partir de **1 200** itérations. Ce sont les deux valeurs que nous retenons pour le couple *(shrinkage, n.trees)*.

Au final, les valeurs optimales des quatre hyperparamètres utilisées pour le modèle de projection des sinistres en-cours déclarés en 2017 sont les suivantes :

	n.minobsinnode	*interaction.depth*	*n.trees*	*shrinkage*
Valeurs optimales	25	6	1 200	0,006

FIGURE 4.9 – Valeurs optimales des hyperparamètres du *Gradient Boosting* pour le 1er modèle de projection

4.4 Résultats de la méthode de projection individuelle

Pour rappel, notre méthode de projection individuelle nécessite de construire un modèle distinct en fonction de la profondeur d'historique des sinistres en-cours au 31/12/2017. En adoptant un pas de temps annuel, l'application de notre méthode consiste à projeter les sinistres en fonction de leur année de déclaration.

Le processus de calibrage des hyperparamètres de chaque algorithme est répété pour chaque modèle. Les valeurs optimales sont représentées en annexe D de ce mémoire.

Dans cette partie, nous analysons les erreurs de prédictions individuelles commises sur chacune des bases test à l'aide des trois indicateurs suivants (le RMSE étant déjà introduit dans la partie 4.3.1) :

En notant :

- y_i le coût observé du sinistre i
- $\overline{y}$ la moyenne des coûts observés
- $f(x_i)$ l'estimation du coût du sinistre i
- N le nombre d'observations

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

Les indicateurs de performance des deux algorithmes de *Machine Learning* sont également comparés aux valeurs obtenues en utilisant la dernière charge dossier/dossier connue comme prédiction du coût ultime.

Cette comparaison nous permet de justifier l'intérêt de notre méthode de projection individuelle. En effet, dans le cas où la charge dossier/dossier est suffisamment précise pour estimer le coût ultime des sinistres, la construction d'un modèle prédictif ne semble pas nécessaire.

Pour faciliter la lecture graphique des résultats, *RF* désignera l'algorithme des forêts aléatoires, *GBM* le *Gradient Boosting Machine* et *Charge D/D* les estimations obtenues à l'aide de la dernière charge dossier/dossier connue.

Après avoir jugé de la performance de la méthode de projection individuelle à l'aide de ces trois indicateurs, les erreurs de prédictions seront analysées individuellement. Ensuite, les variables les plus influentes dans la construction des modèles selon l'algorithme utilisé seront présentées.

4.4.1 Comparaison des indicateurs de performance sur les bases test

La valeur des trois indicateurs (RMSE, R^2 et le MAE) obtenus pour chaque modèle sur chacune des bases test sont présentés en Figure 4.10.

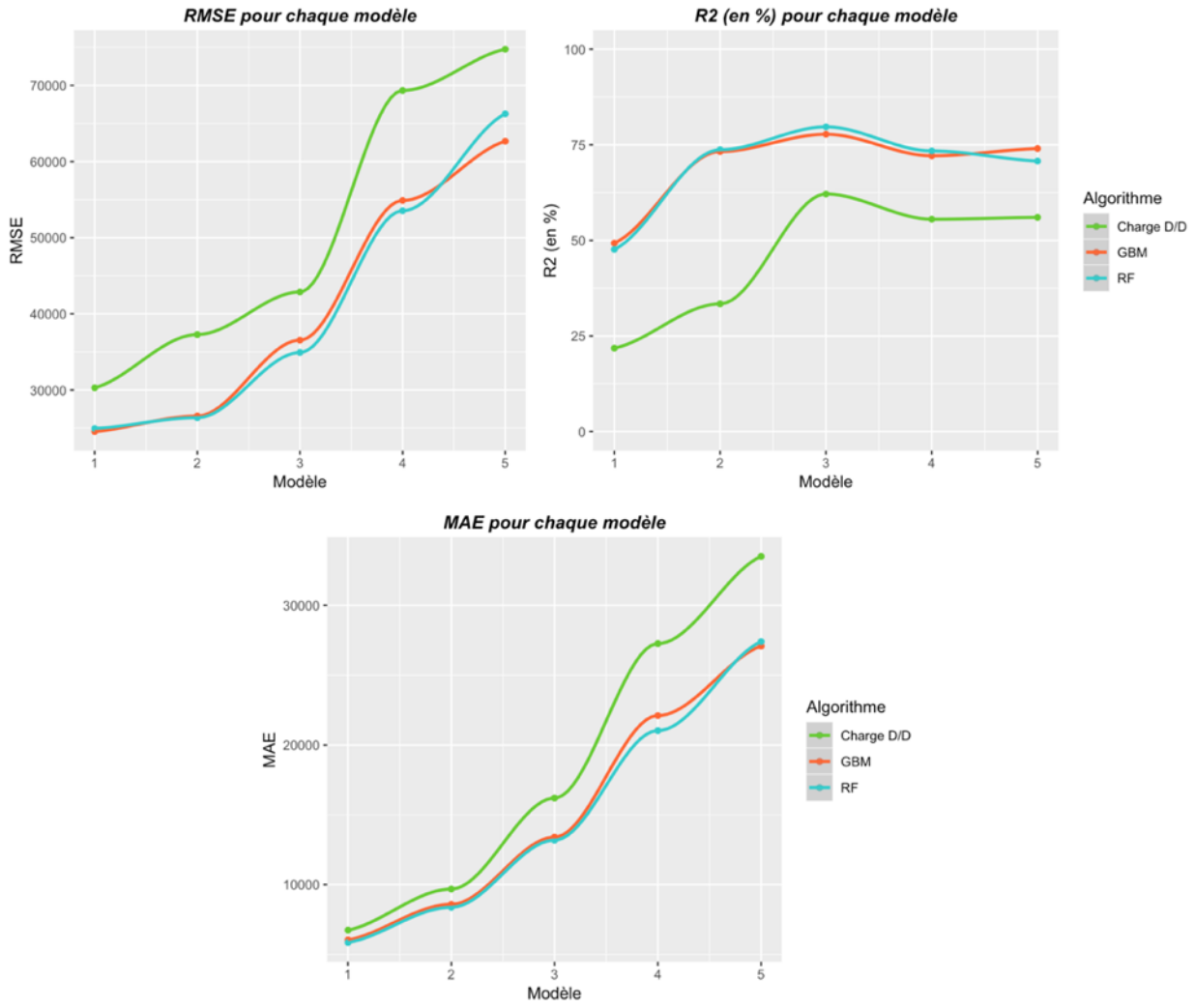


FIGURE 4.10 – RMSE/R^2/MAE obtenus pour chaque modèle

Tout d'abord, nous pouvons commenter l'évolution du RMSE ainsi que du MAE selon le modèle construit. Ces deux indicateurs deviennent de plus en plus élevés à mesure où la connaissance des sinistres s'améliore (passage du modèle 1 à 5). Les erreurs de prédictions commises sont en moyenne plus importantes et plus hétérogènes lors de la projection à l'ultime des sinistres ouverts en gestion depuis plusieurs années que pour ceux déclarés plus récemment.

Ce constat doit cependant être nuancé au regard de la relation croissante entre le coût des sinistres et leur durée de règlement que nous avions observé lors de la présentation de la base d'étude (partie 2.8.1). De plus, contrairement au R^2, l'analyse du RMSE et du MAE ne permet pas de comparer la qualité d'ajustement des modèles compte tenu du fait qu'ils sont construits sur 5 bases test distinctes.

La deuxième observation concerne l'évolution de la qualité d'ajustement mesurée par le R^2. Elle s'améliore nettement à partir du deuxième modèle. Pour les sinistres observés après au moins deux années de développement, la prédiction du coût ultime issue de la méthode de projection individuelle semble relativement précise avec un R^2 supérieur à 70%. Cet indicateur peut même atteindre 80% pour le troisième modèle en utilisant les forêts aléatoires.

L'estimation du coût ultime se révèle être plus délicate lorsque le sinistre est observé au maximum un an après son ouverture. En effet, le R^2 du 1er modèle est inférieur à 50% pour les deux algorithmes, témoignant d'une qualité d'ajustement modérée. Nous disposons peu d'informations sur ces sinistres ce qui explique en toute logique des erreurs d'estimation plus importantes que pour les autres modèles.

D'autre part, l'estimation réalisée à partir de la dernière charge dossier/dossier connue du sinistre est nettement moins précise que celle obtenue à l'aide des deux algorithmes de *Machine Learning*, quel que soit le modèle de projection. Compte tenu de la valeur de ces trois indicateurs de performance, la construction de notre méthode de projection individuelle est totalement justifiée.

Pour davantage de lisibilité, le Tableau 4.1 et le Tableau 4.2 présentent respectivement les valeurs des trois indicateurs (RMSE, R^2 et le MAE) obtenues pour chaque modèle implémenté à l'aide des forêts aléatoires et du *Gradient Boosting*.

Modèle	*1*	*2*	*3*	*4*	*5*
RMSE	24 940	26 379	34 922	53 584	66 348
R^2	47,6%	73,7%	79,7%	73,4%	70,7%
MAE	5 856	8 369	13 186	20 994	27 431

TABLE 4.1 – Résultats des forêts aléatoires

Modèle	*1*	*2*	*3*	*4*	*5*
RMSE	24 543	26 613	36 550	54 856	62 517
R^2	49,3%	73,2%	77,8%	72,1%	74,0%
MAE	6 048	8 589	13 368	22 075	27 107

TABLE 4.2 – Résultats du *Gradient Boosting*

En termes de RMSE et de R^2, les forêts aléatoires prédisent avec une plus grande précision le coût ultime des sinistres pour les modèles 2, 3 et 4. Ce constat semble au premier abord surprenant compte tenu de la capacité du *boosting* à réduire à la fois le biais et la variance des estimations tandis que le *bagging* réduit uniquement la variance. On pouvait s'attendre à obtenir de meilleurs résultats avec le *Gradient Boosting* qu'avec les forêts aléatoires.

Cette intuition doit toutefois être nuancée en fonction des données utilisées et de la problématique étudiée. L'algorithme des forêts aléatoires peut dans certains cas générer de meilleures prédictions que le *Gradient Boosting*.

D'autre part, toujours en termes de RMSE et de R^2, l'algorithme du *Gradient Boosting* est plus précis pour les modèles 1 et 5. A l'exception du modèle 5, les forêts aléatoires affichent un MAE plus faible traduisant une erreur de prédiction (en valeur absolue) moyenne légèrement moins importante que celle commise par le *Gradient Boosting*.

4.4.2 Analyse des erreurs de prédictions individuelles

Après analyse du RMSE, du R^2 et du MAE obtenus sur les bases test, il semble pertinent de s'intéresser aux erreurs de prédiction individuelles.

Dans un premier temps, une comparaison entre les coûts ultimes prédits et observés est réalisée afin d'observer la qualité prédictive des deux algorithmes selon le modèle de projection.

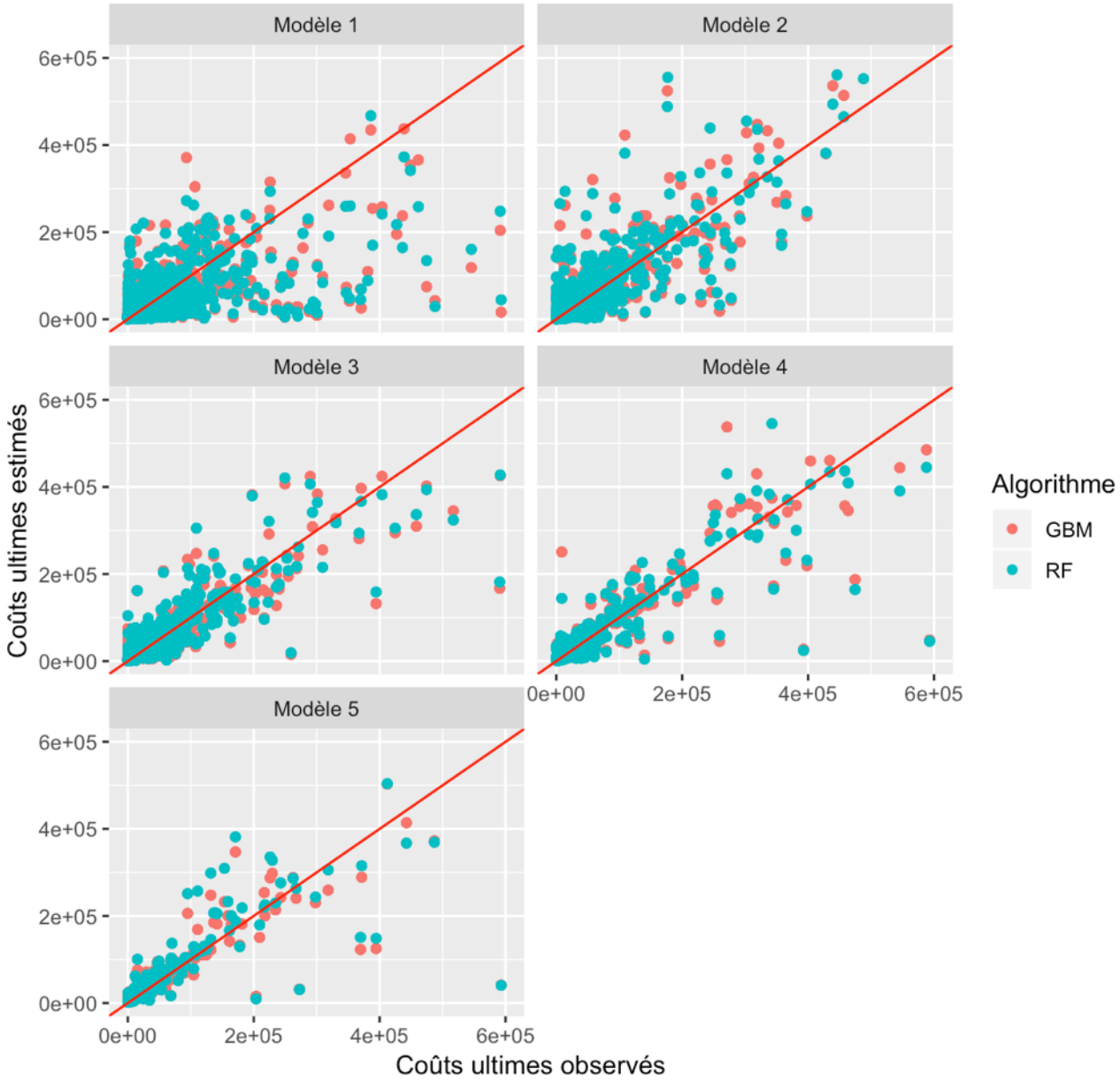


FIGURE 4.11 – Coûts ultimes prédits VS observés pour chaque modèle

Pour une meilleure lisibilité, seuls les sinistres dont le coût ultime est inférieur à 600 000 euros sont intégrés dans la représentation graphique en Figure 4.11. La droite rouge représente la première bissectrice, les sinistres situés sur celle-ci correspondent aux cas où le coût ultime estimé est égal au coût ultime observé.

Le modèle de projection des sinistres à la fin de la première année de développement (modèle 1) a tendance à sous-estimer le coût ultime des sinistres dépassant les 200 000 euros, et ce quel que soit l'algorithme utilisé.
En se focalisant sur les sinistres de moins de 200 000 euros, nous retrouvons une proportion non négligeable de sinistres surestimés et sous-estimés (situés au-dessous et en-dessus de la première bissectrice). Même si d'un point de vue individuel, le premier modèle de projection procure une qualité d'estimation modérée, les surestimations et sous-estimations sont vraisemblablement de même ampleur. Les erreurs de prédictions semblent se compenser, traduisant une bonne estimation du coût ultime des sinistres de moins de 200 000 euros à un niveau agrégé.

En accord avec l'analyse du R^2 réalisée précédemment, l'estimation du coût ultime semble nettement s'améliorer pour les sinistres observés après deux années de développement. En effet, nous retrouvons davantage de sinistres autour de la première bissectrice. Notons tout de même que la sous-estimation quasi systématique des sinistres supérieurs à 200 000 euros n'est plus présente dans les modèles 2 à 5. L'actualisation des informations sur les sinistres telle que la mise à jour de l'évaluation dossier/dossier ou bien la prise en compte des règlements passés améliorent la prédiction de l'ultime. Les erreurs de prédictions pour ces sinistres supérieurs à 200 000 euros restent toutefois très volatiles même après les avoir observés sur 4 années de développement.
Après 5 années de développement (modèle 5), les prédictions semblent globalement proches du coût ultime même si certains sinistres restent mal estimés.

En résumé, le coût ultime des sinistres en-cours à la fin de la première année de développement reste délicat à estimer. A ce stade de développement, les informations dont nous disposons sur les sinistres ne suffisent à estimer précisément l'ultime des sinistres. Bien souvent, les caractéristiques des sinistres sont restées identiques à celles observées à leur ouverture dans le système de gestion (évaluation forfaitaire, aucun règlement etc..).
La qualité prédictive de la méthode de projection individuelle s'améliore considérablement dès lors que les sinistres sont observés sur au moins deux années de développement. Après 5 années de développement, la méthode procure une estimation relativement précise des sinistres, même pour les plus coûteux.

Le coût ultime de certains sinistres restent toutefois difficiles à prédire et ce, malgré les avoir observés sur plusieurs années de développement. Pour ces sinistres, nous ne disposons à priori pas d'informations dans notre base d'étude nous permettant d'estimer précisément leur coût ultime.

La confrontation entre les prédictions et les observations permet difficilement d'analyser précisément l'impact de l'algorithme utilisé au niveau individuel. La Figure 4.12 représentent la densité empirique des erreurs de prédictions commises selon l'algorithme utilisé pour chaque modèle.

A noter que l'erreur de prédiction relative sinistre i (notée E_i) est calculée de la façon suivante :

$$E_i = f(x_i) - y_i$$

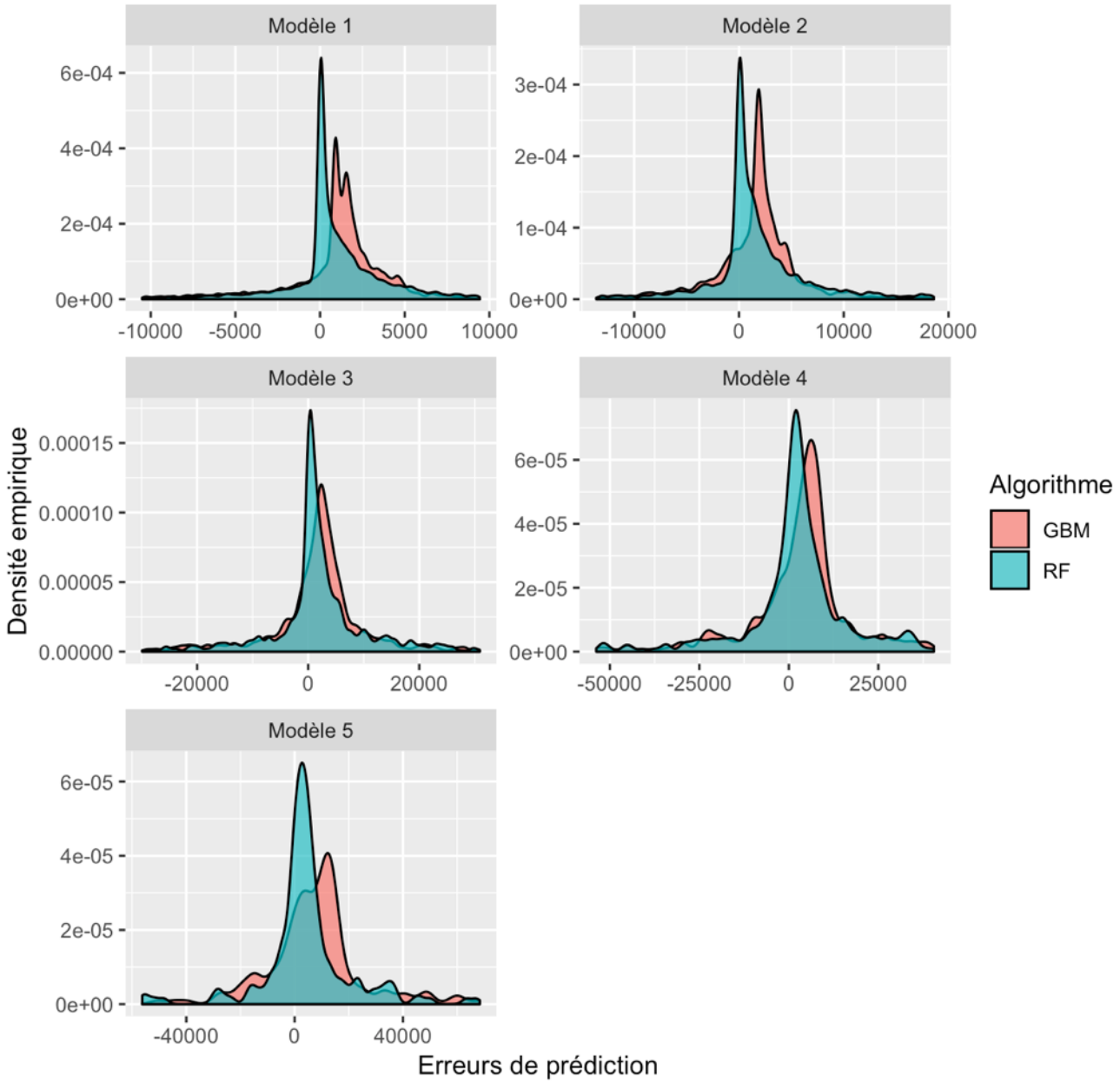


FIGURE 4.12 – Densité empirique des erreurs de prédiction pour chaque modèle

Comme énoncé précédemment, l'erreur d'estimation commise peut être relativement importante et ce, quel que soit l'algorithme utilisé. Par soucis de lisibilité, les erreurs de prédiction présentées en Figure 4.12 sont bornées entre le 5ème et 95ème centile.

Quel que ce soit le modèle de projection, les deux distributions des erreurs de prédictions présentent une asymétrie positive (i.e une queue de distribution plus étalée vers la droite du graphique) traduisant une tendance à surestimer le coût ultime des sinistres. Cette asymétrie semble toutefois plus marquée pour l'algorithme des forêts aléatoires. De plus, le *Gradient Boosting* a une légère tendance à sous-estimer plus fréquemment le coût ultime des sinistres que les forêts aléatoires. En effet, la densité des erreurs de prédictions est légèrement plus élevée sur la partie gauche du graphique.

D'autre part, les erreurs de prédiction commises par le *Gradient Boosting* ne sont pas centrées en 0. Le pic de la distribution se situe pour des valeurs strictement positives. Ce phénomène traduit vraisemblablement la présence d'un biais d'estimation en utilisant cet algorithme.

En termes de gestion des risques, l'algorithme utilisé pour la projection des sinistres à l'ultime a un impact significatif. En effet, les sous-estimations plus fréquentes du *Gradient Boosting* (partie gauche des graphiques en Figure 4.12) par rapport aux forêts aléatoires n'est pas souhaitable pour une compagnie d'assurance, que ce soit dans le cadre d'une problématique de provisionnement ou bien du pilotage de la rentabilité d'une branche d'activité.

D'autre part, la présence d'un biais d'estimation positif conduit l'assureur à fréquemment surestimer le coût de ses sinistres, et donc à sur-provisionner. Là aussi, ce n'est pas une situation souhaitable pour un assureur, il devra immobiliser davantage de capital pour faire face à ses engagements.

Dans le cas présent, les forêts aléatoires présentent une meilleure qualité d'estimation que le *Gradient Boosting*. Toutefois, ce constat est uniquement valable en analysant les prédictions individuelles. Selon la granularité retenue, il est tout à fait possible que le *Gradient Boosting* apporte une meilleure prédiction du montant de sinistres à un niveau agrégé. En effet, par agrégation du coût des sinistres, les surestimations et sous-estimations peuvent tout à fait se compenser.

4.4.3 Importance des variables dans la construction des modèles

Bien qu'efficaces, les forêts aléatoires et le *Gradient Boosting* sont bien souvent définis comme des « boîtes noires » en raison de la difficulté à expliquer les résultats obtenus. De manière générale, ce point figure parmi les inconvénients majeurs des méthodes de *Machine Learning*.

Néanmoins, les packages implémentant les forêts aléatoires et le *Gradient Boosting* proposent différentes mesures pour analyser l'importance des variables explicatives dans la construction des modèles prédictifs. Ces mesures peuvent également être utilisées à des fins de sélection de variables explicatives lors de l'étape de construction des modèles en supprimant celles qui apportent peu d'informations. Une application est réalisée dans le mémoire d'actuariat de D.Fabre Rudelle [11].

Afin de mesurer l'importance d'une variable explicative, Breiman [7] propose pour les forêts aléatoires de permuter aléatoirement les valeurs prises par cette variable et de quantifier l'augmentation des erreurs de prédiction sur l'échantillon de validation suite à cette permutation.
Cette mesure d'importance est également proposée pour le *Gradient Boosting* dans le package « *caret* » (utilisant le package « *gbm* ») à la différence près que l'augmentation de l'erreur de prédiction suite à une permutation aléatoire de la variable explicative en question est mesurée sur l'ensemble de la base d'apprentissage [7].

L'importance des variables explicatives pour les forêts aléatoires (resp. le *Gradient Boosting*) pour le premier modèle est présentée Figure 4.13 (resp. Figure 4.14).

7. https ://cran.r-project.org/web/packages/caret/caret.pdf

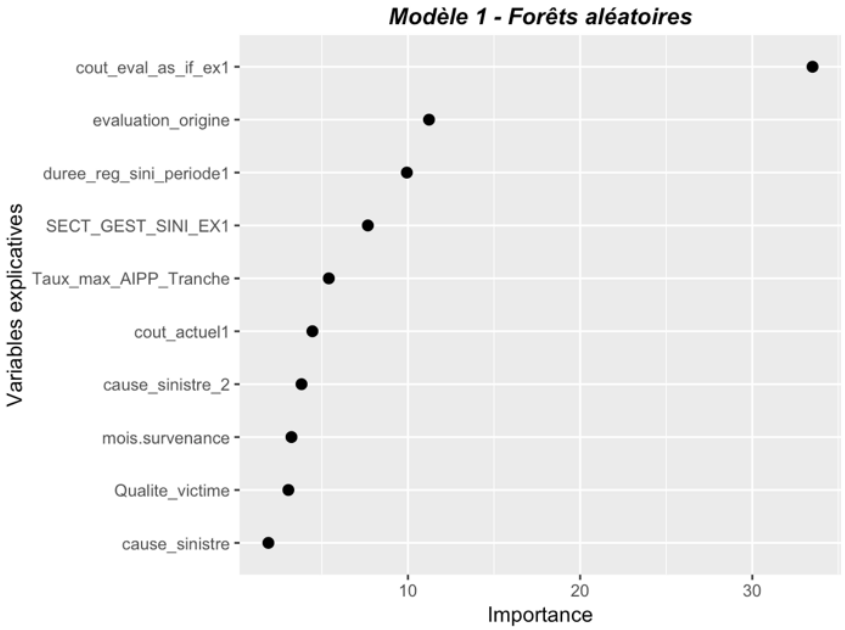


FIGURE 4.13 – Importance des variables pour le 1er modèle implémenté à l'aide des forêts aléatoires

Sans surprise, la charge dossier/dossier connue à la fin de la première année de développement (c'est-à-dire la dernière évaluation connue du sinistre pour ce modèle) est la variable la plus importante (cout_eval_as_if_ex1). C'est également le cas pour le *Gradient Boosting* (Figure 4.14).

Viennent ensuite des variables relatives au développement du sinistre (coût et durée de règlement des sinistres à la fin de la première année), la section de gestion du sinistre (SECT_GEST_SINI_EX1) puis des informations victimes (Taux d'AIPP, qualité de la victime). Les causes du sinistre impactent également les prédictions du coût ultime issues des forêts aléatoires.

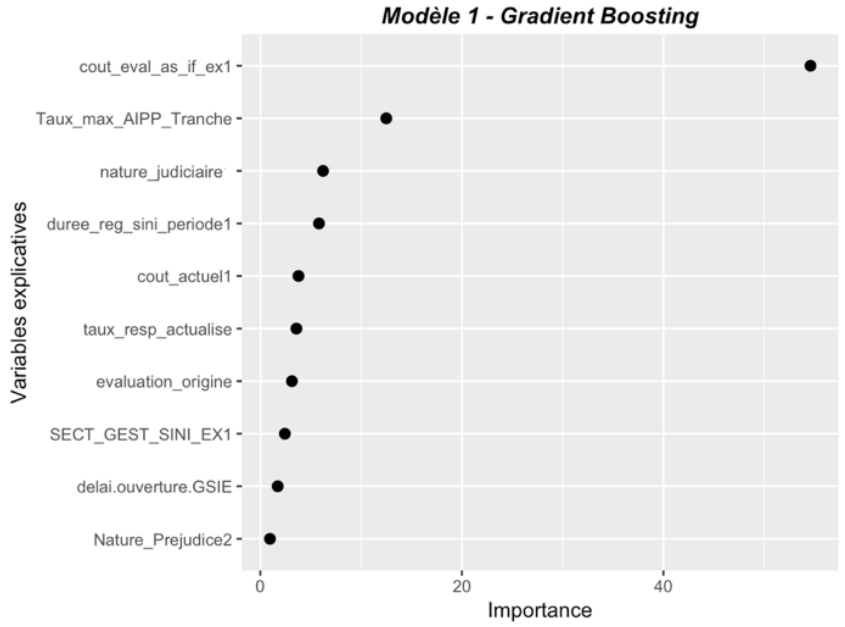


FIGURE 4.14 – Importance des variables pour le 1er modèle implémenté à l'aide du *Gradient Boosting*

En comparant avec les forêts aléatoires, l'importance de la dernière charge dossier/dossier connue est relativement plus marquée pour le *Gradient Boosting*. Certaines variables sont en commun mais dans un ordre d'importance différent, notamment le taux d'AIPP de la victime qui figure comme la deuxième

variable la plus importante. A noter ensuite l'impact significatif de la nature judiciaire du sinistre dans la construction du *GBM*, variable qui n'est pas présente pour les forêts aléatoires (Figure 4.13).

Il paraît ensuite intéressant d'analyser l'évolution de l'importance des variables entre le premier et le deuxième modèle (Figure 4.15 et 4.16).

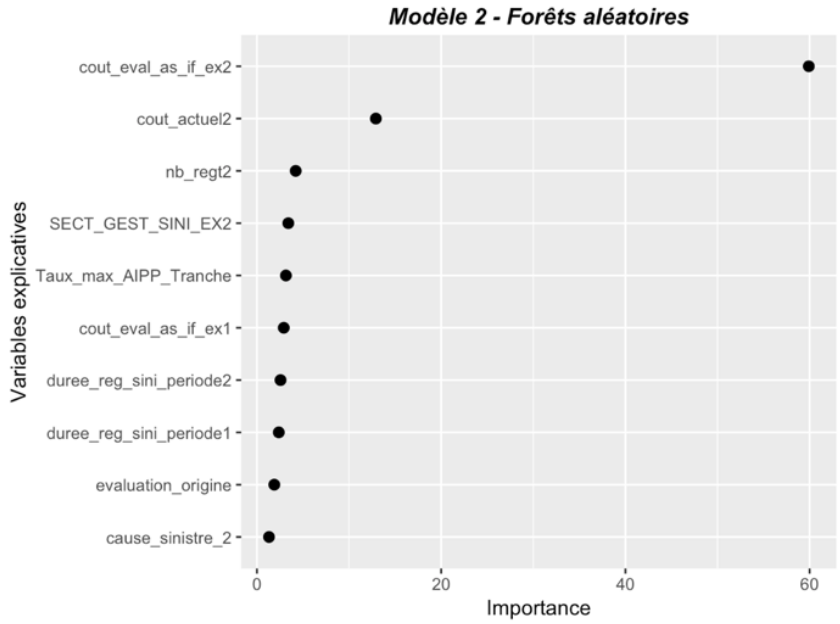


FIGURE 4.15 – Importance des variables pour le 2ème modèle implémenté à l'aide des forêts aléatoires

Outre le taux d'AIPP de la victime et la cause du sinistre, les variables ayant le plus d'impact sur les prédictions des forêts aléatoires concernent le développement du sinistre : les charges dossier/dossier et les coûts observés (cout_actuel) à la fin de chaque année de développement, le nombre de règlements (nb_regt2) ainsi que les durées de règlement (duree_reg_sini_periode).

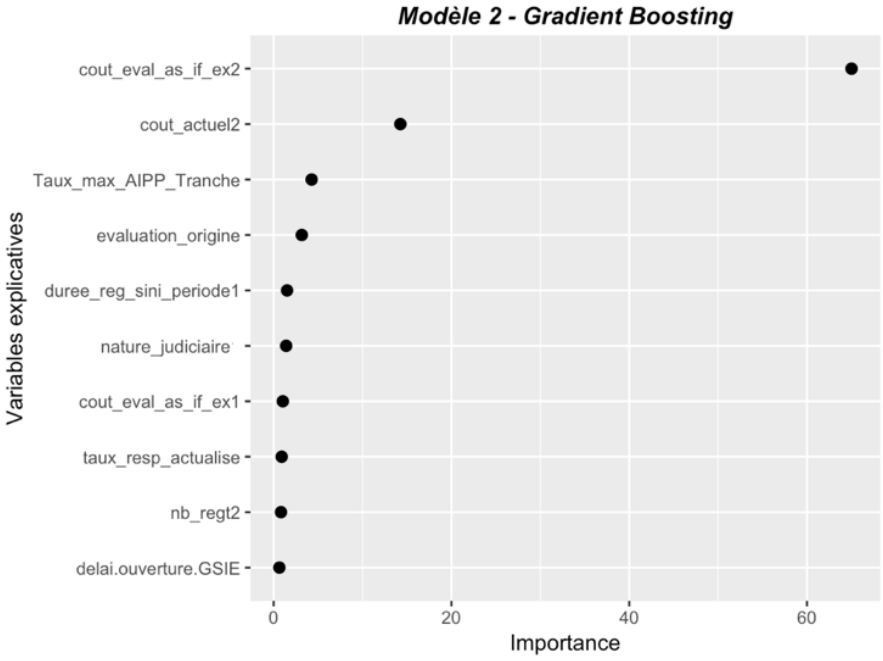


FIGURE 4.16 – Importance des variables pour le 2ème modèle implémenté à l'aide du *Gradient Boosting*

Comme pour les forêts aléatoires, les variables relatives au développement du sinistre figurent parmi les plus discriminantes. A mesure où la connaissance du sinistre s'améliore, l'importance de ces variables devient de plus en plus marquée. Cette tendance est vérifiée pour les modèles 3 à 5, c'est pourquoi la présentation de l'importance des variables dans la construction de ces derniers n'apporte que peu d'intérêt.

On remarquera par ailleurs l'absence des circonstances sinistres intégrés dans la base d'étude à l'aide du *Text Mining* dans le classement des variables les plus importantes. Ces informations ne se sont pas révélées pertinentes dans la construction des prédictions.

En résumé, la projection des sinistres observés à la fin de la première année de développement (modèle 1) s'avère relativement délicate, comme en témoigne la valeur des indicateurs de performance. Ce constat se justifie compte tenu du peu d'informations dont nous disposons sur ces sinistres à la date de projection.
La qualité prédictive de notre méthode individuelle s'améliore nettement lorsque les sinistres sont observés sur au moins deux périodes, bien que certains d'entre eux restent mal estimés même après plusieurs années de développement.

Les deux algorithmes affichent des résultats du même ordre de grandeur avec toutefois des comportements différents lorsque l'on s'intéresse aux prédictions individuelles (biais d'estimation pour le *Gradient Boosting*).

En toute logique, les prédictions issues des deux algorithmes s'appuient principalement sur des informations relatives au développement des sinistres (charge dossier/dossier, règlements) mais aussi sur le taux d'AIPP des victimes. A mesure où la connaissance du sinistre s'améliore, les charges dossier/dossier se rapprochent des coûts ultimes, justifiant leur importance prépondérante dans la construction des algorithmes.

Chapitre 5

Backtesting et comparaison avec Chain Ladder

La partie précédente nous a permis de juger de la qualité des prédictions fournies par la méthode de projection individuelle.

Nous proposons ici de réaliser dans un premier temps un benchmark en comparant la méthode de projection individuelle avec Chain Ladder, la méthode agrégée la plus utilisée à ce jour pour projeter les sinistres à l'ultime. Les prédictions issues de ces deux méthodes seront ensuite comparées aux coûts des sinistres réellement observés.

Pour ce faire, nous nous plaçons dans le cadre de la réalisation d'un reporting sur la liquidation de la sinistralité sur la branche RC corporelle automobile en utilisant notre base d'étude. D'un point de vue opérationnel, ce reporting est réalisé par les équipes actuariat de MMA en utilisant une méthode dérivée du Chain Ladder classique.

A noter que la méthode Chain Ladder ne sera pas présentée dans ce mémoire, le lecteur intéressé peut se référer à l'article de Thomas Mack [16].

5.1 Présentation du backtesting

Pour réaliser une comparaison aux coûts réellement observés, la démarche consiste ici à appliquer la méthode de projection individuelle ainsi que Chain Ladder à une date antérieure au 31/12/2017, fixée au 31/12/2015. Les sinistres déclarés après cette date de vision de données fictive sont par conséquent retirés de la base d'étude.

Une date de vision des données fixée au 31/12/2015 implique que les sinistres sont observés sur au maximum 7 années de développement, une période relativement courte pour une branche comme la RC corporelle. En effet, nous avons observé précédemment que certains sinistres survenus en 2009 ne sont, en principe, toujours pas clos au 31/12/2017 [1] (et donc au 31/12/2015).

1. Les sinistres déclarés en 2009 ont été clôturés pour appliquer la méthode de projection individuelle (partie 3.3.1).

Pour comparer la qualité prédictive des deux méthodes, les prédictions doivent être, en théorie, comparées aux coûts ultimes des sinistres.

Une première solution serait donc de se restreindre uniquement aux sinistres clos au 31/12/2017. Compte tenu de la relation croissante entre le coût des sinistres et le délai de règlement [2], ce choix conduirait à ne pas retenir les sinistres en-cours au 31/12/2017 dont le coût moyen est relativement plus élevée que celui des sinistres clos à cette même date.

Le fait d'écarter les sinistres en-cours au 31/12/2017 modifierait sensiblement les cadences de développement pour les années de survenance ou de déclaration récentes. Pour ces raisons, cette hypothèse est écartée.

L'alternative retenue ici consiste à utiliser l'ensemble des sinistres présents dans la base d'étude en retenant le coût observé au 31/12/2017 comme variable à prédire, lorsque l'ultime n'est pas connu.

Cependant, cette variable cible correspond respectivement pour les exercices 2009 et 2010 aux montants observés à la fin de la 9ème et 8ème année de développement alors qu'à fin 2015, les sinistres sont observés sur au maximum 7 années de développement. L'intégration d'un *tail factor* à la méthode Chain Ladder permettrait d'obtenir une estimation des coûts observés à fin 2017 pour ces deux années.

Cependant, la méthode de projection individuelle doit être construite de façon à analyser la liquidation des sinistres en-cours à la fin de la 7ème année de développement jusqu'à la 9ème année de développement pour projeter à l'ultime les sinistres de l'exercice 2009 (resp. 6ème et 8ème année de développement pour les sinistres de 2010). Au 31/12/2015, le développement de tels sinistres n'est pas observable, il n'est donc pas possible de projeter au 31/12/2017 par cette méthode les montants observés à fin 2015 pour ces deux exercices.

En conséquence, le coût des sinistres de 2009 ne sont pas projetés à fin 2017, aucune comparaison n'est réalisée pour cet exercice. Pour les sinistres de 2010, nous retenons le coût observé à fin 2016 comme variable cible (correspondant au montant observé à la fin de 7ème année de développement). Cet exercice de backtesting porte donc sur la dernière diagonale connue du triangle de liquidation construit à fin 2017.

Un autre point important concerne la prise en compte des sinistres IBNR dans la projection à l'ultime. La méthode individuelle développée dans ce mémoire n'intègre pas les IBNR, elle permet uniquement de projeter les sinistres RBNS. Par construction, les triangles de liquidation par année de survenance intègre implicitement les sinistres RBNS et IBNR. En choisissant ce type de triangle, les estimations ne seraient par conséquent pas comparables avec les prédictions de la méthode de projection individuelle.

Pour pallier à ce problème, nous construisons des triangles de liquidation par année de déclaration afin d'obtenir une cadence de développement de sinistres RBNS.

2. Relation observée lors de la présentation des statistiques descriptives de la base d'étude (partie 2.8.1)

Une alternative présentée dans le mémoire de M.Barbaste [2] serait de construire un triangle de liquidation par exercice de survenance en venant corriger les montants des sinistres IBNR afin de ne pas les projeter. Cette méthode implique de construire un triangle différent pour chaque exercice de survenance, soit 6 triangles dans notre cas.

Bien qu'intéressante, cette approche est plus lourde à implémenter, elle n'a donc pas été retenue.

La qualité prédictive des deux méthodes est analysée par année de déclaration et au global à l'aide de l'erreur relative définie comme suit :

$$Erreur\ relative = \frac{Coût\ agrégé\ prédit - Coût\ agrégé\ observé}{Coût\ agrégé\ observé}$$

La Figure 5.1 synthétise la démarche adoptée pour réaliser cette comparaison.

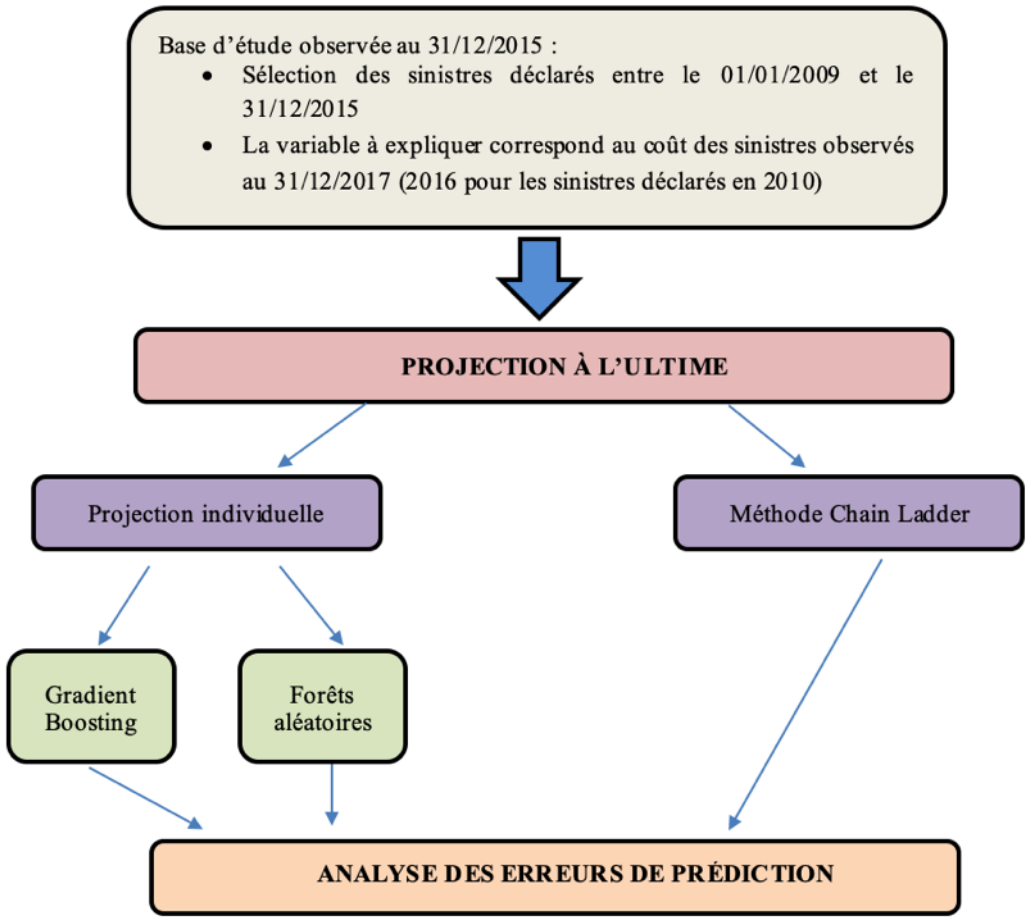


FIGURE 5.1 – Comparaison de la méthode de projection individuelle avec Chain Ladder

L'application de Chain Ladder est réalisée sur la base d'étude globale (hors millionnaires), sans distinction entre sinistres attritionnels et graves. Notons qu'en pratique, cette distinction est souvent réalisée en appliquant Chain Ladder sur les sinistres attritionnels, les sinistres graves sont traités séparément.

Néanmoins, la volumétrie de sinistres ne permet pas toujours d'effectuer cette séparation selon le périmètre considéré. De plus, l'intégration des sinistres majeurs (hors millionnaires) dans le processus de projection à l'ultime figure parmi les motivations du développement d'une méthode individuelle. La distinction entre sinistres attritionnels et graves n'est par conséquent pas réalisée ici.

5.2 Application de la méthode Chain Ladder

Nous commençons par construire notre triangle de liquidation par année de déclaration au 31/12/2015, la date de vision des données fictive. Après analyse des charges dossier/dossier et des règlements par année de déclaration, nous optons ici pour un triangle de paiements. Ce dernier est présenté en Figure 5.2.

Année de déclaration	Année de développement						
	1	2	3	4	5	6	7
2009	13 626 396	36 538 508	49 484 079	56 381 165	60 167 127	63 972 017	66 662 768
2010	14 581 451	38 754 146	53 391 624	61 284 818	66 620 492	70 028 847	
2011	15 626 631	40 314 619	55 309 386	64 805 725	71 438 477		
2012	16 531 401	39 937 704	53 093 961	60 410 893			
2013	15 400 632	40 866 814	58 403 634				
2014	15 562 927	41 249 515					
2015	14 369 077						

FIGURE 5.2 – Triangle de liquidation par année de déclaration au 31/12/2015

5.2.1 Validation des hypothèses de Chain Ladder

Avant d'appliquer Chain Ladder, rappelons l'hypothèse fondamentale devant être vérifiée au préalable :

Pour une année de développement donnée ($\forall\ j \in \{1, ..., J\}$), les facteurs de développement sont indépendants de l'année de survenance i ($\forall\ i \in \{1, ..., I\}$). Les facteurs de développement sont obtenus de la façon suivante :

$$f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \qquad (*)$$

Dans notre cas, nous devons vérifier que les facteurs de développement sont indépendants de l'année de déclaration.

Avant cela, nous devons nous assurer de l'existence d'une telle relation entre les paiements cumulées de l'année j et $j+1$.

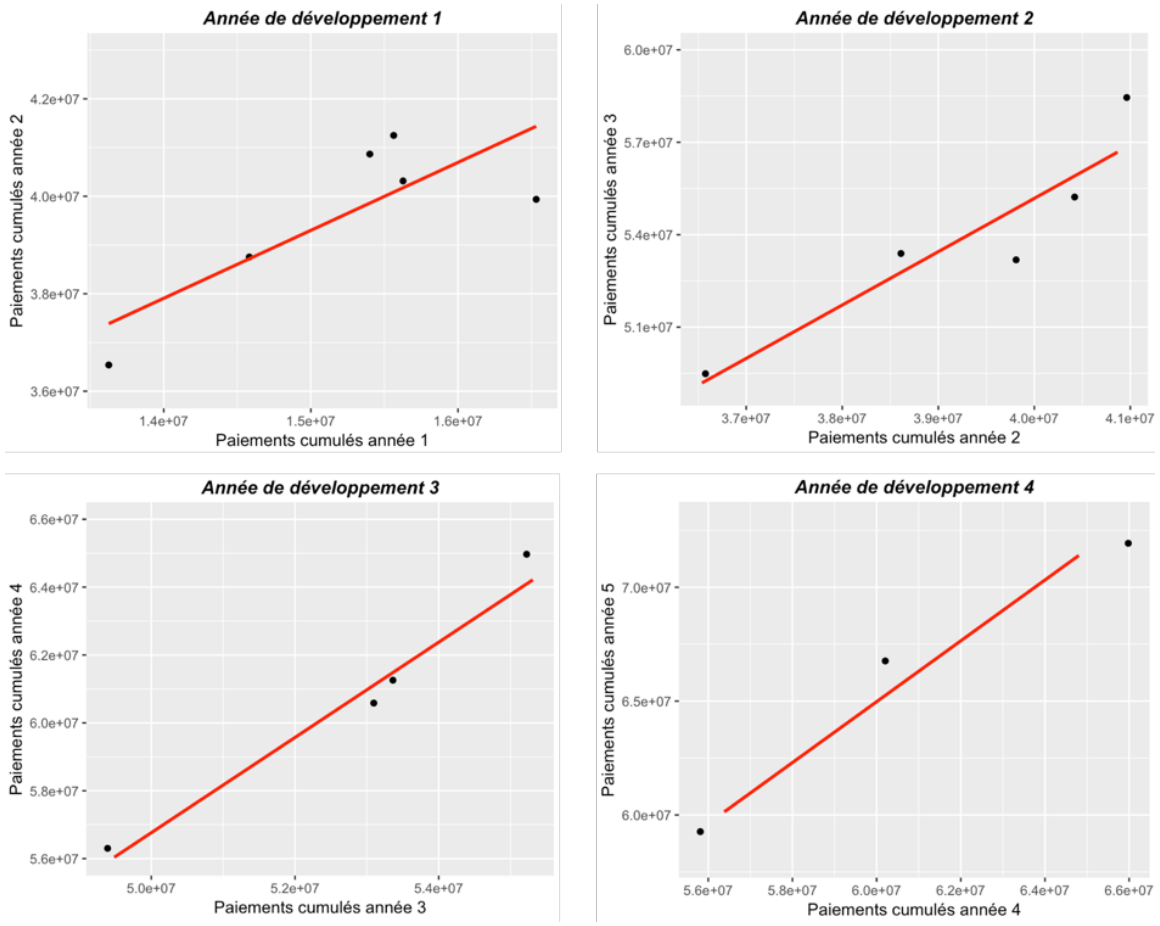


FIGURE 5.3 – Vérification de l'existence des facteurs de développement pour les années de développement 1 à 4

Au vu des quatre graphiques présentés sur la Figure 5.3, nous constatons une relation croissante entre les paiements cumulés de l'année j et $j+1$ ($j \in \{1, ..., 4\}$). Pour l'année de développement 5, une représentation graphique n'apporte que peu d'intérêt puisque seuls deux points seraient représentés. L'hypothèse concernant l'existence de la relation (*) est vérifiée empiriquement.

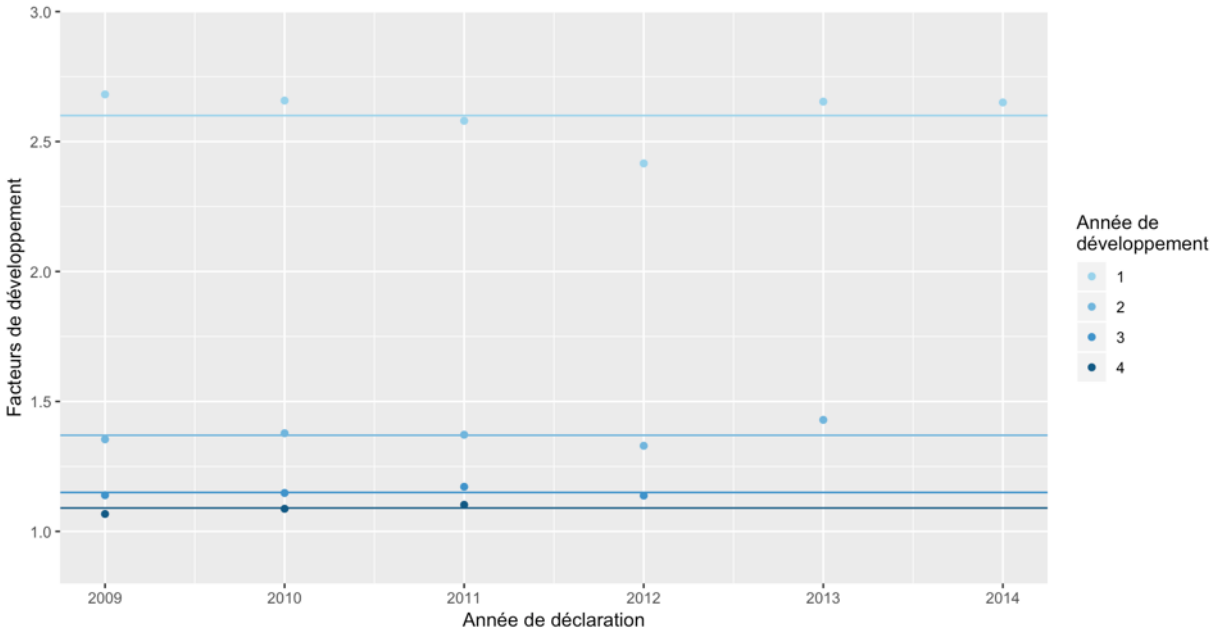


FIGURE 5.4 – Indépendance des facteurs de développement

La Figure 5.4 illustre les facteurs de développement empiriques en fonction de l'année de déclaration pour chaque année de développement. Quelle que soit l'année de déclaration, les facteurs de développement sont globalement constants pour une année de développement donnée.

Notons tout de même que de manière générale, les facteurs de développement de la première année de développement sont relativement plus volatiles que pour les autres années. En particulier, le facteur de développement de l'année 2012 se distingue légèrement des autres, nous choisissons donc de l'écarter pour l'estimation du facteur de développement par Chain Ladder.

Outre cette exception, l'hypothèse d'indépendance des facteurs de développement est vérifiée empiriquement.

5.2.2 Résultats

Après vérification des hypothèses, la méthode Chain Ladder est appliquée sur notre triangle de liquidation présentée en Figure 5.5 en écartant l'année de déclaration 2012 pour estimer le facteur de développement associé à la 1ère année de développement.

Année de déclaration	Année de développement						
	1	2	3	4	5	6	7
2009	13 626 396	36 538 508	49 484 079	56 381 165	60 167 127	63 972 017	66 662 768
2010	14 581 451	38 754 146	53 391 624	61 284 818	66 620 492	70 028 847	72 974 356
2011	15 626 631	40 314 619	55 309 386	64 805 725	71 438 477	75 502 780	78 678 530
2012	16 531 401	39 937 704	53 093 961	60 410 893	65 626 697	69 360 354	72 277 745
2013	15 400 632	40 866 814	58 403 634	67 139 769	72 936 535	77 086 066	80 328 412
2014	15 562 927	41 249 515	56 637 536	65 109 494	70 730 969	74 755 020	77 899 319
2015	14 369 077	37 983 694	52 153 409	59 020 388	64 116 136	67 763 854	70 614 095

FIGURE 5.5 – Application de la méthode Chain Ladder

Les facteurs de développement obtenus (avant et après correction) sont présentés dans le Tableau 5.1.

Année de développement	*1*	*2*	*3*	*4*	*5*	*6*
Facteurs de développement Chain Ladder	2,602	1,373	1,150	1,086	1,057	1,042
Facteurs de développement Chain Ladder après correction	2,643	1,373	1,150	1,086	1,057	1,042

TABLE 5.1 – Facteurs de développement obtenus par Chain Ladder par année déclaration

Rappelons que l'objectif ici est de comparer les estimations obtenues par les deux méthodes avec les coûts observés à fin 2017 (fin 2016 pour les sinistres déclarés en 2010). Chaque diagonale supplémentaire du triangle correspond aux estimations des montants de paiements cumulés observés à la fin de l'année suivante. Dans notre cas, l'exercice de comparaison est réalisé en projetant au 31/12/2017 les sinistres en-cours à fin 2015. Nous nous intéressons donc aux estimations présentées en rouge sur la deuxième

diagonale du triangle en Figure 5.5 (sauf pour l'année 2010). Ces montants correspondent pour les années de déclaration 2011 à 2015 à la dernière diagonale connue du triangle construit au 31/12/2017.

Les montants encadrés en gris correspondent à des prédictions postérieures au 31/12/2017, par conséquent, ils ne sont pas analysés dans le cadre de cette comparaison.

La Figure 5.6 et le Tableau 5.7 présentent les erreurs d'estimation obtenues en appliquant la méthode Chain Ladder.

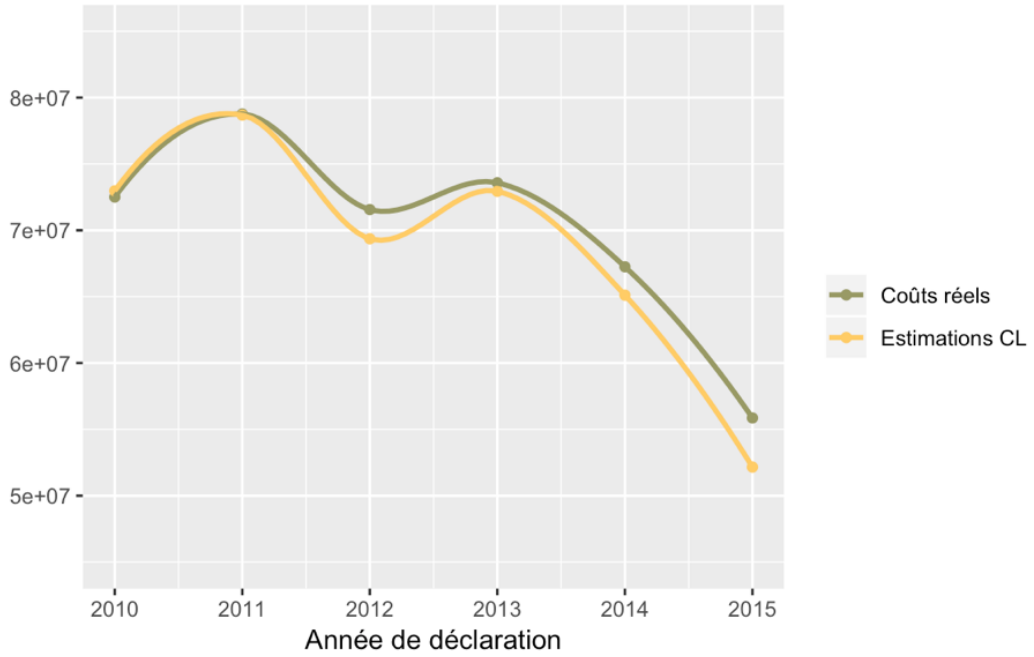


FIGURE 5.6 – Comparaison des estimations Chain Ladder avec les montants observés par année de déclaration

	2010	*2011*	*2012*	*2013*	*2014*	*2015*	*Global*
Coûts observés (en millions d'euros)	72,5	78,8	71,6	73,6	67,2	55,9	419,5
Estimations Chain Ladder (CL)	73	78,7	69,4	73	65,1	52,2	411,2
Erreurs relatives CL (en %)	0,7 %	-0,1 %	-3,1 %	-0,9 %	-3,2 %	-6,6 %	-2 %

FIGURE 5.7 – Erreurs de prédiction de la méthode Chain Ladder

Globalement, la méthode Chain Ladder a tendance à sous-estimer le développement des sinistres pour chacune des années de développement. L'erreur d'estimation globale est d'approximativement 2% soit un peu plus de 8 millions d'euros. Pour les années de déclaration 2012 et 2014, la méthode Chain Ladder sous-estime d'un peu plus de 3% le coût des sinistres observés à fin 2017.

Pour 2015, l'erreur commise par Chain Ladder est conséquente puisqu'elle s'élève à 6,6%, une telle sous-estimation n'est pas souhaitable pour une compagnie d'assurance. Pour en comprendre les raisons, les facteurs de développement observés à fin 2017 sont représentés Figure 5.8.

Année de déclaration	Année de développement					
	1	2	3	4	5	6
2009	2,681	1,354	1,139	1,067	1,063	1,042
2010	2,658	1,378	1,148	1,087	1,051	1,035
2011	2,580	1,372	1,172	1,102	1,059	1,041
2012	2,416	1,329	1,138	1,135	1,043	
2013	2,654	1,429	1,178	1,070		
2014	2,650	1,393	1,171			
2015	2,790	1,393				

FIGURE 5.8 – Facteurs de développement observés au 31/12/2017 pour les années de déclaration 2009 à 2015

Nous avions observé sur la Figure 5.4 que les facteurs de développement de la 1ère année sont relativement plus volatiles que pour les autres années. Les facteurs représentés en rouge sont à comparer avec ceux estimés par Chain Ladder (Tableau 5.1) pour l'année 2015. Pour la 1ère année, nous observons que le facteur de développement pour l'année de déclaration 2015 est légèrement atypique puisqu'il est supérieur à tous les facteurs associés aux années de déclaration passées.

La méthode Chain Ladder sous-estime de 5,6% le facteur de développement observé la première année pour l'année 2015. A noter qu'en intégrant l'année 2012, le facteur de développement estimé s'élève à 2,602. L'erreur commise aurait été de 7,2% sur la première année entraînant une sous-estimation du coût à fin 2017 de 8,1%.

L'année de déclaration 2015 illustre une des limites de la méthode Chain Ladder : le développement des sinistres passés ne reflète pas toujours le développement futur des sinistres.

Reste à savoir si l'intégration de données individuelles dans la projection des sinistres à l'ultime permet d'améliorer la qualité des prédictions des coûts observés à fin 2017, notamment pour l'année de déclaration 2015.

5.3 Application de la méthode de projection individuelle

Pour réaliser ce backtesting, les coûts des sinistres observés à fin 2017 (fin 2016 pour l'année 2010) sont comparés avec les prédictions issues des deux algorithmes. Pour la méthode Chain Ladder, cet exercice de comparaison consistait à analyser les montants présentés en rouge sur la deuxième diagonale du triangle en Figure 5.5 (mis à part pour 2016). En toute cohérence, la méthode de projection individuelle doit être adaptée pour obtenir des prédictions comparables avec celles issues de Chain Ladder.

5.3.1 Adaptation de la méthode de projection individuelle

Jusqu'à présent, la méthode de projection était construite de façon à prédire directement le coût ultime des sinistres en fonction de leur année de déclaration (i.e la profondeur d'historique dont nous disposons). Cependant, les coûts observés à fin 2017 des sinistres en-cours au 31/12/2015 correspondent seulement pour une partie d'entre eux à leur coût ultime.

Dans le cas présent, la méthode de projection individuelle doit être adaptée afin de prédire l'évolution du coût des sinistres entre le 31/12/2015 et le 31/12/2017.

Le premier impact concerne la variable réponse des deux algorithmes de *Machine Learning*, elle est remplacée par le coût observé à fin 2017. Les variables explicatives utilisées dans chaque modèle restent inchangées.

D'autre part, la construction des bases d'apprentissage des différents modèles est modifiée. Tout d'abord, étant donné que dans le cas présent nous ne cherchons pas à prédire un coût ultime, les sinistres utilisés pour entraîner les différents modèles ne sont pas nécessairement clos.

Pour projeter les sinistres en-cours au 31/12/2015 observés sur p années de développement [3] (modèle p), nous retenons pour la phase d'apprentissage des modèles tous les sinistres déclarés avant l'année 2015 - $p+1$ qui étaient en-cours à la fin de l'année de développement p.

Pour bien comprendre ce changement, nous pouvons reprendre la Figure 3.3 (partie 3.3.1) illustrant la méthode de projection à l'ultime des sinistres observés après deux années de développement ($p = 2$) en l'adaptant dans le cadre de ce backtesting :

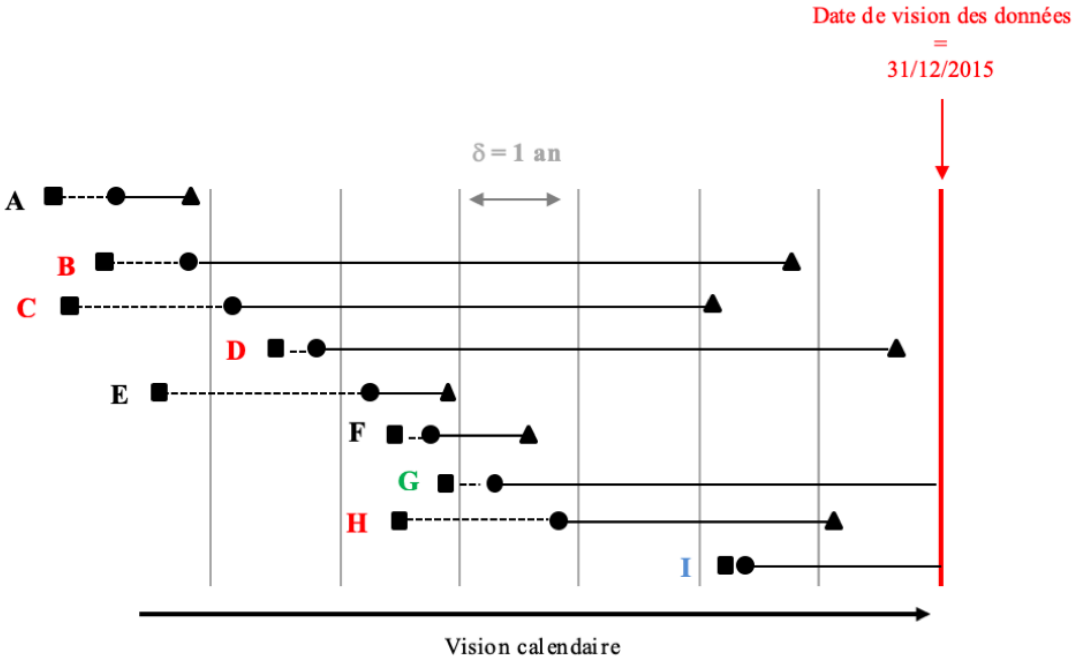


FIGURE 5.9 – Modification de la méthode de projection individuelle pour le backtesting : exemple avec les sinistres déclarés en 2014

3. C'est-à-dire dans le cas présent les sinistres déclarés au cours de l'année 2015 – $p+1$.

Pour projeter le sinistre I déclaré en 2014, le sinistre G n'avait pas servi à la construction du modèle étant donné que celui-ci est toujours en-cours à la date de vision des données (son coût ultime n'est pas connu).

Dans le cadre de ce backtesting, l'évolution du sinistre G entre la 2ème et 4ème année de développement peut être utilisée pour projeter à fin 2017 les sinistres déclarés en 2014.

D'autre part, le découpage de la base d'étude entre base d'apprentissage et base test n'est plus nécessaire ici. En effet, les sinistres en-cours au 31/12/2015 répartis par année de déclaration constituent naturellement les bases test des différents modèles.

Le Tableau 5.2 présente le nombre de sinistres utilisés pour l'apprentissage des modèles et le nombre de sinistres à projeter par modèle.

Modèle	*1*	*2*	*3*	*4*	*5*	*6*
Base en entrée du modèle p	39 305	12 570	3 933	1 649	678	204
Bases test	6 724	2 840	1 019	566	363	252

TABLE 5.2 – Volumes de sinistres pour chaque modèle (apprentissage et test)

Comme lors de la présentation des bases en entrée des modèles (partie 3.3.3), nous faisons face à un problème de volumétrie de sinistres pour les modèles 5 et 6 (c'est-à-dire pour la projection des sinistres déclarés en 2010 et 2011).

Pour réaliser ce backtesting, nous supposons que le volume de sinistres est suffisant pour projeter ces sinistres. Néanmoins, une attention toute particulière sera portée aux résultats issus de ces deux modèles. Même dans le cas où les prédictions seraient de bonne qualité, cette faible volumétrie de sinistre pour certains modèles de projection constitue une des limites de notre méthode individuelle. Ce point sera plus longuement discuté en conclusion de ce mémoire.

Outre ces différents changements, le coût des sinistres sont projetés à fin 2017 en suivant la démarche présentée dans la partie 3.3.3.

5.3.2 Résultats

Après avoir adapté la méthode de projection individuelle pour réaliser ce backtesting et comparer les prédictions avec la méthode Chain Ladder, le processus d'optimisation des hyperparamètres présenté dans la partie 4.3 est répété pour les forêts aléatoires et le *Gradient Boosting*.

La Figure 5.10 et le Tableau 5.3 présentent les prédictions obtenues par année de déclaration en utilisant la méthode de projection individuelle.

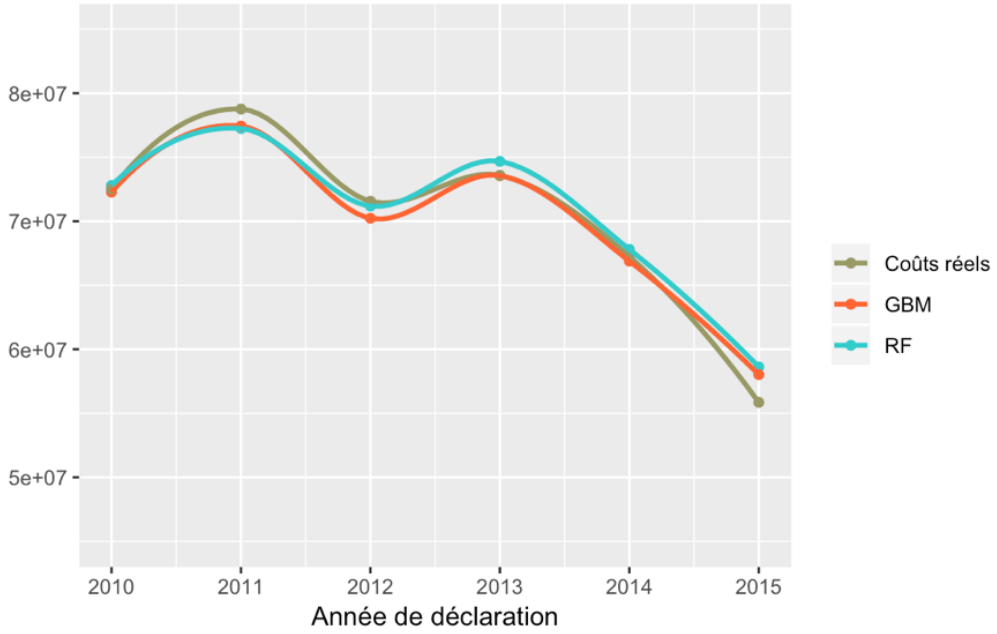


FIGURE 5.10 – Comparaison des estimations issues de la méthode de projection individuelle avec les coûts observés par année de déclaration

	2010	*2011*	*2012*	*2013*	*2014*	*2015*	*Global*
Coûts observés (en millions d'euros)	72,5	78,8	71,6	73,6	67,2	55,9	419,5
Prédictions RF	72,8	77,2	71,2	74,7	67,8	58,6	422,3
Prédictions GBM	72,3	77,4	70,2	73,6	66,9	58	418,4
Erreurs relatives RF (en %)	0,4 %	-1,9 %	-0,5 %	1,5 %	0,8 %	5 %	0,7 %
Erreurs relatives GBM (en %)	-0,3 %	-1,7 %	-1,8 %	0 %	-0,5 %	3,9 %	-0,3 %

TABLE 5.3 – Erreurs de prédiction de la méthode de projection individuelle par année de déclaration

Tout d'abord, l'impact du choix de l'algorithme est moindre à un niveau agrégé. En effet, les prédictions obtenues sont relativement proches, que ce soit en utilisant les forêts aléatoires ou le *Gradient Boosting*. Outre l'année 2015, l'erreur d'estimation en valeur absolue ne dépasse pas 2%, les prédictions par année de déclaration sont mêmes parfois très proches des coûts observés. Pour les années 2010 et 2012 (resp. 2010, 2013 et 2014) l'erreur commise en valeur absolue par les forêts aléatoires (resp. *Gradient Boosting*) est inférieure à 0,5%. Pour ces quatre années de déclaration (2010, 2012, 2013 et 2014), la méthode de projection individuelle fournie des prédictions de très bonne qualité.

Pour l'année de déclaration 2011, les deux algorithmes commettent une erreur proche de 2% représentant une sous-estimation du coût des sinistres de près de 1,5 millions d'euros. Celle-ci peut être due à la faible volumétrie de sinistres utilisée pour entraîner le modèle de projection (cf Tableau 5.2). Nous pouvons donc supposer qu'une base d'apprentissage plus volumineuse aurait permis d'améliorer la qualité prédictive du modèle. Néanmoins, l'année de déclaration 2010 est quant à elle très bien estimée et ce malgré une base d'apprentissage constituée de très peu de sinistres.

La méthode de projection individuelle surestime de 5% (resp. 3,9%) le coût au 31/12/2017 des sinistres déclarés au cours de l'année 2015 lorsque celle-ci est implémentée à l'aide des forêts aléatoires (resp. du *Gradient Boosting*). Le constat est conforme à celui observé lors de la prédiction du coût ultime des sinistres à la fin de la 1ère année de développement (partie 4.4.2) : Les sinistres déclarés récemment restent délicats à estimer compte tenu du peu d'informations dont nous disposons sur ces derniers.

Même si les erreurs de prédictions en valeur absolue sont globalement plus faibles en utilisant le *Gradient Boosting*, en termes de gestion des risques l'algorithme des forêts aléatoires pourrait être privilégié. En effet, les prédictions sont globalement prudentes puisqu'elles sont supérieures aux coûts observés (de 0,7%).

5.4 Comparaison avec Chain Ladder et discussion sur l'apport des méthodes individuelles

Cette dernière partie est consacrée à la comparaison de la méthode de projection individuelle avec Chain Ladder. La Figure 5.11 présente les estimations obtenues à partir des deux méthodes ainsi que les coûts observés à fin 2017 [4] par année de déclaration. Les erreurs relatives de prédiction sont observées sur le Tableau 5.4.

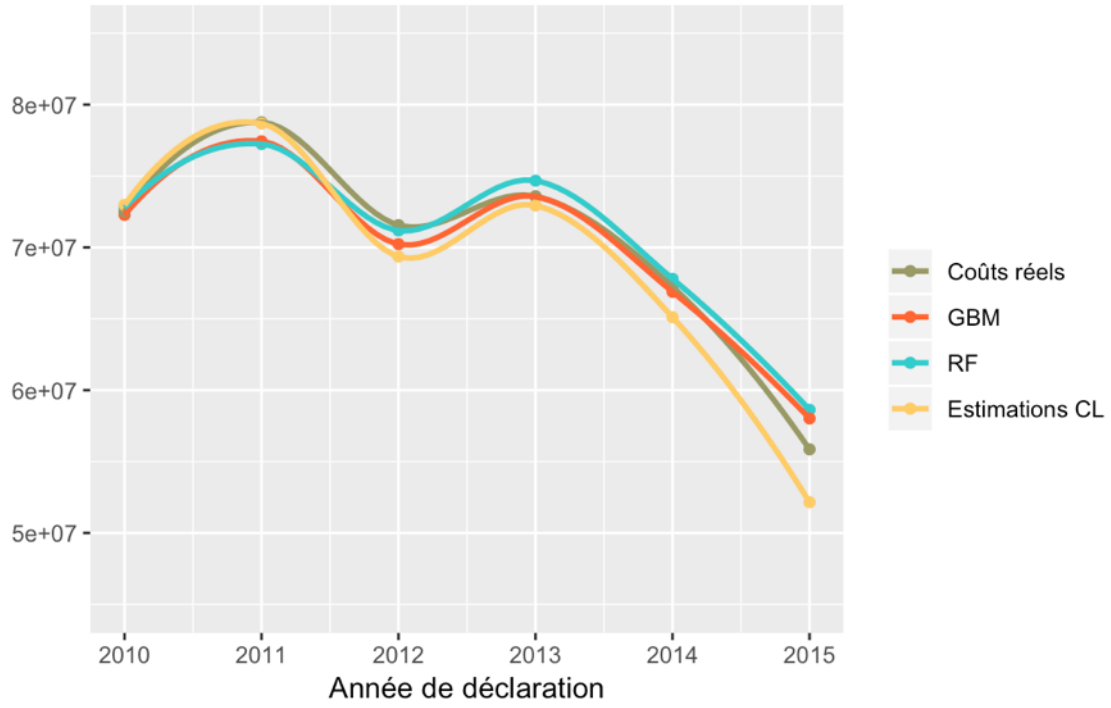


FIGURE 5.11 – Comparaison des estimations issues de la méthode de projection individuelle et Chain Ladder avec les coûts observés par année de déclaration

4. A fin 2016 pour l'année de déclaration 2010

	2010	*2011*	*2012*	*2013*	*2014*	*2015*	*Global*
Chain-Ladder (CL)	0,7 %	-0,1 %	-3,1 %	-0,9 %	-3,2 %	-6,6 %	-2 %
Forêts aléatoires (RF)	0,4 %	-1,9 %	-0,5 %	1,5 %	0,8 %	5 %	0,7 %
Gradient Boosting (GBM)	-0,3 %	-1,7 %	-1,8 %	0 %	-0,5 %	3,9 %	-0,3 %

TABLE 5.4 – Erreurs d'estimation relatives par année de déclaration obtenues par Chain Ladder et la méthode de projection individuelle

La méthode de projection individuelle se révèle plus précise que Chain Ladder sur toutes les années de déclaration, exceptée 2011. L'intégration de données individuelles a permis d'améliorer la prédiction des coûts observés. Toutefois, les erreurs d'estimation de la méthode Chain Ladder pour les années 2010 et 2013 restent relativement faibles, même inférieures à celles commises par la méthode de projection individuelle implémentée à l'aide des forêts aléatoires pour l'année 2013.

Ce constat permet de s'interroger sur la pertinence de développer une méthode de projection individuelle. Bien qu'elles puissent être plus précises que les méthodes agrégées, elles sont généralement beaucoup plus lourdes à implémenter. En prenant le cas de la méthode développée dans ce mémoire, le processus d'optimisation des hyperparamètres nécessitent plusieurs heures de traitement. Outre la problématique de temps de calculs, les modèles restent sensibles au risque de sur-apprentissage et au mauvais calibrage des hyperparamètres.

De plus, malgré les outils mis à disposition pour comprendre la construction des modèles de *Machine Learning*, les résultats restent tout de même difficiles à expliquer. D'un point de vue opérationnel, il est essentiel de pouvoir comprendre et présenter clairement les prédictions obtenues.

Néanmoins, les méthodes de projection individuelle offrent la possibilité d'utiliser une grande quantité d'informations sur les sinistres. En conséquence, elles permettent de pallier à certaines lacunes des méthodes agrégées, notamment lors de l'estimation du coût des sinistres déclarés récemment ou bien lorsqu'une année de déclaration se révèle légèrement atypique.

En reprenant le Tableau 5.4, le *Gradient Boosting* surestime les coûts observés à fin 2017 des sinistres déclarés au cours de l'année 2015 de près de 4%, contre une sous-estimation de 6,6% de Chain Ladder. De plus, cet exercice de backtesting consistait à projeter le coût des sinistres à deux ans (de fin 2015 à fin 2017). Les effets multiplicatifs de la méthode Chain Ladder peuvent potentiellement entraîner d'importantes erreurs d'estimation du coût ultime des sinistres déclarés au cours de l'année 2015.

Enfin, les méthodes individuelles fournissent des estimations du coût ultime pour chaque sinistre ce qui peut se révéler particulièrement utiles pour appliquer des traités de réassurance en excédent de sinistres ou bien réaliser des reportings sur des périmètres plus réduits.

Cette discussion au sujet des avantages et inconvénients des méthodes vient clôturer la dernière partie de ce mémoire.

Conclusion

Dans ce mémoire, nous avons développé puis implémenté une méthode de projection individuelle en ayant pour objectif d'améliorer la prédiction du coût des sinistres en intégrant des données individuelles. Cette méthode a été appliquée sur la sinistralité en Responsabilité Civile Corporelle (RCC) issue du marché des professionnels de l'automobile et des particuliers sur la période 2009-2017.

Cette méthode se révèle innovante dans le sens où elle permet de projeter les sinistres individuellement en fonction de la quantité d'informations disponibles sur ces derniers. D'autre part, la réalisation d'une étude de *Text Mining* sur les circonstances des sinistres a permis d'ajouter de l'information dans le modèle de projection, même si au final, elles ne se sont pas révélées pertinentes dans la construction des prédictions. La démarche de récupération de données non structurées entreprise dans ce mémoire illustre (partiellement) la capacité des outils actuels à intégrer une multitude de sources de données présentées sous différentes formes. C'est aussi en cela que l'approche adoptée dans ce mémoire est originale.

Nous avons dans un premier temps comparé le pouvoir prédictif de notre méthode en l'implémentant à l'aide de deux algorithmes de *Machine Learning* : les forêts aléatoires et le *Gradient Boosting*. Une attention toute particulière a été portée sur l'optimisation des hyperparamètres dans le but d'améliorer leur pouvoir prédictif sur de nouvelles données. Même si les indicateurs de performance ont témoigné d'une bonne qualité d'ajustement des différents modèles, certains sinistres coûteux restent difficiles à prédire même après plusieurs années de développement.

Ensuite, les résultats ont été comparés à ceux obtenus avec une méthode de place pour l'estimation des coûts ultimes, la méthode Chain-Ladder.
Globalement, la méthode de projection individuelle s'est avérée plus performante, en particulier pour les exercices légèrement atypiques (cf année de déclaration 2015, partie 5.4) et pour les sinistres déclarés récemment dont les coûts se révèlent être plus délicats à estimer. En ce sens, l'implémentation d'une méthode individuelle apporte une réelle plus-value dans l'estimation des coûts observés par rapport aux méthodes agrégées.

Néanmoins, ce gain en précision est à mettre au regard de la difficulté à implémenter ce type de méthode. En effet, les étapes de construction des différents modèles en passant par le paramétrage des algorithmes nécessitent beaucoup de ressources informatiques et plusieurs heures de calculs. Dans le cas où seules les prédictions à un niveau agrégé nous intéressent, la construction d'une méthode individuelle semble peu justifiée lorsque les estimations coïncident avec la méthode Chain-Ladder (cf années de déclaration 2010, 2011 et 2014, partie 5.4).

La démarche entreprise dans ce mémoire peut être améliorée sur plusieurs aspects. Tout d'abord, certaines données en entrée des modèles de projection sont restées manquantes (notamment les données victimes) et ce, malgré les différents retraitements réalisés. De plus, certaines informations n'ont pu être observées qu'à une date fixe (lors de la survenance des sinistres ou bien lors de l'extraction des données) alors qu'elles ont pu évoluer au cours de la liquidation des sinistres (par exemple le taux d'AIPP de la victime ou bien la nature judiciaire du sinistre).

Ajouté à cela, plusieurs modèles de projection ont été entraînés à partir d'une faible volumétrie de sinistres. Tous ces éléments ont pu potentiellement diminuer la qualité des résultats obtenus et constituent un axe d'amélioration.

D'autre part, les sinistres tardifs (IBNR) ne sont pas intégrés dans la méthode de projection individuelle développée dans ce mémoire. Par définition, nous ne disposons pas d'informations sur ces sinistres à la date de vision des données. M.Baudry et C.Y.Robert [3] proposent d'estimer le coût de ces sinistres à l'aide d'une decomposition fréquence-coût moyen en utilisant des informations propres aux contrats (exposition, la date de souscription etc..). Ne disposant pas de telles informations dans la base d'étude, cette méthode a été écartée.

D'autres algorithmes de *Machine Learning* auraient également pu être considérés pour améliorer la prédiction du coût ultime des sinistres, en citant par exemple les réseaux de neurones (cf M.V.Wüthrich [24], D.Fabre Rudelle [11], P.Ottou [18]) et les *Random Extra Trees* (cf M.Baudry et C.Y.Robert [3]). L'introduction d'un "meta-regresseur" combinant les predictions issues des forêts aléatoires et du *Gradient Boosting* permettrait également d'améliorer le pouvoir prédictif de chaque modèle de projection (*Stacking*).

Bibliographie

[1] Antonio, K. and Plat, R. [2014], *Micro-level stochastic loss reserving for general insurance*, Scandinavian Actuarial Journal, 7, 649-669.

[2] Barbaste, M. [2017], *Une méthode de provisionnement individuel par apprentissage automatique*, Mémoire d'actuariat.

[3] Baudry, M. and Robert, C. [2017], *Non parametric individual claim reserving in insurance*, Université Lyon 1 ISFA – Chaire DAMI.

[4] Billore, C. [2016], *Application de l'apprentissage automatique au provisionnement ligne à ligne en assurance non-vie*, Mémoire d'actuariat.

[5] Bishop, C. [2006], *Pattern Recognition and Machine Learning*, Springer.

[6] Breiman, L. [1996], *Bagging Predictors*, Machine Learning, 24, 123–140.

[7] Breiman, L. [2001], *Random Forests*, Machine Learning, 45, 5–32.

[8] Breiman, L., Friedman, J., Stone, C. and Olshen, R. [1984], *Une méthode de provisionnement individuel par apprentissage automatique*, The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

[9] Chalnot, J., Gremillet, M., Miehe, P. and Trufin, J. [2015], *Implementing the Individual Claims Reserving method, a new approach in non-life reserving*, ASTIN Bulletin.

[10] Dinh, N. and Chau, G. [2012], *Mesures de provisions cohérentes et méthode ligne à ligne pour des risques non-vie*, Mémoire d'actuariat.

[11] Fabre Rudelle, J. [2018], *Apport des methodes d'apprentissage statistique pour le provisionnement individuel en assurance non-vie*, Mémoire d'actuariat.

[12] Freund, Y. [1996], *Experiments with a new boosting algorithm*, Proceedings of the Thirteenth International Conference on Machine Learning.

[13] Friedman, J. [2001], *Greedy Function Approximation : A Gradient Boosting Machine*, The Annals of Statistics, 29, No. 5, 1189-1232, Institute of Mathematical Statistics.

[14] Friedman, J. [2002], *Stochastic Gradient Boosting*, Computational Statistics and Data Analysis, 38, 367-378.

[15] Larsen, C. [2007], *An individual claims reserving models*, ASTIN Bulletin, 37, 113-132.

[16] Mack, T. [1993], *Distribution-free calculation of the standard error of chain ladder reserve estimates*, ASTIN Bulletin, 23, No 2., 213 :225.

[17] Norberg, R. [1993], *Prediction of outstanding liabilities in non-life insurance*, ASTIN Bulletin.

[18] Ottou, P. [2017], *Méthode d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie*, Mémoire d'actuariat.

[19] Pigeon, K., Antonio, K. and Denuit, M. [2013], *Individual loss reserving with the multivariate skew normal framework*, ASTIN Bulletin, 43, 399-428.

[20] Rakotomalala, R. [2008], *Comparaison de population. Tests non paramétriques*, Université Lyon 2.

[21] Rakotomalala, R. [2016], *Gradient Boosting*, Université Lyon 2.

[22] Rouvière, L. [n.d.], *Introduction aux methodes d'agrégation : boosting, bagging et forêts aléatoires. Illustrations avec R*, Université de Rennes.

[23] Wüthrich, M. [2016], *Machine Learning in Individual Claims Reserving*, Research Paper Series n°16-67, ETH Zurich - Swiss Finance Institute.

[24] Wüthrich, M. [2018], *Neural networks applied to chain-ladder reserving*, European Actuarial Journal (2018) 8 :407-436.

Table des figures

1.1 Composition du marché des professionnels de l'automobile à fin 2015 9
1.2 Structure du produit MMA Pros de l'Auto 10

2.1 Densité du coût des sinistres bornés entre 0 et le 95ème centile en fonction du périmètre 17
2.2 Résultats du test de Mann - Withney 18
2.3 Représentation du phénomène de long tail 23
2.4 Nuage de mots 25
2.5 Nuage de mots comparatif 26
2.6 Synthèse des différents retraitements réalisés sur la nature du préjudice 31
2.7 Evolution du coût moyen d'indemnisation des sinistres corporels à gravité constante sur la période 2002-2011 34
2.8 Répartition des sinistres par exercice de survenance selon leur état 38
2.9 Répartition des sinistres par exercice de survenance selon leur état 39
2.10 Distribution de la durée de règlement des sinistres en fonction de leur état 39
2.11 Coût moyen des sinistres en fonction de la durée de règlement et selon leur état 40
2.12 Coût moyen des sinistres en fonction du taux d'AIPP max des victimes 41
2.13 Les 10 évaluations à l'origine les plus fréquentes dans la base d'étude 42
2.14 Coût moyen des sinistres en fonction de la section de gestion sinistres au 31/12/2017 . . 43
2.15 Coût moyen des sinistres en fonction de la nature judiciaire (Oui/Non) 43

3.1 Représentation d'un sinistre RBNS 46
3.2 Exemple de grille de taille $P \times \delta$ pour $P = 7$ 48
3.3 Présentation de la méthode pour $p = 2$ 49
3.4 Découpage de la base en entrée des modèles 51

4.1 Etapes de construction d'une forêt aléatoire 55
4.2 Présentation de l'algortihme du *Gradient Boosting* 57
4.3 RMSE sur l'échantillon *Out-Of-Bag* en fonction du nombre d'arbres *(num.trees)* et du nombre de variables candidates *(mtry)* 62
4.4 RMSE en fonction du nombre d'arbres pour $mtry = 8$ 63
4.5 Illustration de la *K-fold Cross Validation* pour $K = 5$ 65
4.6 Initialisation du nombre d'itérations 66
4.7 Calibrage du couple *(interaction.depth,n.minobsinnode)* 67
4.8 Calibrage du couple *(shrinkage,n.trees)* 68
4.9 Valeurs optimales des hyperparamètres du *Gradient Boosting* pour le 1er modèle de projection 69
4.10 RMSE/R^2/MAE obtenus pour chaque modèle 71

4.11 Coûts ultimes prédits VS observés pour chaque modèle 73
4.12 Densité empirique des erreurs de prédiction pour chaque modèle 75
4.13 Importance des variables pour le 1er modèle implémenté à l'aide des forêts aléatoires . . 77
4.14 Importance des variables pour le 1er modèle implémenté à l'aide du *Gradient Boosting* . 77
4.15 Importance des variables pour le 2ème modèle implémenté à l'aide des forêts aléatoires . 78
4.16 Importance des variables pour le 2ème modèle implémenté à l'aide du *Gradient Boosting* 78

5.1 Comparaison de la méthode de projection individuelle avec Chain Ladder 82
5.2 Triangle de liquidation par année de déclaration au 31/12/2015 83
5.3 Vérification de l'existence des facteurs de développement pour les années de développement 1 à 4 84
5.4 Indépendance des facteurs de développement 84
5.5 Application de la méthode Chain Ladder 85
5.6 Comparaison des estimations Chain Ladder avec les montants observés par année de déclaration 86
5.7 Erreurs de prédiction de la méthode Chain Ladder 86
5.8 Facteurs de développement observés au 31/12/2017 pour les années de déclaration 2009 à 2015 87
5.9 Modification de la méthode de projection individuelle pour le backtesting : exemple avec les sinistres déclarés en 2014 88
5.10 Comparaison des estimations issues de la méthode de projection individuelle avec les coûts observés par année de déclaration 90
5.11 Comparaison des estimations issues de la méthode de projection individuelle et Chain Ladder avec les coûts observés par année de déclaration 91

Annexes

A Recours forfaitaires des conventions inter-assureurs en Responsabilité Civile automobile

- Forfaits IRCA dans le cas d'un taux AIPP nul depuis 2009

Année	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*
Forfaits IRCA	1490	1490	1490	1490	1518	1518	1518	1518	1480
Variation annuelle (en %)		0 %	0 %	0 %	+ 1,9 %	+ 2,7 %	0 %	0 %	- 2,5 %

- Forfaits IRSA depuis 2009

Année	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*
Forfaits IRSA	1204	1204	1236	1236	1242	1276	1308	1354	1420
Variation annuelle (en %)		0 %	+ 2,7 %	0 %	+ 0,5 %	+ 2,7 %	+ 2,5 %	+ 3,5 %	+ 4,9 %

B Test statistique de Mann - Whitney

La présentation du test de Mann - Withney est issue du support de R.Rakotomalala [20].

La première étape consiste à associer à chaque observation i un rang noté r_i sur l'ensemble des données. On peut alors calculer la somme des rangs (notée S_k) pour chaque sous-échantillon k de taille n_k :

$$S_k = \sum_{i=1}^{n_k} r_{i,k}$$

Notons ensuite θ le paramètre de translation reflètant le décalage entre les fonctions de répartition des deux échantillons.

Le test bilatéral est fomulé de la façon suivante :

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

Les deux quantités suivantes sont calculées :

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2} \qquad\qquad U_2 = S_2 - \frac{n_2(n_2+1)}{2}$$

La statistique du test de Mann - Whitney correspond au minimum des deux quantités U_1 et U_2 :

$$U = min(U_1, U_2)$$

Sous l'hypothèse H_0 et dans le cas où les échantillons sont suffisamment grands, la statistique U converge vers une loi normale de paramètres :

$$\mathbb{E}[U] = \frac{1}{2}n_1n_2 \qquad\qquad Var(U) = \frac{1}{12}(n_1+n_2+1)n_1n_2$$

Dans ce cas précis, la statistique centrée réduite du test bilatéral s'exprime de la façon suivante :

$$Z = \frac{U - \mathbb{E}[U]}{\sqrt{Var(U))}} = \frac{U - \frac{1}{2}n_1n_2}{\frac{1}{12}(n_1+n_2+1)n_1n_2}$$

La région critique du test (notée $R.C.$) pour un seuil de significativité α peut ainsi être obtenue :

$$R.C. : |Z| \geq q_{1-\frac{\alpha}{2}}$$

Avec $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

C Liste des variables explicatives

Variables explicatives	*Source*
Mois_survenance	Base sinistres
Sinistres_ida	Base sinistres
Taux_resp_IDA	Base sinistres
Taux_resp_actualise	Base sinistres
Cause_sinistre	Base sinistres
Cause_sinistre2	Base sinistres
Nature_judiciaire	Base sinistres
Delai_ouverture_sinistre	Base sinistres
Evaluation_origine	Base sinistres
Sect_gestion_sini_ouverture	Base sinistres
Moto	Text Mining
Judiciaire	Text Mining
Alcool_drogue	Text Mining
Perte_de_controle	Text Mining
Sens_inverse	Text Mining
Grave	Text Mining
Type_victime	Base victimes
Nb_vicitmes	Base victimes
Type_Vehicule	Base victimes
Age_moyen_victime	Base victimes
Nb_ayant_droits	Base victimes
Qualite_victime	Base victimes
Nature_prejudice	Base victimes
Taux_max_AIPP	Base victimes
Cout_actuel_i	Base sinistres
Cout_eval_as_if_i	Base sinistres
Nb_reglements_i	Base compable
Nature_reglements_i	Base compable
Info_IRCA_i	Base compable
Sect_gest_sinistre_ex_i	Base sinistres
Duree_reg_sini_periode_i	Base sinistres
Presence_recours_i	Base compable

Les variables présentées en bleu ne sont pas historisées, elles restent identiques dans tous les modèles de projection. Les variables présentées en rouge sont historisées, elles sont observées à la fin de chaque année de développement (notée i ici).

D Valeurs optimales des hyperparamètres pour chaque modèle de projection

- Forêts aléatoires

Modèle	*1*	*2*	*3*	*4*	*5*
num.trees	1 440	1 250	1 200	1 250	1 000
mtry	8	19	23	23	27

- *Gradient Boosting*

Modèle	*1*	*2*	*3*	*4*	*5*
n.minobsinnode	25	0	0	25	0
interaction.depth	6	8	5	5	6
n.trees	1 200	1 300	1 100	1 400	1 100
shrinkage	0,006	0,004	0,006	0,003	0,003